TU Delft

Delft University of Technology

Document Version
Final published version

Licence
Dutch Copyright Act (Article 25fa)

Citation (APA)
Zhao, Y., Kellman, P., Xue, H., Yang, T., Zhang, Y., Han, Y., Simonetti, O., & Tao, Q. (2026). Reverse Imaging for Wide-Spectrum Generalization of Cardiac MRI Segmentation. In J. C. Gee, J. Hong, C. H. Sudre, P. Golland, D. C. Alexander, J. E. Iglesias, A. Venkataraman, & J. H. Kim (Eds.), *Medical Image Computing and Computer Assisted Intervention , MICCAI 2025 - 28th International Conference, 2025, Proceedings* (pp. 555-565). (Lecture Notes in Computer Science; Vol. 15962 LNCS). Springer. https://doi.org/10.1007/978-3-032-04947-6_53

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

James C. Gee · Daniel C. Alexander · Jaesung Hong ·
Juan Eugenio Iglesias · Carole H. Sudre ·
Archana Venkataraman · Polina Golland ·
Jong Hyo Kim · Jinah Park (Eds.)

LNCS 15962

Medical Image Computing and Computer Assisted Intervention – MICCAI 2025

28th International Conference
Daejeon, South Korea, September 23–27, 2025
Proceedings, Part III





MICCAI

Springer

MOREMEDIA

Lecture Notes in Computer Science **15962**

Founding Editors

Gerhard Goos
Juris Hartmanis

Editorial Board Members

Elisa Bertino, *Purdue University, West Lafayette, IN, USA*
Wen Gao, *Peking University, Beijing, China*
Bernhard Steffen, *TU Dortmund University, Dortmund, Germany*
Moti Yung, *Columbia University, New York, NY, USA*

The series Lecture Notes in Computer Science (LNCS), including its subseries Lecture Notes in Artificial Intelligence (LNAI) and Lecture Notes in Bioinformatics (LNBI), has established itself as a medium for the publication of new developments in computer science and information technology research, teaching, and education.

LNCS enjoys close cooperation with the computer science R & D community, the series counts many renowned academics among its volume editors and paper authors, and collaborates with prestigious societies. Its mission is to serve this international community by providing an invaluable service, mainly focused on the publication of conference and workshop proceedings and postproceedings. LNCS commenced publication in 1973.

James C. Gee · Daniel C. Alexander ·
Jaesung Hong · Juan Eugenio Iglesias ·
Carole H. Sudre · Archana Venkataraman ·
Polina Golland · Jong Hyo Kim · Jinah Park
Editors

Medical Image Computing and Computer Assisted Intervention – MICCAI 2025

28th International Conference
Daejeon, South Korea, September 23–27, 2025
Proceedings, Part III

Springer

Editors
James C. Gee
University of Pennsylvania
Philadelphia, PA, USA

Jaesung Hong
DGIST
Daegu, Korea (Republic of)

Carole H. Sudre
University College London
London, UK

Polina Golland
MIT
Cambridge, MA, USA

Jinah Park
KAIST
Daejeon, Korea (Republic of)

Daniel C. Alexander
University College London
London, UK

Juan Eugenio Iglesias
Massachusetts General Hospital and Harvard Medical School
Charlestown, MA, USA

Archana Venkataraman
Boston University
Boston, MA, USA

Jong Hyo Kim
Seoul National University
Seoul, Korea (Republic of)

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-032-04946-9 ISBN 978-3-032-04947-6 (eBook)
https://doi.org/10.1007/978-3-032-04947-6

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

We are pleased to present the proceedings of the 28th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2025). For the first time in its history, the MICCAI conference was held in South Korea, taking place at the Daejeon Convention Center in Daejeon from September 23 to 27, 2025. The conference included a variety of Satellite Events held on September 23 and 27, featuring 99 workshops, tutorials, and challenges. This year marked a significant milestone for the Clinical Translation in Medical Image Computing and Computer-Assisted Intervention (CLINICCAI) conference, which held its 5th edition on September 25. Unlike in the past when it was organized independently, this year's CLINICCAI was prepared in close collaboration with the MICCAI Organizing Committee for the first time, further strengthening the conference's emphasis on clinical translation and engagement. Open science was another continuing focus for advancement at the meeting as the conference hosted the 2nd Open Data Session on September 26.

MICCAI 2025 set a new record for submissions, with a 28% increase compared to 2024. We received 3677 full submissions, of which 3447 advanced to full review. One hundred fifty-nine papers were desk-rejected as they did not comply with submission guidelines, and 71 papers were withdrawn, including 60 that violated the double submission rules. The number of papers accepted for presentation after review was 1027, which represents an acceptance rate of about 30%. These papers comprise the 16 volumes of these Lecture Notes in Computer Science (LNCS) proceedings as follows:

- Part I, LNCS Volume 15960: Multimodal Fusion and Contextual Reasoning in Medical Imaging
- Part II, LNCS Volume 15961: Surgical Navigation, Scene Understanding, and Video Modeling
- Part III, LNCS Volume 15962: Learning and Augmented Reality for Surgical and Endoscopic Applications (I)
- Part IV, LNCS Volume 15963: Learning and Augmented Reality for Surgical and Endoscopic Applications (II)
- Part V, LNCS Volume 15964: Graph-Based Methods in Medical Imaging
- Part VI, LNCS Volume 15965: Datasets and Methods for Image Quality Enhancement
- Part VII, LNCS Volume 15966: Trustworthy and Responsible AI for Medical Imaging
- Part VIII, LNCS Volume 15967: Multimodal Learning for Diagnosis, Risk Prediction, and Survival Analysis
- Part IX, LNCS Volume 15968: Core Techniques in Medical Imaging: Segmentation, Registration, Synthesis, Reconstruction, and Other Emerging Methods (I)
- Part X, LNCS Volume 15969: Core Techniques in Medical Imaging: Segmentation, Registration, Synthesis, Reconstruction, and Other Emerging Methods (II)
- Part XI, LNCS Volume 15970: Core Techniques in Medical Imaging: Segmentation, Registration, Synthesis, Reconstruction, and Other Emerging Methods (III)
- Part XII, LNCS Volume 15971: Core Techniques in Medical Imaging: Segmentation, Registration, Synthesis, Reconstruction, and Other Emerging Methods (IV)

- Part XIII, LNCS Volume 15972: Adapting Foundation Models for Medical Imaging: LLMs, VLMs, and Cross-Domain Generalization (I)
- Part XIV, LNCS Volume 15973: Adapting Foundation Models for Medical Imaging: LLMs, VLMs, and Cross-Domain Generalization (II)
- Part XV, LNCS Volume 15974: Adapting Foundation Models for Medical Imaging: LLMs, VLMs, and Cross-Domain Generalization (III)
- Part XVI, LNCS Volume 15975: Statistical Techniques in Medical Imaging: Causality, Imputation, Weak Supervision, and Other Methods

The papers for the proceedings were selected after a rigorous double-blind peer-review process with three reviews per paper on average. The MICCAI 2025 Program Committee (PC) consisted of 187 area chairs (ACs), over 2400 scientific reviewers, and 12 dual-submissions reviewers, with representation from 61 countries across all major continents. About a quarter (22%) of area chairs self-identified as women. Each area chair managed the review of roughly 34 papers, and the average workload was four papers per reviewer. Area chairs and reviewers were selected following public calls to the community, and were vetted by the program chairs. As in the past, the Microsoft Conference Management Toolkit (CMT) was used for paper management and organization of the overall review process. The Toronto Paper Matching System (TPMS) was, again, employed to help ensure papers were evaluated by area chairs and reviewers with the appropriate expertise. While special effort was made this year to improve the quality of paper matches by reworking both TPMS formulae and CMT categorization of papers and PC member profiles, we recognize the ongoing challenge of securing high-quality reviews in the face of ever-increasing numbers of submissions.

Based on the double-blinded reviews, area chairs' recommendations, and program chairs' decisions, 9% (310) of papers were provisionally accepted without rebuttal, 48% (1663) were rejected, and 43% (1474) proceeded to the rebuttal stage. Final decisions were made based on all available information, including paper screening, adherence to the author submission guidelines, reviews, rebuttals, 2–3 independent meta-reviews, communications during the review process, and, in some cases, additional assessment by the program chairs. A dedicated team was formed to enforce dual submissions policy. The program committee scrutinized papers flagged during reviews, manuscripts of special interest, potential plagiarism, and appeals from authors. The committee was particularly attentive to suggestions of misuse of LLMs both in submissions and reviews - an increasingly vexing issue not only at MICCAI but in science in general.

The 1027 accepted papers that constitute this year's scientific program originated from 43 countries. Seventy-nine percent of these papers are from the following five countries, as identified by a paper's first author: China, including Hong Kong SAR (501), USA (118), Germany (66), South Korea (64), and UK (62). The next five countries in descending order of the number of accepted papers are France (26), Australia (22), Canada (22), Switzerland (17), and Singapore (15). Nineteen papers were from lower and middle-income countries in Asia, a region highlighted by this MICCAI edition toward the conference's continuing efforts to broaden its global impact.

To build the oral program of 76 presentations, accepted papers were first clustered by technical theme. For each theme, oral presentation candidates were chosen from the papers that received overwhelming enthusiasm among the area chairs. These candidates

were then narrowed down to the final list by ensuring the program was representative of the range of clinical applications covered by the papers and reflected both geographic and demographic inclusivity among its presenters. Additionally, 36 posters were highlighted through rapid talks ahead of their presentation at subsequent poster sessions. These pieces of work were recommended by ACs for their likelihood to provoke discussion and for specific contributions to the community in the form of new open-source datasets.

MICCAI 2025 centered around two overarching themes. The first, "From Concept to Clinical," emphasized the translational journey of MICCAI methodologies: from foundational research to real-world clinical application. To support this theme, we actively encouraged industry engagement through dedicated panels and presentations, and featured mentoring programs with industry leaders in collaboration with the MICCAI Student Board.

The second theme, "Pan-Asian Connections," highlighted the growing collaboration within Asian research communities and reinforced MICCAI's ongoing commitment to inclusion and outreach. Initiatives under this theme included targeted mentorship programs for young researchers, from writing technical papers to navigating the submission and rebuttal process. Another key initiative involved real-time broadcasts of the conference by selected young researchers across Asia who facilitated local community discussions, particularly in countries historically underrepresented at MICCAI.

Furthermore, MICCAI 2025 spotlighted focused discussions on healthcare challenges across Asia, with special attention to underrepresented countries in Area 3. These efforts culminated in a special plenary session titled "Challenges and Opportunities of Medical AI in Asia." With over half of the world's population residing in Asia, the region represents immense potential and need for developing scalable and generalizable healthcare AI solutions. The recent increase in cross-border research collaboration further positions Asia as a fertile ground for piloting innovative medical technologies. This highlight session brought together leading experts from South Korea, Japan, China, Singapore, Mongolia, and Nepal to share insights on the current landscape, region-specific challenges, and the future of AI integration in healthcare. These collective efforts were made possible through the unique platform provided by the MICCAI meeting.

To further the mission of enhancing inclusion at MICCAI, we offered travel grants to candidates from lower-to-middle-income countries as well as from countries historically underrepresented in the MICCAI community, including participants from Africa, Asia, and Latin America. We also provided partial support to additional participants through the MICCAI Society Registration Grants. Funding was made possible by generous support from the MICCAI Society, donations from individual groups within the MICCAI community, and financial support from non-commercial sponsors. Support for US-based students and early-career investigators was also provided in part by the National Institutes of Health (NIH) through conference grant R13EB035922 from the National Institute of Biomedical Imaging and Bioengineering. The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the NIH; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government. We additionally supported young Asian researchers through a conference outreach program

that provided registration and accommodation support, made possible by the generous support of the City of Daejeon.

A meeting of MICCAI's scale and impact is only possible through the collective dedication and tireless efforts of many individuals, most notably, the members of this year's organizing committees. While every member played a vital role, we would like to especially recognize several key contributors. Luping Zhou, Fernando Bello, and Nicholas Heller led the Workshops, Tutorials, and Challenges teams, respectively, in reviewing a total of 191 Satellite Event proposals and assembling a diverse, high-quality program. We are particularly grateful to our Satellite Program Chair, Michal Kozubek, for skillfully coordinating the 99 accepted events into a cohesive and well-structured program. We also recognize Lena Maier-Hein for her primary role in launching the Lighthouse Challenges this year. The Industry Engagement Committee, headed by Stephen Aylward, established a strong industry presence that brought our conference themes to life. Tanveer Syeda-Mahmood coordinated both the industry and clinical panel sessions. The Young Researchers Initiative Committee, under the leadership of Minjeong Kim, introduced a new and impactful program that further enriched the participation and engagement of early-career researchers.

We also appreciate the exemplary work of our Sponsorship team, Kwang Gi Kim, Mohammad Yaqub, Jihoon Cho, and Maria Valdés Hernández, for their efforts in securing conference financial support despite the challenging economic climate. Joonbeom Seo and Jonghye Woo, together with the CLINICCAI Program Committee, ensured the continued success of CLINICCAI as a distinctive flagship of MICCAI. Nassir Navab and Seong Tae Kim provided consummate oversight of the Outreach Program, helping to further expand MICCAI's reach into underrepresented regions. Martijn Starmans, Apostolia Tsirikoglou, and Namkug Kim led with distinction in advancing the Open Data Initiative.

A tremendous amount of work was carried out by the Local Organizing Committee, chaired by Sanghyun Park, whose commitment and meticulous attention to every detail were critical to the smooth execution of all parts of the meeting. We are also grateful to Jaeil Kim, Finance Chair, as well as our Communications and Signage Chairs, Keri Kim, Won Hwa Kim, and Sangrok Jin, whose expert support in managing various elements of the local venue helped ensure a pleasant experience for all participants. Seungryong Cho directed the Student Volunteers with enthusiasm, while Hyunjoo Jenny Lee successfully organized KAIST Day, creating a memorable and festive pre-conference event. Jae Sung Lee, Ilwoo Lyu, and Joonho Seo provided impeccable support to facilitate networking and foster meaningful connections and engagement among the community throughout the event.

We are deeply grateful to our local supporting institutions, KAIST (Korea Advanced Institute of Science and Technology) and KSIIM (Korea Society of Imaging Informatics in Medicine), for their invaluable organizational and logistical support. We also wish to extend our heartfelt thanks to the City of Daejeon and the Daejeon Tourism Organization for their generous contributions, which made it possible to bring MICCAI 2025 to South Korea.

The MICCAI Student Board, chaired by Naren Akash, contributed significantly by organizing highly popular networking and social events, including the Ph.D. Thesis

Madness as well as the Doctoral Consortium, which was held for the first time and offered a valuable forum for recent graduates to present their work and engage with the broader community. Likewise, the Women in MICCAI and RISE communities, under the guidance of their respective presidents, Ruogu Fang and Esther Puyol Anton, presented topical events that further strengthened the inclusivity and depth of our technical program.

We would also like to express our sincere gratitude to the MICCAI Society Board of Directors and in particular to its Chair, Marius George Linguraru, and former Chair, Caroline Essert, for their support and constructive feedback. We gratefully acknowledge, too, the contributions of the MICCAI Society secretariat personnel, Janette Wallace and Johanne Langford, who oversaw logistics and budgets, and Diana Cunningham, Anna Van Vliet, and John Baxter for the timely conference announcements in the MICCAI Society newsletters and MICCAI Morning issues. The proficient on-site organization of the conference, including budget management and event operations, was the commendable work of our Professional Conference Organizer, Dekon Congress & Tourism, led by Mehmet Eldegez. Dekon's essential efforts were carried out in partnership with Lorraine Han and her team at the MICE Bureau of the Daejeon Tourism Organization, to whom we offer our wholehearted thanks.

The proceedings of MICCAI 2025 are the fruit of the extraordinary work of our entire community. Kitty Wong, the MICCAI Submission System Manager, deserves special mention and our profound thanks for her crucial support throughout the paper submission, review, program planning, and proceedings preparation process. We are particularly grateful, too, to the members of the Program Committee for their dedicated and diligent work in the reviewer assignments and final paper selection, and to all the reviewers for their commitment and support during the paper evaluation process. Congratulations to the outstanding area chairs and reviewers! Finally, and most importantly, we reserve our deepest appreciation and gratitude for all the authors for their effort in preparing their manuscripts and supporting the conference through their submissions. The success of MICCAI 2025 is a direct reflection of the high-quality work presented at the meeting. Our warmest congratulations to all!

In closing, we would like to thank once more every individual and team involved in the organization of MICCAI 2025. Their inspiring dedication, professionalism, and collaborative spirit made this year's conference a truly outstanding and memorable

event. We hope you enjoyed MICCAI 2025 in Daejeon, the conference's first meeting in South Korea. We now look forward to seeing you next year in Abu Dhabi at MICCAI 2026.

September 2025

James C. Gee
Daniel Alexander
Jaesung Hong
Juan Eugenio Iglesias
Carole Sudre
Archana Venkataraman
MICCAI 2025 Program Chairs

Jinah Park
Polina Golland
Jonghyo Kim
MICCAI 2025 General Chairs

Organization

General Chairs

Jinah Park	Korea Advanced Institute of Science and Technology, South Korea
Polina Golland	Massachusetts Institute of Technology, USA
Jonghyo Kim	Seoul National University, South Korea

Program Committee Chairs

James C. Gee	University of Pennsylvania, USA
Daniel C. Alexander	University College London, UK
Juan Eugenio Iglesias	Massachusetts General Hospital and Harvard Medical School, USA
Jaesung Hong	Daegu Gyeongbuk Institute of Science and Technology, South Korea
Carole H. Sudre	University College London, UK
Archana Venkataraman	Boston University, USA

Satellite Events Chairs

Michal Kozubek	Masaryk University, Czech Republic
Masahiro Oda	Nagoya University, Japan

Workshops

Luping Zhou (Chair)	University of Sydney, Australia
Masahiro Oda	Nagoya University, Japan
Heung-Il Suk	Korea University, South Korea
Guorong Wu	University of North Carolina at Chapel Hill, USA

Challenges

Lena Maier-Hein (Chair)	German Cancer Research Center (DKFZ), Germany
Nicholas Heller (Co-chair)	Cleveland Clinic, USA
Suyash Awate	Indian Institute of Technology Bombay, India
Sophia Bano	University College London, UK
Jun Ma	University Health Network, Canada
Annika Reinke	German Cancer Research Center (DKFZ), Germany

Tutorials

Fernando Bello (Chair)	Imperial College London, UK and Duke-NUS Medical School, Singapore
Diana Mateus	Centrale Nantes LS2N, France
Daniel H. Pak	Yale University, USA
Chang Min Park	Seoul National University College of Medicine, South Korea
Hoo-Chang Shin	NVIDIA, USA

MICCAI-Clinical (CLINICCAI)

Joon Beom Seo (Chair)	University of Ulsan College of Medicine, South Korea
Steven Wong	Sengkang General Hospital, Singapore
Jonghye Woo	Massachusetts General Hospital and Harvard Medical School, USA

Industry Engagement

Stephen Aylward (Chair)	NVIDIA, USA
Kyu-Hwan Jung (Co-chair)	Sungkyunkwan University, South Korea
Alex Lemm	Amazon Web Services, Germany
Ma Luo	Medtronic, USA
Brendon Lutnick	Johnson & Johnson Innovative Medicine, USA

Sponsorship

Kwang Gi Kim (Co-chair)	Gachon University, South Korea
Mohammad Yaqub (Co-chair)	Mohamed bin Zayed University of Artificial Intelligence, UAE
Jihoon Cho	Samsung Research, South Korea
Maria Valdés Hernández	University of Edinburgh, UK

Young Scholars Initiative

Minjeong Kim (Chair)	University of North Carolina at Greensboro, USA
Ehsan Adeli	Stanford University, USA
Seong Jae Hwang	Yonsei University, South Korea
Anees Kazi	Massachusetts General Hospital and Harvard Medical School, USA
Islem Rekik	Imperial College London, UK

Outreach

Nassir Navab (Co-chair)	Technical University of Munich, Germany
Seong Tae Kim (Co-chair)	Kyung Hee University, South Korea
Shadi Albarqouni	Universität Bonn and Helmholtz Munich, Germany
Shoji Kido	Osaka University, Japan
Namkug Kim	University of Ulsan College of Medicine, South Korea
Anna Seo	SEEANN Solution Co., South Korea

Open Data Chairs

Martijn Starmans	Erasmus University Medical Center, Netherlands
Apostolia Tsirikoglou	Karolinska Institutet, Sweden
Namkug Kim	University of Ulsan College of Medicine, South Korea
Lidia Garrucho Moras	Universitat de Barcelona, Spain
Kaouther Mouheb	Erasmus University Medical Center, Netherlands

Panels Chair

Tanveer Syeda-Mahmood	IBM Research, USA

Local Organization Committee

Sang Hyun Park (Chair)	Daegu Gyeongbuk Institute of Science and Technology, South Korea

Finance Chair

Jaeil Kim	Kyungpook National University, South Korea

Communications and Signage Chairs

Keri Kim	Korea Institute of Science and Technology, South Korea
Won Hwa Kim	Pohang University of Science and Technology, South Korea
Sangrok Jin	Pusan National University, South Korea

Student Volunteers Chair

Seungryong Cho	Korea Advanced Institute of Science and Technology, South Korea

KAIST Day Chair

Hyunjoo Jenny Lee	Korea Advanced Institute of Science and Technology, South Korea

Network and Arrangements

Jae Sung Lee	Seoul National University College of Medicine, South Korea
Ilwoo Lyu	Pohang University of Science and Technology, South Korea
Joonho Seo	Korea Institute of Machinery & Materials, South Korea

Convention Liaison

Lorrain Han	Daejeon Tourism Organization, South Korea
Jinny Seo	Daejeon Tourism Organization, South Korea
Hazel Seok	Daejeon Tourism Organization, South Korea

Women in MICCAI President

Ruogu Fang	University of Florida, USA

RISE Committee President

Esther Puyol-Antón	Heartflow and King's College London, UK

Submission Platform Manager

Kitty Wong	MICCAI Society, Canada

Program Committee

Pablo Arbelaez	Universidad de los Andes, Colombia
Angelica I. Aviles-Rivero	Tsinghua University, China
Wenjia Bai	Imperial College London, UK
Yaël Balbastre	University College London, UK
Abhirup Banerjee	University of Oxford, UK
Omri Bar	Theator, Israel
Neslihan Bayramoglu	University of Oulu, Finland

Cosmin I. Bercea	Technical University Munich and Helmholtz AI and Helmholtz Center Munich, Germany
Ryoma Bise	Kyushu University, Japan
Esther E. Bron	Erasmus University Medical Center, Netherlands
Mariano Cabezas	University of Sydney, Australia
Weidong Cai	University of Sydney, Australia
Chen Chen	University of Sheffield, UK
Cheng Chen	Chinese University of Hong Kong, China
Elvis C. S. Chen	Robarts Research Institute, Canada
Hao Chen	Hong Kong University of Science and Technology, China
Jianan Chen	University College London, UK
Zhen Chen	Yale University, USA
Jang-Hwan Choi	Ewha Womans University, South Korea
Zhiming Cui	ShanghaiTech University, China
Kang Dang	Xi'an Jiaotong-Liverpool University, China
Eleonora D'Arnese	University of Edinburgh, UK
Charles Delahunt	Global Health Labs, USA
Ruining Deng	Weill Cornell Medicine, USA
Christian Desrosiers	ÉTS Montréal, Canada
Blake E. Dewey	Johns Hopkins University, USA
Jose Dolz	ÉTS Montréal, Canada
Niharika S. D'Souza	IBM Research, USA
Nicha Dvornek	Yale University, USA
Dmitry V. Dylov	Skoltech, Russia
Mostafa El Habib Daho	University of Western Brittany, France
Deng-Ping Fan	Nankai University, China
Ruogu Fang	University of Florida, USA
Azade Farshad	Technical University of Munich, Germany
Aasa Feragen	Technical University of Denmark, Denmark
Huazhu Fu	A*STAR Institute of High Performance Computing, Singapore
Jin Kyu Gahm	Pusan National University, South Korea
Adrian Galdran	Universitat Pompeu Fabra, Spain
Mingchen Gao	University at Buffalo, SUNY, USA
Elodie Germani	Universitätsklinikum Bonn, Germany
Syed Zulqarnain Gilani	Edith Cowan University, Australia
Karthik Gopinath	Massachusetts General Hospital and Harvard Medical School, USA
Pietro Gori	Télécom Paris, France
Yun Gu	Imperial College London, UK
Xiaoqing Guo	University of Oxford, UK

Ilker Hacihaliloglu	University of British Columbia, Canada
Stathis Hadjidemetriou	University of Limassol, Cyprus
Hu Han	Institute of Computing Technology, Chinese Academy of Sciences, China
Yi Hong	Shanghai Jiao Tong University, China
Yoonmi Hong	University of North Carolina at Chapel Hill, USA
Mobarak I. Hoque	University College London, UK
Benjamin Hou	Imperial College London, UK
Xiaoling Hu	Harvard Medical School, USA
Xiaowei Hu	South China University of Technology, China
Junzhou Huang	University of Texas at Arlington, USA
Jana Hutter	King's College London, UK
Seong Jae Hwang	Yonsei University, South Korea
Abdullah-Al-Zubaer Imran	University of Kentucky, USA
Amir Jamaludin	University of Oxford, UK
Yuanfeng Ji	University of Hong Kong, China
Jue Jiang	Memorial Sloan Kettering Cancer Center, USA
Meirui Jiang	Chinese University of Hong Kong, China
Tingting Jiang	Peking University, China
Xiajun Jiang	Rowan University, USA
Zhifan Jiang	Children's National Hospital, USA
Dakai Jin	Alibaba USA Inc., USA
Yueming Jin	National University of Singapore, Singapore
Leo Joskowicz	Hebrew University of Jerusalem, Israel
Davood Karimi	Harvard University, USA
Youngwook Kee	Stony Brook University, USA
Marta A. Kersten-Oertel	Concordia University, Canada
Jinman Kim	University of Sydney, Australia
Minjeong Kim	University of North Carolina at Greensboro, USA
Won Hwa Kim	Pohang University of Science and Technology, South Korea
Jin Tae Kwak	Korea University, South Korea
Christian Ledig	University of Bamberg, Germany
Baiying Lei	Shenzhen University, China
Gang Li	University of North Carolina at Chapel Hill, USA
Guang Li	Hokkaido University, Japan
Hongwei Bran Li	Harvard Medical School, USA
Lei Li	National University of Singapore, Singapore
Xiaoxiao Li	University of British Columbia, Canada
Zhen Li	Chinese University of Hong Kong, Shenzhen, China
Hongxiang Lin	Zhejiang Lab, China

Daochang Liu	University of Western Australia, Australia
Hong Liu	Institute of Computing Technology, Chinese Academy of Sciences, China
Peirong Liu	Johns Hopkins University, USA
Jingpei Lu	Intuitive Surgical, USA
Le Lu	Alibaba Group, USA
Gongning Luo	Harbin Institute of Technology, China
Jax Luo	Cleveland Clinic/Harvard Medical School, USA
Luyang Luo	Harvard University, USA
Xiangde Luo	Stanford University, USA
Andreas K. Maier	FAU Erlangen-Nuremberg, Germany
Diana Mateus	Centrale Nantes, France
Qingjie Meng	University of Birmingham, UK
Yanda Meng	University of Exeter, UK
Deepak Mishra	Indian Institute of Technology Jodhpur, India
Mehdi Moradi	McMaster University, Canada
Rodrigo Moreno	KTH Royal Institute of Technology, Sweden
Saad Nadeem	Memorial Sloan Kettering Cancer Center, USA
Dong Nie	University of North Carolina at Chapel Hill, USA
Masahiro Oda	Nagoya University, Japan
Yoshito Otake	Nara Institute of Science and Technology, Japan
Cheng Ouyang	University of Oxford, UK
Johannes C. Paetzold	Weill Cornell Medicine, USA
Yongsheng Pan	Northwestern Polytechnical University, China
Tingying Peng	Helmholtz AI, Germany
Terry Peters	Robarts Research Institute, Canada
Szymon Płotka	University of Warsaw, Poland
Prateek Prasanna	Stony Brook University, USA
Yulei Qin	Tencent, China
Shi Qiu	Chinese University of Hong Kong, China
Wu Qiu	Huazhong University of Science and Technology, China
Liangqiong Qu	University of Hong Kong, China
Daniele Ravì	University of Messina, Italy
Monjoy Saha	National Institutes of Health, USA
Lalithkumar Seenivasan	Johns Hopkins University, USA
Hongming Shan	Fudan University, China
Yiyu Shi	University of Notre Dame, USA
Keewon Shin	Korea University Anam Hospital, South Korea
Sahar S. Soussa	Nile University, Egypt
Vinkle Kumar Srivastav	University of Strasbourg, France
Ruisheng Su	Eindhoven University of Technology, Netherlands

Yao Sui	Peking University, China
Kyung Hyun Sung	University of California, Los Angeles, USA
Tanveer Syeda-Mahmood	IBM Research, USA
Lin Tian	Massachusetts General Hospital and Harvard Medical School, USA
Aleksei Tiulpin	University of Oulu, Finland
Martin Urschler	Medical University Graz, Austria
Marta Varela	Imperial College London, UK
Harini Veeraraghavan	Memorial Sloan Kettering Cancer Center, USA
Chengjia Wang	Heriot-Watt University, UK
Guotai Wang	University of Electronic Science and Technology of China, China
Hu Wang	Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
Li Wang	University of North Carolina at Chapel Hill, USA
Liansheng Wang	Xiamen University, China
Qian Wang	ShanghaiTech University, China
Renzhen Wang	Xi'an Jiaotong University, China
Shujun Emma Wang	Hong Kong Polytechnic University, China
Shuo Wang	Fudan University, China
Xiaosong Wang	Shanghai AI Laboratory, China
Yan Wang	East China Normal University, China
Donglai Wei	Boston College, USA
Matthias Wilms	University of Calgary, Canada
Jelmer Wolterink	University of Twente, Netherlands
Ye Wu	Nanjing University of Science and Technology, China
Yicheng Wu	Monash University, Australia
Yong Xia	Northwestern Polytechnical University, China
Yiming Xiao	Concordia University, Canada
Yutong Xie	Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
Fuyong Xing	University of Colorado Denver, USA
Xiaohan Xing	Stanford University, USA
Chenchu Xu	Anhui University, China
Xuanang Xu	Rensselaer Polytechnic Institute, USA
Yanwu Xu	South China University of Technology, China
Yanyu Xu	Shandong University, China
Zhe Xu	Chinese University of Hong Kong, China
Wufeng Xue	Shenzhen University, China
Zengqiang Yan	Huazhong University of Science and Technology, China

Jiancheng Yang	Swiss Federal Technology Institute of Lausanne, Switzerland
Qiushi Yang	City University of Hong Kong, China
Yanwu Yang	University of Tübingen, Germany
Jiawen Yao	DAMO Academy, Alibaba Group, China
Qingsong Yao	Stanford University, USA
Chuyang Ye	Beijing Institute of Technology, China
Dong Hye Ye	Georgia State University, USA
Zhaozheng Yin	Stony Brook University, USA
Chenyu You	Stony Brook University, USA
Fatemeh Zabihollahy	University of Toronto, Canada
Jianpeng Zhang	Alibaba DAMO Academy, China
Jiong Zhang	Cixi Institute of Biomedical Engineering, Ningbo Institute of Materials Technology and Engineering, Chinese Academy of Sciences, China
Lichi Zhang	Shanghai Jiao Tong University, China
Yanfu Zhang	William and Mary, USA
Yi Zhang	Sichuan University, China
Can Zhao	Nvidia, USA
Fenqiang Zhao	DAMO Academy, Alibaba Group, China
Yitian Zhao	Cixi Institute of Biomedical Engineering, Ningbo Institute of Industrial Technology, Chinese Academy of Sciences, China
Yue Zhao	Chongqing University of Posts and Telecommunications, China
Yushan Zheng	Beihang University, China
Zichun Zhong	Wayne State University, USA
Qin Zhou	East China University of Science and Technology, China
S. Kevin Zhou	University of Science and Technology of China, China
Tao Zhou	Nanjing University of Science and Technology, China
Tianfei Zhou	Beijing Institute of Technology, China
Dajiang Zhu	University of Texas at Arlington, USA
Lei Zhu	Hong Kong University of Science and Technology (Guangzhou), China
Xiaofeng Zhu	University of Electronic Science and Technology of China, China
Ke Zou	Sichuan University, China

Outstanding Area Chairs

Angelica I. Aviles-Rivero	Tsinghua University, China
Mariano Cabezas	University of Sydney, Australia
Martin Urschler	Medical University Graz, Austria
Fuyong Xing	University of Colorado Denver, USA

Outstanding Reviewers

Syed Farhan Abbas	Norwegian University of Science and Technology, Norway
Tajamul Ashraf	Indian Institute of Technology Delhi, India
Matthew Baugh	Imperial College London, UK
Benjamin Billot	Inria, France
Katharina Breininger	University of Würzburg, Germany
Sebastian Doerrich	University of Bamberg, Germany
Catalin Fetita	Télécom SudParis, France
Isabel Funke	National Center for Tumor Diseases (NCT) Dresden, Germany
Zdravko Marinov	Karlsruhe Institute of Technology, Germany
Chinedu Nwoye	University of Strasbourg, France
Gasper Podobnik	University of Ljubljana, Slovenia
Isabella Poles	Politecnico di Milano, Italy
Hatem Rashwan	Rovira i Virgili University, Spain
Amirhossein Sabour	McMaster University, Canada
Weiyi Zhang	Hong Kong Polytechnic University, China
Lianrui Zuo	Vanderbilt University, USA

Honorable Mentions (Reviewers)

Seyed-Ahamad Ahmadi	Technical University of Munich, Germany
Yasmina Al Khalil	Eindhoven University of Technology, Netherlands
Mehdi Astaraki	Karolinska Institutet, Sweden
Mohammad Saber Azimi	Shahid Beheshti University, Iran
Joao Barreto	University of Coimbra, Portugal
Víctor M. Batlle	I3A, Universidad de Zaragoza, Spain
Federico Bolelli	Università degli Studi di Modena e Reggio Emilia, Italy
Ester Bonmati Coll	University of Westminster, UK
Yi Hao Chan	Nanyang Technological University, Singapore

Shengjia Chen	Icahn School of Medicine at Mount Sinai, USA
Ziyang Chen	Northwestern Polytechnical University, China
Philip Chikontwe	Harvard University, USA
Allison Clement	University of Oxford, UK
Francesco Cremonesi	Inria, France
Meghane Decroocq	RIKEN Center for Brain Science, Japan
Yu Deng	King's College London, UK
Kerol Djoumessi	University of Tübingen, Germany
Reuben Dorent	Inria, France
Peiyu Duan	Yale University, USA
Benoit Dufumier	CEA, France
Felix Dülmer	Technical University of Munich, Germany
Jim Duncan	Yale University, USA
Marawan Elbatel	Hong Kong University of Science and Technology, China
Huihui Fang	Nanyang Technological University, Singapore
Greg Fleishman	Howard Hughes Medical Institute, USA
Shangqi Gao	University of Cambridge, UK
Yifan Gao	University of Science and Technology of China, China
Negin Ghamsarian	University of Bern, Switzerland
Catalina Gomez	Johns Hopkins University, USA
Alvaro Gonzalez-Jimenez	University Hospital of Basel, Switzerland
Pengfei Gu	University of Notre Dame, USA
Ida Häggström	Chalmers Tekniska Högskola, Sweden
Shahad Hardan	Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
Hideaki Hayashi	Osaka University, Japan
Ruobing Huang	Shenzhen University, China
Juntao Jiang	Zhejiang University, China
Amelia Jiménez-Sánchez	IT University of Copenhagen, Denmark
Shantanu Joshi	University of California, Los Angeles, USA
Bernhard Kainz	Imperial College London, UK & FAU Erlangen-Nürnberg, Germany
Yunsoo Kim	University College London, UK
Tobias Klinder	Philips, Germany
Nicholas Konz	Duke University, USA
Daniel Lang	Helmholtz Center Munich, Germany
Moon Hwan Lee	Daegu Gyeongbuk Institute of Science and Technology, South Korea
Won Hee Lee	Kyung Hee University, South Korea
Fan Li	Shanghai Jiao Tong University, China

Haiqing Li	University of Texas at Arlington, USA
Jun Li	Technical University of Munich, Germany
Wei Liu	Tongji University, China
Mingyuan Luo	Shenzhen University, China
Ilwoo Lyu	Pohang University of Science and Technology, South Korea
Klaus H. Maier-Hein	German Cancer Research Center (DKFZ), Germany
Mayur Mallya	University of British Columbia, Canada
Pablo Márquez Neila	University of Bern, Switzerland
Evangelos Mazomenos	University College London, UK
Raghav Mehta	Imperial College London, UK
Lia Morra	Politecnico di Torino, Italy
Kaouther Mouheb	Erasmus University Medical Center, Netherlands
Enamundram Naga Karthik	Mila/Polytechnique Montréal, Canada
Sahil Nalawade	Dana-Farber Cancer Institute, USA
Gia Ngo	Cornell University, USA
Ajibola Oladokun	University of Cape Town, South Africa
Antoine Olivier	Owkin, France
Suraj Pai	Mass General Brigham & Harvard Medical School, USA
Matthias Perkonigg	Medical University of Innsbruck, Austria
Shihan Qiu	Siemens Healthineers, USA
Sandro Queiros	University of Minho, Portugal
Samuel Remedios	Johns Hopkins University, USA
Robert Rohling	University of British Columbia, Canada
Numan Saeed	Mohamed Bin Zayed University of Artificial Intelligence, United Arab Emirates
Mostafa Salem	Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
Thomas Sanchez	University of Lausanne, Switzerland
Gary Sarwin	ETH Zurich, Switzerland
Lulin Shi	Hong Kong University of Science and Technology, China
Julio Silva-Rodríguez	ÉTS Montreal, Canada
Emma Stanley	University of Calgary, Canada
Mohammod Naimul Islam Suvon	University of Sheffield, UK
Tomasz Szczepański	Sano - Centre for Computational Personalised Medicine, Poland
Fenghe Tang	University of Science and Technology of China, China
Thomas Tavolara	Mayo Clinic, USA
Kwang-Hyun Uhm	Gachon University, South Korea

Francisco Vasconcelos	University College London, UK
Hooman Vaseli	University of British Columbia, Canada
Trinh Thi Le Vuong	Korea University, South Korea
Thomas Wendler	Technical University of Munich, Germany
Markus Wenzel	Fraunhofer Institute for Digital Medicine MEVIS, Germany
Julia Wolleb	Yale University, USA
McKell Woodland	University of Texas MD Anderson Cancer Center, USA
Fuping Wu	University of Oxford, UK
Jianghao Wu	University of Electronic Science and Technology of China, China
Yunheng Wu	Nagoya University, Japan
Zijian Wu	University of British Columbia, Canada
Geng-Xin Xu	Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
Zikang Xu	University of Science and Technology of China, China
Yingyu Yang	University of Oxford, UK
Chenfei Ye	Harbin Institute of Technology at Shenzhen, China
Siyeop Yoon	Massachusetts General Hospital and Harvard Medical School, USA
Zhen Yu	Monash University, Australia
Ziqi Yu	Fudan University, China
Paul Yushkevich	University of Pennsylvania, USA
Bodong Zhang	University of Utah, USA
Dong Zhang	University of British Columbia, Canada
Tianyi Zhang	Agency for Science, Technology and Research, Singapore
Weihang Zhang	Beijing Institute of Technology, China
Yushan Zheng	Beihang University, China

Mentorship Program (Mentors)

Sèmèvo A. R. M. Ahouandjinou	University of Abomey-Calavi, Benin
Joseph Akinyemi	University of York, UK
Shadi Albarqouni	University of Bonn & Helmholtz AI, Germany
Sharib Ali	University of Leeds, UK
Adel Alii	University of Sfax, Tunisia
Amine Amyar	Harvard Medical School, USA

Syed Anwar	Children's National Hospital, USA
Zhen Chen	Yale University, USA
Kathleen Curran	University College Dublin, Ireland
Idy Diop	Cheikh Anta Diop University, Senegal
Mostafa El Habib Daho	University of Western Brittany, France
Mohammed El Hassouni	Mohammed V University in Rabat, Morocco
Lamiaa Elfadaly	October University for Modern Science and Arts, Egypt
Parastoo Farnia	Tehran University of Medical Sciences, Iran
Huazhu Fu	A*STAR Institute of High Performance Computing, Singapore
Florin-Cristian Ghesu	Siemens Healthineers, Germany
Sanaa Ghouzali	International University of Rabat, Morocco
Syed Zulqarnain Gilani	Edith Cowan University, Australia
Prashnna Gyawali	West Virginia University, USA
Kaiyuan Hu	McGill University, Canada
Uzair Iqbal	University of Malaya, Malaysia
Mobarakol Islam	University College London, UK
Eric Jagwara	Tufuna Technologies, Uganda
Amal Jlassi	University of Tunis El Manar, Tunisia
Wonsik Jung	Korea University, South Korea
Manjunath K. N.	Manipal Institute of Technology, India
Dmitrii Kaplun	China University of Mining and Technology, China
Baiying Lei	Shenzhen University, China
Jianming Liang	Arizona State University, USA
Meharban M. S.	Sheridian, Canada
Michail Mamalakis	University of Cambridge, UK
Subhamoy Mandal	Indian Institute of Technology Kharagpur, India
Mounia Mikram	École des Sciences de l'Information, Morocco
Ahmed Nebli	Heirich Heine University, Düsseldorf, Germany
Akash Parvatikar	HistoWiz Inc., USA
Mengyun Qiao	Imperial College London, UK
Rizwan Qureshi	University of Central Florida, USA
Tariq Rahim	Kingston University London, UK
Maryem Rhanoui	Université Claude Bernard Lyon 1, France
Juan A. Sánchez-Margallo	Jesús Usón Minimally Invasive Surgery Centre, Spain
Sahar Selim Soussa	Nile University, Egypt
Keewon Shin	Korea University College of Medicine, South Korea
Kyung Sung	University of California, Los Angeles, USA

Jiacheng Wang	Manteia Technology Co. Ltd, China
Ying Weng	University of Nottingham Ningbo China, China
Ye Wu	Nanjing University of Science and Technology, China
Bôa Djavph Yoganguina	École Supérieure Polytechnique de Dakar, Senegal
S. Kevin Zhou	University of Science and Technology of China, China
Zongwei Zhou	Johns Hopkins University, USA
Lei Zhu	HKUST (GZ) and HKUST, China

Reviewers

Syed Farhan Abbas
Hamid Abbasi
Zeinab Abboud
Lamiaa Abdel-Hamid
Sanne Abeln
Kumar Abhishek
Purang Abolmaesumi
Burak Acar
Hamidreza Aftabi
Vinti Agarwal
Harshit Agrawal
Antonio Agudo
Seyed-Ahamad Ahmadi
Ninad Aithal
Naren Akash
Faranak Akbarifar
Emre Akbas
Zeynettin Akkus
Fatih Aksu
Dawood Al Chanti
Wafa Al Ghallabi
Yasmina Al Khalil
Khaoula Alaoui Belghiti
Julia Alekseenko
Hassan Alhajj
Shahzad Ali
Sharib Ali
Sikandar Ali
Omar Al-Kadi
Simeon Allmendinger
Mohammed Al-masni
Pablo Alvarez
Charlems Alvarez-Jimenez
Jennifer Alvén
Victor Alves
Lidia Al-Zogbi
Dayeong An
Haill An
Elsa Angelini
Anna Anikina
André Anjos
Sameer Antani
Michela Antonelli
Michel Antunes
Syed Anwar
Yoshimitsu Aoki
Dooman Arefan
Ignacio Arganda-Carreras
Mohammad Ali Armin
Abu Bakor Hayat Arnob
Chetan Arora
Dewa Made Sri Arsa
Muhammad Arsalan
Janan Arslan
Marius Arvinte
Md Ashikuzzaman
Tajamul Ashraf
Mehdi Astaraki
Arda Atalık
Chinmayee Athalye
Thomas Athey
Benjamin Aubert
Marc Aubreville
Chloé Audigier
Ignacio Avellino
Akash Awasthi
Suyash Awate
Ishak Ayad
Nicolás Ayobi
Sana Ayromlou
Md Abulkalam Azad
Mohammad Farid Azampour
Mehran Azimbagirad
Mohammad Saber Azimi
Abu Zahid Bin Aziz
Britty Baby
Hyeon-Min Bae
Seung Baek
Ulas Bagci
Harrison Bai
Jieyun Bai
Long Bai
Muhammad Naseer Bajwa
Eduard Bakstein
Guha Balakrishnan
Abdourahmane Balde
Abhirup Banerjee
Imon Banerjee
Soumyanil Banerjee

Chandan Ganesh Bangalore Yogananda
Jordan Bannister
Jaume Banus
Jamal Banzi
Shunxing Bao
D. M. Bappy
Joao Barreto
Hritam Basak
Tamer Basha
Deepti R. Bathula
Víctor M. Batlle
Matthew Baugh
Christian Baumgartner
John Baxter
Nourhan Bayasi
Siming Bayer
Roza Bayrak
Roozbeh Bazargani
Clément Beitone
Soufiane Belharbi
Simone Bendazzoli
Sean Benson
Mariana Bento
Guy Ben-Yosef
Christos Bergeles
Gabriel Bernardino
Riddhish Bhalodia
Nithya Bhasker
Moinak Bhattacharya
Binod Bhattarai
Lei Bi
Ning Bi
Qi Bi
Yuan Bi
Zhangxing Bian
Benjamin Billot
Manuel Birlo
George Biros
Pinar Bisgin
Koushik Biswas
Sebastian Bodenstedt
Lea Bogensperger
Patrick Bolan
Federico Bolelli
Mahyar Bolhassani
Fabian Bongratz
Ester Bonmati Coll
Sébastien Bougleux
Sylvain Bouix
Emmanuelle Bourigault
Alexandre Bousse
Arnaud Boutillon
Maria Boyko
Afshin Bozorgpour
Nadia Brancati
Kit Bransby
Biagio Brattoli
Diego Bravo
Katharina Breininger
Clara Brémond Martin
Michael Briden
Christopher Bridge
Tom Brosch
Michael Brudno
Michal Brzus
Nhat-Tan Bui
Phuoc-Nguyen Bui
Junghyun Bum
Arianna Bunnell
Stefano Buoso
Ninon Burgos
Nikolay Burlutskiy
Joshua Butke
Yunsu Byeon
Guido Caccianiga
Noa Cahan
Jiangdong Cai
Linghan Cai
Lishan Cai
Ming Cai
Shangyan Cai
Tongan Cai
Zhiyuan Cai
Michael Camilleri
Gabriele Campanella
Victor Campello
Liane Canas
Dongliang Cao
Guogang Cao
Jianguo Cao
Weiguo Cao
Xu Cao
Yankun Cao
Yuzhu Cao
Daniel Capellán-Martín
Gianluca Carloni
Owen Carmichael
Patrick Carnahan
Jacob Carse
Alessandro Casella
Eduardo Castañeda
Philippe C. Cattin
Stefano Cerri
Kate Cevora
Richard Cha
Matthieu Chabanas
Dong-Kyu Chae
Krishna Chaitanya
Souradeep Chakraborty
Tapabrata Chakraborty
Arunava Chakravarty
Kwan Ho Ryan Chan
S.-H. Gary Chan
Yi Hao Chan
Shekhar Chandra
Yao-Jen Chang
Young Hwan Chang
Yuchou Chang
Simon Chatelin
Soumitri Chattopadhyay
Muhammad Faizyab Ali Chaudhary
Shivesh Chaudhary
Joël Chavas
Yiming Che
Antong Chen
Boqi Chen
Chuangquan Chen
Dongdong Chen
Geng Chen
Guannan Chen
Han Chen
Haoyuan Chen
Hongbo Chen

Jiahe Chen
Jianxu Chen
Jiayi Chen
Jiazhou Chen
Jie Chen
Jieneng Chen
Jun Chen
Junxiang Chen
Junzhou Chen
Kai Chen
Kanghao Chen
Kevin Chen
Lei Chen
Li Chen
Lixuan Chen
Pengcheng Chen
Pingjun Chen
Pingyi Chen
Qi Chen
Qiang Chen
Qishen Chen
Shanxiong Chen
Shaobin Chen
Shengchao Chen
Shengjia Chen
Tao Chen
Tingting Chen
Tong Chen
Wanwen Chen
Wenting Chen
Wenxuan Chen
Xiang Chen
Xin Chen
Xinrong Chen
Xiumei Chen
Xueli Chen
Yaxi Chen
Yifei Chen
Yiliang Chen
Yinda Chen
Yinran Chen
Yixiong Chen
Yufei Chen
Yu-Jen Chen
Yunjie Chen
Zhihao Chen
Zhu Chen
Zhuangzhuang Chen
Zifan Chen
Zihao Chen
Zijian Chen
Zijiao Chen
Ziyang Chen
Chang-Chieh Cheng
Chi-Tung Cheng
Jiale Cheng
Jing Cheng
Jun Cheng
Kun Cheng
Pujin Cheng
Teja Krishna Cherukuri
Aichi Chien
Mark Chiew
Philip Chikontwe
Debarghya China
Ching-Hao Chiu
Seungryong Cho
Sue Min Cho
Woo-Jin Cho Kim
Tomáš Chobola
Changyong Choi
Hyoungshin Choi
Kyu Sung Choi
Min-Kook Choi
Sang-Il Choi
Yoonseok Choi
Young Sang Choi
Jakub Chojnacki
Kyobin Choo
Yu-Cheng Chou
Argyrios Christodoulidis
Hung Chu
Lingyang Chu
Miao Chu
Yuetan Chu
Yusung Chu
Kai-Cheng Chuang
Hyungjin Chung
Minyoung Chung
Yunsung Chung
Ozkan Cigdem
Davide Ciraolo
Allison Clement
Bryan Clifford
Robert Cobb
Dana Cobzas
Jaume Coll-Font
Olivier Colliot
Marco Colussi
Cong Cong
Runmin Cong
Laura Connolly
William Consagra
Pierre-Henri Conze
Tim Cootes
Valentina Corbetta
Teresa Correia
Aurélien Corroyer-Dulmont
Baris Coskunuzer
Stephane Cotin
Francesco Cremonesi
Alex Crimi
Ines Alejandro Cruz Guerrero
Hengfei Cui
Jianan Cui
Jiaqi Cui
Xiaoxiao Cui
Xiaoyu Cui
Yue Cui
Zhuo-Xu Cui
Sasa Cukovic
Tolga Cukur
Ariel Curiale
Tobias Czempiel
Duwei Dai
Pengyu Dai
Wei Dai
Sixtus Dakurah
Tingting Dan
Maik Dannecker
Eleonora D'Arnese
Dhritiman Das
Debashis Das Chakladar

Ananyananda Dasari
Aigerim Dautkulova
Palak Dave
Tareen Dawood
Riccardo De Feo
Jonas De Vylder
Bram de Wilde
Meghane Decroocq
Andreas Deistung
Herve Delingette
Liang-Jian Deng
Tianpeng Deng
Yu Deng
Zengtian Deng
Zhipeng Deng
Zhongying Deng
Felix Denzinger
Adrien Depeursinge
Hrishikesh Deshpande
Thomas DeSilvio
Neel Dey
Rohan Raju Dhanakshirur
Maxime Di Folco
Daniel Di Giovanni
Francesco Di Salvo
Songhui Diao
Alina Dima
Hao Ding
Hongxu Ding
Jiaqi Ding
Jun-En Ding
Kexin Ding
Xinpeng Ding
Ying Ding
Zhengyao Ding
Nicola Dinsdale
Nico Disch
Kerol Djoumessi
Konstantin Dmitriev
Daniel Dobkin
Sebastian Doerrich
Melanie Dohmen
Mischa Dombrowski
Ines Domingues
Caixia Dong
Di Dong
Haoyu Dong
Hexin Dong
Mengjin Dong
Min Dong
Shoubin Dong
Siyuan Dong
Xingbo Dong
Zijian Dong
Reuben Dorent
Gianfranco Doretto
Sven Dorkenwald
Haoran Dou
Yimeng Dou
Jason Dowling
Simon Drouin
Karen Drukker
Niharika D'Souza
Chenhe Du
Guodong Du
Honghui Du
Jie Du
Jing Du
Shaoyi Du
Siyi Du
Xia Du
Yuexi Du
Bin Duan
Peiyu Duan
Wenting Duan
Shikha Dubey
Nicolas Duchateau
David Dueñas Gaviria
Benoit Dufumier
Felix Dülmer
Jim Duncan
Joshua Durso-Finley
Pallabi Dutta
Oleh Dzyubachyk
Jan Egger
Gudmundur Einarsson
Favour Ekong
Nafie El Amrani
Mohammed El Hassouni
Ahmed Elazab
Marawan Elbatel
Mohamed El Amine Elforaici
Shireen Elhabian
David Ellis
Mohamed Elsharkawy
Daniel Elson
Solayman Hossain Emon
Sandy Engelhardt
Justin Engelmann
Andinet Enquobahrie
Ertunc Erdil
Marius Erdt
Ayhan Can Erdur
Selman Erel
Meltem Esengönül
Hooman Esfandiari
Nazila Esmaeili
Marco Esposito
Oscar Esteban
Alfonso Estudillo-Romero
Alessio Fagioli
Di Fan
Jianan Fan
Jiansong Fan
Jingfan Fan
Lei Fan
Liang Fan
Xiaoyao Fan
Xin Fan
Huihui Fang
Xi Fang
Yingying Fang
Yuqi Fang
Zhenghan Fang
Zhengqing Fang
Benito Farina
Aiman Farooq
Syed Hasib Akhter Faruqui
Jamil Fayyad
Botond Fazekas
Manman Fei
Paula Feldman
Lina Felsner
Anqi Feng

Mengling Feng
Wei Feng
Weiwei Feng
Zishun Feng
Zunlei Feng
Philipp Feodorovici
Abdesselam Ferdi
André Ferreira
Catalin Fetita
Jean Feydy
Leo Fillioux
Martina Finocchiaro
Alessio Fiorin
Ahmad Firjani
Lukas Fischer
Paul Fischer
Stefan Fischer
Greg Fleishman
Kevin Flores
Kyriakos Flouris
Stefano Fogarollo
Bernhard Föllmer
Nils Daniel Forkert
Jean-Rassaire Fouefack
Kevin François-Bouaou
Jannik Franzen
Moti Freiman
Friedhelm Schwenker
Magda Friedjungová
Paul Friedrich
Yannik Frisch
Jia Fu
Jingru Fu
Qiang Fu
Suzhong Fu
Wei Fu
Xianghua Fu
Xueming Fu
Xueyang Fu
Yu Fu
Yunguan Fu
Moritz Fuchs
Gareth Funka-Lea
Isabel Funke
Pedro Furtado
Ryo Furukawa
Lucas Gago
Mélanie Gaillochet
Gurur Gamgam
Fan Gan
Yu Gan
Melanie Ganz
Bingchen Gao
Chang Gao
Cong Gao
Fei Gao
Linlin Gao
Ruiliang Gao
Shangqi Gao
Shuo Gao
Xiaoru Gao
Xiaoyi Gao
Yibo Gao
Yifan Gao
Yuan Gao
Yue Gao
Zeyu Gao
Zheyao Gao
Sumita Garai
Utpal Garain
Piotr Garbat
Carles Garcia Cabrera
Lin Ge
Yuyan Ge
Zongyuan Ge
Eloy Geenjaar
Manuela Geiss
Elodie Germani
Tijana Geroski
Salah Ghamizi
Negin Ghamsarian
Yazan Gharaibeh
Ghazal Ghazaei
Behnaz Gheflati
Florin Ghesu
Shantanu Ghosh
Yosr Ghozzi
Yongha Gi
Diana Giraldo
Yannik Glaser
Ben Glocker
Bharti Goel
Wesley Gohn
Negar Golestani
Mayank Golhar
Alberto Gomez
Catalina Gomez
Santiago Gómez Hernández
Estibaliz Gómez-de-Mariscal
Mikhail Goncharov
Andrea Gondova
Haifan Gong
Xun Gong
Yongyi Gong
Ricardo Gonzales
Camila Gonzalez
German Gonzalez
Miguel Angel González Ballester
Alvaro Gonzalez-Jimenez
Anisleidy Gonzalez-Mitjans
Sharath Gopal
Vivek Gopalakrishnan
Philippe Gottfrois
Michael Götz
Maged Goubran
Robert Graf
Alejandro Granados
Eric Granger
Vicente Grau
Michael Green
Thomas Greer
Thomas Grenier
Fabian Gröger
Michal Grzeszczyk
Hanxue Gu
Leslie Gu
Pengfei Gu
Ran Gu
Yi Gu
Zaiwang Gu
Francesco Guarnera

Malitha Gunawardhana
Changlu Guo
Fumin Guo
Jintao Guo
Jixiang Guo
Jun Guo
Song Guo
Ting Guo
Xiaojuan Guo
Xueqi Guo
Yi Guo
Yuanbo Guo
Zhanqiang Guo
Gagan Gupta
Danna Gurari
Simon Gutwein
Lisa Guzzi
Sung Min Ha
Benjamin Haeffele
Zahra Hafezi Kafshgari
Paul Hager
Ida Häggström
Justin Haldar
Ghassan Hamarneh
Hua Han
Jinyoung Han
Jiyong Han
John Han
Kangfu Han
Luyi Han
Wenchao Han
Xiangmin Han
Yoseob Han
Lasse Hansen
Degan Hao
Jing Hao
Luoying Hao
Qin Hao
You Hao
Nazim Haouchine
Shota Harada
Shahad Hardan
Michael Hardisty
Mohamed Harmanani
Jeffry Hartanto
Septian Hartono
Md Kamrul Hasan
Ravi Hassanaly
Hideaki Hayashi
Dongdong He
Hongliang He
Huiguang He
Jianzhong He
Mengshen He
Tiantian He
Xiaoxiao He
Xinwei He
Xinzi He
Yifei He
Zhenqi He
Zhibin He
Moein Heidari
Mattias Heinrich
Nicholas Heller
Anja Hennemuth
Pit Henrich
Julius Hense
Hwan Heo
Keun-Soo Heo
Alessa Hering
Monica Hernandez
Daniel Hieber
John Hipwell
David Ho
Edmond S. L. Ho
Matthew Holden
Felix Holm
David Holmes III
Gregory Holste
Barmak Honarvar Shakibaei Asli
Jaeseong Hong
Jaesung Hong
Qingqi Hong
Nicolas Honnorat
Fabian Hörst
Tonmoy Hossain
Golriz Hosseinimanesh
Peyman Hosseinzadeh Kassani
Benjamin Hou
Junlin Hou
Qingshan Hou
Xuege Hou
Franko Hrzic
Dan Hu
Dianlin Hu
Jihong Hu
Kyle Hu
Minhao Hu
Qiang Hu
Qingqiao Hu
Renjiu Hu
Shishuai Hu
Xinrong Hu
Yan Hu
Yang Hu
Zhehao Hu
Shengyi Hua
Mengdi Huai
Baoru Huang
Bin Huang
Cheng Huang
Huifang Huang
Jiahao Huang
Jiashuang Huang
Jiaxing Huang
Kai Huang
Kuan Huang
Liqin Huang
Qinwen Huang
Ruobing Huang
Shuo Huang
Weitian Huang
Wenhui Huang
Wentao Huang
Xinrui Huang
Yanyan Huang
Yi-Jie Huang
Yijin Huang
Yiming Huang
Ying Huang
Yongsong Huang
Yuhao Huang
Yuliang Huang

Zengan Huang
Zhe Huang
Zhen Huang
Zheng Huang
Zhenxing Huang
Arnaud Huaulmé
Alex Hung
Jiayu Huo
Mohammad Arafat Hussain
Alexander Hustinx
Yasin Ibrahim
Ilknur Icke
Tali Ilovitsh
Abdullah-Al-Zubaer Imran
Veera Venkata Raghunath Indugu
Samra Irshad
Aisha Isa
Kh Tohidul Islam
Md Shazid Islam
Mohammad Tariqul Islam
Lukman Ismaila
Koichi Ito
Hayato Itoh
Krithika Iyer
Nabil Jabareen
Mohammad Jafari
Bahram Jafrasteh
Clément Jailin
Amoon Jamzad
Ananya Jana
Ikbeom Jang
Joseph Janizek
Pierre Jannin
Uditha Jarayathne
Ronnachai Jaroensri
Chathurangi Shyalika Jayakody Kankanamalage
Nivetha Jayakumar
Alissa Jell
Alexander Jenke
Mark Jenkinson
Rachid Jennane
Duhee Jeon
Wangsu Jeon
Hwihun Jeong
Ji-Hoon Jeong
Jiwoong Jeong
Won-Ki Jeong
Ranjeet Ranjan Jha
Zexin Ji
Zexuan Ji
Dengqiang Jia
Haozhe Jia
Bailiang Jian
Baichuan Jiang
Hao Jiang
Haotian Jiang
Hongxu Jiang
Juntao Jiang
Ning Jiang
Shancheng Jiang
Songhan Jiang
Xiaoyi Jiang
Xixi Jiang
Yufeng Jiang
Yuncheng Jiang
Zhencun Jiang
Zhongliang Jiang
Yining Jiao
Amelia Jiménez-Sánchez
Qiangguo Jin
Qiuye Jin
Qunchao Jin
Ruinan Jin
Yinzhu Jin
Jie Jing
Yongcheng Jing
Pierre-Marc Jodoin
Craig Jones
Anand Joshi
Shantanu Joshi
Smriti Joshi
Samuel Joutard
Ce Ju
Lie Ju
Vincent Jugnon
Gao Jun
Yohan Jun
Jaehee Jung
Wonsik Jung
Manjunath K. N.
Ali Kafaei Zad Tehrani
Joshua Kaggie
Dagmar Kainmueller
Bernhard Kainz
Mohammad Mahdi Kalantari
Niveditha Kalavakonda
John Kalkhof
Gabriel Kalweit
Maria Kalweit
Tae-Eui Kam
Sreeram Kamabattula
Naoki Kamiya
Michael C. Kampffmeyer
Renuga Kanagavelu
Dongwoo Kang
Junghwa Kang
Ming Kang
Mingon Kang
Qingbo Kang
Sohyun Kang
Wenjie Kang
Arunkumar Kannan
Dmitrii Kaplun
Saarthak Kapse
Siddhartha Kapuria
Tamás Karácsony
Mokshagna Sai Teja Karanam
Neerav Karani
Mert Asim Karaoglu
Turkay Kart
Tushar Kataria
Eytan Kats
Alexander Katzmann
Gurwinder Kaur
Matthias Keicher
Elif Keles
Erwan Kerrien
Ali Keshavarzi
Sara Ketabi

Mohsen Khadem
Firas Khader
Abbas Khan
Hassan Khan
Izhar Ahmed Khan
Ufaq Khan
Bidur Khanal
Bishesh Khanal
Benjamin Killeen
Daeseung Kim
Daewoon Kim
Dohoon Kim
Dohyun Kim
Hyeongyu Kim
Hyun Gi Kim
Hyun-Woo Kim
Jeongchul Kim
Jin Kim
Jinhee Kim
Jonghun Kim
JongWoo Kim
Kiduk Kim
Kyungsu Kim
Mansu Kim
MinWoo Kim
Sangwook Kim
Seong Tae Kim
Soopil Kim
Suhyeon Kim
Tae Soo Kim
Young-Min Kim
Yunsoo Kim
Alvin Kimbowa
Benjamin Kimia
Andrew King
Atilla Kiraly
Ayesha Kiran
Yannick Kirchhoff
Gabriel Kiss
Andreas Kist
Itaru Kitahara
Yoshiro Kitamura
Jens Kleesiek
Stefan Klein
Tobias Klinder
Kazuma Kobayashi
Lisa Koch
Yeong Jun Koh
Taha Koleilat
Anudeep Konda
Satoshi Kondo
Ming Kong
Ender Konukoglu
Aishik Konwer
Nicholas Konz
Mohamad Koohi-Moghadam
Steven Korevaar
Kivanc Kose
Sebastian Kozerke
Anna Kreshuk
Axel Krieger
Murali Krishnamoorthy
AnithaPriya Krishnan
Florian Kromp
Frithjof Kruggel
Henry Krumb
Hulin Kuang
Nanyu Kuang
David Kügler
Aaron Kujawa
Pranav Kulkarni
Amar Kumar
Ashwin Kumar
Nishant Kumar
Logiraj Kumaralingam
Kassymzhomart Kunanbayev
Soumya Kundu
Manuela Kunz
Po-Chih Kuo
Jan Kybic
Charlotte Kyeremah
Aymen Laadhari
Dmitrii Lachinov
Chaymaa Lahmar
Bolin Lai
Shang-Hong Lai
Songning Lai
Yingxin Lai
Brejesh Lall
Bennett Landman
Daniel Lang
Frank Langbein
Georg Langs
Qicheng Lao
Othmane Laousy
Catherine Laporte
Carole Lartizien
Shlomi Laufer
Max-Heinrich Laves
Aonghus Lawlor
Ana Lawry Aguila
An Le
Lawrence Le
Loic Le Folgoc
Bumshik Lee
Dong-Gyu Lee
Eungjoo Lee
Hong Joo Lee
Hyunyeol Lee
Jangho Lee
Jemyoung Lee
Jong-Hwan Lee
Jong-Min Lee
Jungwook Lee
Kyungsu Lee
Moon Hwan Lee
Sangmin Lee
Sangseon Lee
Seoyoung Lee
Seungeun Lee
Seungjun Lee
Sing Chun Lee
Soochahn Lee
Tiarna Lee
Won Hee Lee
Woonghee Lee
Youngjoon Lee
Antoine Legouhy
Long Lei
Minglong Lei
Wen Hui Lei
Yiming Lei
Rogers Jeffrey Leo John

Riccardo Levi
Michael Levine
Caizi Li
Chao Li
Chen Li
Cheng Li
Chengyin Li
Cynthia Xinran Li
Dongyang Li
Fan Li
Fangda Li
Ganping Li
Guangyao Li
Guoshi Li
Haiqing Li
Haitao Li
Han Li
Hanyu Li
Hao Li
Haofeng Li
Haoshen Li
Haoyang Li
Heng Li
Hongming Li
Hongyu Li
Huiqi Li
Jiawen Li
Jingming Li
Jingxiong Li
Jinpeng Li
Jun Li
Juncheng Li
Junjian Li
Jupeng Li
Kang Li
Lei Li
Lin Li
Miao Li
Qi Li
Qian Li
Qiyue Li
Rui Li
Ruizhe Li
Shiwei Li
Shuaibo Li
Shulong Li
Shumeng Li
Shuo Li
Siqi Li
Tiancheng Li
Wen Li
Wenxue Li
Xiangyu Li
Xiao-Xin Li
Xiaoxu Li
Xingjian Li
Xingyu Li
Xinyue Li
Xuesong Li
Yamin Li
Yang Li
Ye Li
Yi Li
Yicong Li
Yingshu Li
Yingtai Li
Yitong Li
Yuexiang Li
Yuheng Li
Zeju Li
Zhe Li
Zhiyuan Li
Zhjin Li
Zhongyu Li
Zhuoan Li
Zi Li
Zihan Li
Chunfeng Lian
Sheng Lian
Dong Liang
Libin Liang
Ruixing Liang
Yaqian Liang
Ziyun Liang
Guiqiu Liao
Hongen Liao
Junda Liao
Wei Liao
Zehui Liao
Marta Ligero
Chern Hong Lim
Gilbert Lim
Ailiang Lin
Chieh-Te Lin
Fangzhou Lin
Huei-Yung Lin
Jianxin Lin
Li Lin
Manxi Lin
Mingquan Lin
Runjia Lin
Xian Lin
Yi Lin
Yiqun Lin
Yiyang Lin
Zehui Lin
Claudia Lindner
Haibin Ling
Xitong Ling
Anran Liu
Baodi Liu
Bo Liu
Chang Liu
Che Liu
Chen Liu
Chenyu Liu
Chi Liu
Chuanbin Liu
Dichao Liu
Fei Liu
Feihong Liu
Guiran Liu
Guoyang Liu
Han Liu
Hao Liu
Hengjie Liu
Huafeng Liu
Huaxiang Liu
Jianan Liu
Jianming Liu
Jiawei Liu
Jia-Wei Liu
Jiaxiang Liu
Jie Liu
Jinduo Liu

Jingsong Liu
Jingxin Liu
Jinhua Liu
Jinyang Liu
Juan Liu
Jun Liu
Junyi Liu
Kang Liu
Meng Liu
Mengting Liu
Mianxin Liu
Mingxin Liu
Mingxuan Liu
Mingyuan Liu
Peng Liu
Shengyuan Liu
Shuting Liu
Sidong Liu
Tiange Liu
Ting Liu
Tingwei Liu
Wei Liu
Wenjie Liu
Xianyuan Liu
Xinyu Liu
Yan Liu
Yang Liu
Yanhong Liu
Yanzhen Liu
Yi Liu
Yifan Liu
Yihao Liu
Yikang Liu
Yue Liu
Yunyi Liu
Yusheng Liu
Yuxiao Liu
Zelong Liu
Zeyu Liu
Zhentao Liu
Zhi Liu
Ziteng Liu
Zixuan Liu
Lisette Lockhart
Andrea Loddo
Herve Lombaert
Nicolas Loménie
Yonghao Long
Giovanni Lonia
Miguel López-Pérez
Marco Lorenzi
Ange Lou
Meng Lou
Wei Lou
Chenggang Lu
Chun-Shien Lu
Guangming Lu
Jiaxuan Lu
Jingfeng Lu
Mengkang Lu
Shengnan Lu
Shuai Lu
Xing Lu
Yao Lu
Yu Lu
Yuhuan Lu
Ziru Lu
Xiao Luan
Ruben Lucassen
Miguel Luna
Guibo Luo
Haozhe Luo
Huoling Luo
Jiaqi Luo
Ma Luo
Mingyuan Luo
Xiangde Luo
Xiaoling Luo
Xinzhe Luo
Yong Luo
Zhiming Luo
Ngoc Ly
Ilwoo Lyu
Mengye Lyu
Qing Lyu
Yiheng Lyu
Deeksha M. Shama
Benteng Ma
Chenglong Ma
Da Ma
DongAo Ma
Hehuan Ma
Jiquan Ma
Jun Ma
Lei Ma
Ming Ma
Qiang Ma
Xinke Ma
Yingliang Ma
Yuanzheng Ma
Yuhui Ma
Zhuoqi Ma
Diego Machado Reyes
Pedro Macias Gordaliza
Anderson Maciel
Derek Magee
Dwarikanath Mahapatra
S. Sara Mahdavi
Ahsan Mahmood
Mohammed Mahmoud
Huayu Mai
Jennifer Maier
Klaus H. Maier-Hein
Lena Maier-Hein
Abrar Majeedi
Roman Malashin
Hashmat Shadab Malik
Mayur Mallya
Norberto Malpica
Constantina Maltezou-Papastylianou
Jan Emily Mangulabnan
Arooba Maqsood
Radu Marculescu
Jan Margeta
Umar Marikkar
Zdravko Marinov
Iman Marivani
Pablo Márquez Neila
Carsten Marr
Robert Marti
Nicolas Martin
Josué Martínez-Martínez
Queralt Martín-Saladich
Najmeh Mashhadi

Martin Maška
Tejas Sudharshan Mathai
Sasan Matinfar
Shinnosuke Matsuo
Hiroki Matsuzaki
Thomas Paul Matthews
Petr Matula
Satyajit Maurya
Dimitrios Mavroeidis
Voichita Maxim
Arnaldo Mayer
Perla Mayo
Evangelos Mazomenos
Badhan Mazumder
Ashery Mbilinyi
Jamie McClelland
Julian McGinnis
Pau Medrano-Gracia
Mostafa Mehdipour Ghazi
Deval Mehta
Raghav Mehta
Liye Mei
Felix Meister
Alaa Melek
Steve Mendoza
Mingyuan Meng
Runqi Meng
Yan Meng
Vivek Menon
Martin Menten
Talha Meraj
Dorit Merhof
Odyssée Merveille
Adrien Meyer
Leo Milecki
Wenwen Min
Zhe Min
Duy Minh Ho Nguyen
Golrokh Mirzaei
Nahal Mirzaie
Virendra Mishra
Ashish Mittal
Tadashi Miyamoto
Pooya Mobadersany
Sovesh Mohapatra
Ramtin Mojtahedi
Tony C. W. Mok
Javier A. Montoya-Zegarra
Mehdi Moradi
José Morano
Kensaku Mori
Shohei Mori
Elena Morotti
Lia Morra
Luca Mossina
Ana M. Mota
Lei Mou
Harold Mouchère
Kaouther Mouheb
Seyed Amir Mousavi
Daniel Moyer
Pritam Mukherjee
Anirban Mukhopadhyay
Johanna Müller
Philip Müller
Sarah Müller
Asbjørn Munk
Aditya Murali
Vittorio Murino
Saif Musluh
Yasmine Mustafa
Inye Na
Shahabedin Nabavi
Sheida Nabavi
Roman Naeem
Manas Nag
Enamundram Naga Karthik
Noor Nakhaei
Sahil Nalawade
Jakub Nalepa
Khashayar Namdar
Yang Nan
Nandakishor Nandakishor
Ahmed Nebli
Fnu Neha
Debanga Raj Neog
Favour Nerrise
Gia Ngo
Binh Nguyen
Hang Nguyen
Huy Nguyen
Khoa Nguyen
Tan Nguyen
Phi Nguyễn Văn
Duong Nhu
Dong Ni
Haomiao Ni
Brennan Nichyporuk
Hannes Nickisch
Xinyu Nie
Joshua Niemeijer
Aditya Nigam
Lipeng Ning
Xiaolin Ning
Hareem Nisar
Kazuya Nishimura
Sijie Niu
Usma Niyaz
Fuad Noman
Mehrdad Noori
Chinedu Nwoye
Emeka Obi
Gilberto Ochoa-Ruiz
Marzieh Oghbaie
Baris Oguz
Hyun-Jic Oh
Esther Okafor
Ajibola Oladokun
Laura Oliva Maza
Arlindo Oliveira
Bruno Miguel Gomes Oliveira
Antoine Olivier
Juan Olmos
Jimena Olveres
Yu Hon On
Doruk Oner
John Onofrey
Hiroko Oshibe
Andreas Østvik
Chubin Ou
Yafei Ou
Muhammad Owais
Utku Ozbulak

Caner Ozer
Ege Özsoy
Kutsev Ozyoruk
Martina Paccini
Nicolas Padoy
José Blas Pagador Carrasco
Suraj Pai
Daniel Pak
Ashkan Pakzad
Basudha Pal
Christoph Palm
Delin Pan
Hongyi Pan
Jiazhen Pan
Liangrui Pan
Wentao Pan
Xiang Pan
Yimu Pan
Zhifang Pan
Fengqian Pang
Haowen Pang
Jiaxuan Pang
Yan Pang
Hsing-Kuo Pao
Tasos Papastylianou
Bartlomiej Papiez
Jay Paranjape
Abhijeet Parida
Hyunjin Park
Joon Sik Park
Robin Park
Sang Hyun Park
Sung-Hong Park
Yongkeun Park
Mario Parreno-Centeno
Srivathsa Pasumarthi
Ashay Patel
Sushant Patkar
Mayank Patwari
Angshuman Paul
Riti Paul
Rasmus Paulsen
Thierry Pécot
João Pedrosa
Paraskevas Pegios
Jialun Pei
Yuchen Pei
Yuru Pei
Himashi Peiris
Chantal Pellegrini
Danni Peng
Linkai Peng
Yige Peng
Yunsong Peng
Zelin Peng
Matteo Pennisi
Tobias Penzkofer
Jacob Peoples
Sergio Pereira
Alejandra Perez
Fernando Pérez-García
Matthias Perkonigg
Denis Peruzzo
Eike Petersen
Jens Petersen
Simon Pezold
Elisabeth Pfaehler
Ashish Phophalia
Thomas Pinetz
Sergey Plis
Sergei Pnev
Gasper Podobnik
Kilian Pohl
Isabella Poles
Dirk H. J. Poot
Mirela Popa
Peter Pott
Alison Pouch
Pranav Poudel
Chinmay Prabhakar
Ferran Prados
Manish Prajapati
Rishav Pramanik
Raphael Prevost
Federica Proietto Salanitri
Sergey Protserov
Lemuel Puglisi
Sergi Pujades
Behrus Puladi
Lorenzo Putzu
Elodie Puybareau
Kehan Qi
Liangce Qi
Lin Qi
Yaolei Qi
Buyue Qian
Yan Qiang
Mengyun Qiao
Yuchuan Qiao
Fangbo Qin
Jiahao Qin
Yiming Qin
Zengyi Qin
Huaqi Qiu
Jie Qiu
Jingna Qiu
Peijie Qiu
Shihan Qiu
Linhao Qu
Tingwei Quan
Sandro Queiros
Sandro Queirós
Rizwan Qureshi
Prashanth R.
Daniel Racoceanu
Mehdi Rahim
Md Masudur Rahman
Md Mostafijur Rahman
Ankita Raj
Ajitha Rajan
Jagath Rajapakse
João Ramalhinho
Sirisha Rambhatla
Jayroop Ramesh
Xuming Ran
Amin Ranem
Bryan Ranger
Aneesh Rangnekar
Carolina Raposo
Hatem Rashwan
Sebastian Rassmann
Aditya Rastogi
Daniele Ravì
Sadhana Ravikumar

Vikram Ravindra
Sudarshan Regmi
Simon Reiß
Samuel Remedios
Qin Ren
Zhihang Ren
Mauricio Reyes
Abel Reyes-Angulo
Hadrien Reynaud
Helena Ribeiro Torres
Daniel Riccio
Anne-Marie Rickmann
Bastian Rieck
Brandalyn Riedel
Tammy Riklin Raviv
Morgan Ringel
Harry Robertshaw
Lucas Robinet
Alfie Roddan
Jessica Rodgers
Ranga Rodrigo
Javier Rodriguez-Puigvert
Robert Rohling
Melanie Roschewitz
Pooneh Roshanitabrizi
Florence Rossant
Alberto Rota
Ayush Roy
Kaushiki Roy
Saikat Roy
Dan Ruan
Jiacheng Ruan
Daniel Rueckert
Daniela Ruiz
Dewinda Rumala
Josef Lorenz Rumberger
Mirabela Rusu
Patryk Rygiel
Kanghyun Ryu
Muralikrishna S. N.
Amirhossein Sabour
Maziar Sabouri
Ario Sadafi
Numan Saeed
Farhang Sahba
Heba Sailem
Hideo Saito
Simone Saitta
Gulfam Saju
Md Sirajus Salekin
Mostafa Salem
Benjamin Salmon
Thomas Sanchez
Yudi Sang
Sara Sangalli
Sivaramakrishnan Sankarapandian
Alice Santilli
Bikash Santra
Nishchal Sapkota
Duygu Sarikaya
Gary Sarwin
Olivier Saut
Ricky Savjani
Cian Scannell
Yoni Schirris
Alexander Schlaefer
Maja Schlereth
Adam Schmidt
Linda-Sophie Schneider
Steffen Schneider
Daniel Scholz
Thomas Schultz
Michael Schwier
Michaël Sdika
Eric Seibel
Matthias Seibold
Jothiraj Selvaraj
Sourya Sengupta
Thilo Sentker
Hyewon Seo
Sumin Seo
Nazmus Shakib Shadin
Ahmed Shaffie
Hina Shah
Keyur Devendra Shah
Nisarg Shah
Nagur Shareef Shaik
S. Shailja
Fereshteh Shakeri
Kun Shang
Wei Shao
Maxim Sharaev
Mostafa Sharifzadeh
Saurav Sharma
Vanshali Sharma
Vatsala Sharma
Bingzhi Shen
Hualei Shen
Linlin Shen
Xiongri Shen
Yiqing Shen
Zhenrong Shen
Zhiqiang Shen
Victor S. Sheng
Enze Shi
Fashun Shi
Ji Shi
Jian Shi
Jun Shi
Junjie Shi
Kangdi Shi
Lulin Shi
Pengcheng Shi
Ruohua Shi
Xiaoshuang Shi
Yaying Shi
Yi Shi
Yinghuan Shi
Yonggang Shi
Zhan Shi
Shu-Fu Shih
Kaito Shiku
Jitae Shin
Yejee Shin
Isaac Shiri
Boris Shirokikh
Hongchao Shu
Yucheng Shu
Zhongyi Shui
Amita Shukla-Dave
Weixin Si
Vasiliki Sideri-Lampretsa
Alberto Signoroni
Monica Sigovan

Saurabh Sihag
Wilson Silva
Julio Silva-Rodríguez
Margarida Silveira
Marco Simoes
Ivor Simpson
Malay Singh
Vivek Singh
Abhishek Singh Sambyal
Rohit Singla
Ashish Sinha
Ayushi Sinha
Vasilis Siomos
Raviteja Sista
Elena Sizikova
Luka Škrlj
Gregory Slabaugh
Paddy Slator
Dane Smith
Kevin Smith
Tiffany So
Roger Soberanis-Mukul
Elanchezhian Somasundaram
Young-Han Son
Sifan Song
Tao Song
Xinrui Song
Tananun Songdechakraiwut
Kristine Sørensen
Moein Sorkhei
Aristeidis Sotiras
Mazen Soufi
Mayssa Soussia
Raissa Souza
Roberto Souza
Georgia Sovatzidi
Michal Španěl
Rachel Sparks
Tyler Spears
William Speier
Ziga Spiclin
Jon Sporring
Arvind Krishna Sridhar
Lawrence Staib
Marc Stamminger
Emma Stanley
Johannes Stegmaier
Jan Steinbrener
Nil Stolt-Ansó
Skylar Stolte
Robin Strand
Moritz Sturm
Martin Styner
Weifeng Su
Yao Su
Zhe Su
Julian Suk
Emese Sükei
Chao Sun
Hongfu Sun
Jianqi Sun
Kaicong Sun
Liyan Sun
Poly Z. H. Sun
Susu Sun
Xiaowu Sun
Xiaoxiao Sun
Yan Sun
Yihua Sun
Yiran Sun
Zhaobin Sun
Zhe Sun
Zhichao Sun
Vaanathi Sundaresan
Wonmo Sung
Mohammod Naimul Islam Suvon
Michele Svanera
Numi Sveinsson Cepero
Tomasz Szczepański
Behrad Taghibeyglou
Pablo Tahoces
Divyanshu Tak
Shumpei Takezaki
Estefania Talavera
Jeremy Tan
Jing Wei Tan
Shan Tan
Tao Tan
Wenjun Tan
Zhiwei Tan
Feilong Tang
Fenghe Tang
Haoteng Tang
He Tang
Jiankai Tang
Xiaoqin Tang
YuCheng Tang
Zhenyu Tang
Zhi-Ri Tang
Zihao Tang
Qian Tao
Tianli Tao
Austin Tapp
Elias Tappeiner
Mickael Tardy
Giacomo Tarroni
Wallapak Tavanapong
Thomas Tavolara
Jacopo Teneggi
Lin Teng
Paul Terrassin
Kaveri Thakoor
Franz Thaler
Antoine Theberge
Belvin Thomas
Sarina Thomas
Elinor Thompson
Stephen Thompson
Felix Thomsen
Alexander Thorley
Jacob Thrasher
Xin Tian
Xuanyu Tian
Yapeng Tian
Yun Tian
Matthew Tivnan
Minh Nguyen Nhat To
Santiago Toledo-Cortés
Lyuyang Tong
Marwan Torki
Nicolas Toussaint
Henry Tregidgo

Jakob Troidl
Yen Nhi Truong Vu
Apostolia Tsirikoglou
Mingxiao Tu
Oezguen Turgut
Aayush Tyagi
Georgios Tziritas
Kwang-Hyun Uhm
Mathias Unberath
Balagopal Unnikrishnan
Hristina Uzunova
Sami Väänänen
Valentina Vadori
Theodoros Vagenas
Anurag Vaidya
Maria Vakalopoulou
Vanya Valindria
Juan Miguel Valverde
Fons van der Sommen
Louis van Harten
Tom van Sonsbeek
Gijs van Tulder
Bino Varghese
Francisco Vasconcelos
Hooman Vaseli
Rajat Vashistha
S. Swaroop Vedula
Matthew Velazquez
Vikram Venkatraghavan
Mitko Veta
Vibujithan Vigneshwaran
Veronica Vilaplana
Pierre-Frederic Villard
Thanaporn Viriyasaranon
Christiaan Viviers
Hung Vo
Wolf-Dieter Vogl
Emiliano Votta
Hung Vu
Milos Vukadinovic
Trinh Thi Le Vuong
Christian Wachinger
Martin Wagner
Jia Wan
Liang Wan
Shouhong Wan
Xiaohua Wan
An Wang
Bin Wang
Changmiao Wang
Chen Wang
Cheng Wang
Chenhui Wang
Chong Wang
Congcong Wang
Diwei Wang
Dongang Wang
Fan Wang
Fu Wang
Fuying Wang
Haifeng Wang
Haolin Wang
Haoran Wang
Hongxiao Wang
Hongyi Wang
Hongyu Wang
Hongzhi Wang
Hualiang Wang
Huayu Wang
Janet Wang
Jiacheng Wang
Jiamu Wang
Jian Wang
Jiaqi Wang
Jiazhen Wang
Jinfeng Wang
Jixin Wang
Jiyao Wang
Jueqi Wang
Jui-Kai Wang
Jun Wang
Junchen Wang
Lehan Wang
Lei Wang
Lin Wang
Linwei Wang
Lyuyang Wang
Manning Wang
Meng Wang
Nan Wang
Peiqi Wang
Qi Wang
Qiuli Wang
Rongfang Wang
Rui Wang
Ruixuan Wang
Runmin Wang
Sen Wang
Shiyi Wang
Shuxian Wang
Tao Wang
Tianyu Wang
Weiming Wang
Xi Wang
Xiaofei Wang
Xin Wang
Xingce Wang
Xinliang Wang
Xu Wang
Yan Wang
Yi Wang
Yifan Wang
Yifeng Wang
Yinsong Wang
Yipei Wang
Yiqing Wang
Yixin Wang
Yiyang Wang
Yiying Wang
Yonghao Wang
Yuheng Wang
Yulin Wang
Yun Wang
Yuxi Wang
Zeyi Wang
Zhaoyang Wang
Zhewei Wang
Zhong Wang
Zhu Wang
Zichen Wang
Ziyi Wang
Ziyue Wang
Zuhui Wang
Valentine Wargnier Dauchelle

Mohammad Wasih
Dong Wei
Guangshun Wei
Jia Wei
Jie Wei
Meng Wei
Ruofeng Wei
Shuwen Wei
Wen Wei
Jonas Weidner
Wolfgang Wein
Oliver Weinheimer
Xuyun Wen
Zheyu Wen
Thomas Wendler
Nina Weng
Wenhai Weng
Markus Wenzel
Julia Werner
Bryan Williams
Brady Williamson
Paul Wilson
Alexander Winkler
Eric Wisotzky
Adam Wittek
Marek Wodzinski
Ivo Wolf
Georg Wölflein
Julia Wolleb
Dong-Ok Won
Ka-Chun Wong
Ken C. L. Wong
McKell Woodland
Stefan Wörz
Binghong Wu
Chengchuan Wu
Chenwei Wu
Fuping Wu
Fuzhi Wu
Ji Wu
Jianghao Wu
Jie Ying Wu
Junde Wu
Junyang Wu
Linshan Wu
Mengqi Wu
Qian Wu
Qianyu Wu
Qing Wu
Ruoyou Wu
Weiyi Wu
Wenbin Wu
Wenxuan Wu
Xian Wu
Xiao Wu
Xinyao Wu
Yihang Wu
Yinhao Wu
Yinzhe Wu
Yunheng Wu
Zhan Wu
Zhengwang Wu
Zhiying Wu
Zihao Wu
Zijian Wu
Julian Wyatt
Magdalena Wysocki
Pengcheng Xi
Chao Xia
Jing Xia
Jun Xia
Peng Xia
Tian Xia
Weihao Xia
Yihao Xia
Jinhai Xiang
Lei Xiang
Nan Xiang
Bin Xiao
Deqiang Xiao
Li Xiao
Qing Xiao
Sa Xiao
Xiaojiao Xiao
Bin Xie
Chi Xie
Hao Xie
Jianyang Xie
Kangxian Xie
Kunzi Xie
Long Xie
Qihang Xie
Xingyu Xie
Yiting Xie
Bowen Xin
Fangxu Xing
Jiarui Xing
Shuwei Xing
Zhaohu Xing
Conghao Xiong
Honglin Xiong
Juming Xiong
Yiheng Xiong
Yujian Xiong
Chi Xu
Di Xu
Dongrong Xu
Fangqiang Xu
Feng Xu
Gelei Xu
Geng-Xin Xu
Hongming Xu
Hongyan Xu
Huihui Xu
Jialang Xu
Jiangchang Xu
Jiaqi Xu
Jiaxing Xu
Jilan Xu
Meilong Xu
Min Xu
Moucheng Xu
Rui Xu
Senyan Xu
Weixin Xu
Wentian Xu
Xiao Xu
Xiaowei Xu
Xiuyuan Xu
Yongchao Xu
Yunqiu Xu
Zhoubing Xu
Ziang Xu
Zikang Xu
Ziyue Xu

Zunnan Xu
Cheng Xue
Jie Xue
Yuyang Xue
Zhong Xue
Fang Yan
Ke Yan
Kun Yan
Rui Yan
Siyuan Yan
Wen Yan
Xiangyi Yan
Yige Yan
Yuguang Yan
Yunlu Yan
Zipei Yan
Baoyao Yang
Carl Yang
Changchun Yang
Chen Yang
Gang Yang
Guanyu Yang
Heran Yang
Huijuan Yang
Jianhua Yang
Jiaqi Yang
Jie Yang
Junlin Yang
Lin Yang
Peng Yang
Qianye Yang
Sejung Yang
Shu Yang
Shuojue Yang
Su Yang
Xiaochen Yang
Xiaofei Yang
Xikai Yang
Xinyi Yang
Xuan Yang
Yan Yang
Yang Yang
Yijun Yang
Yingyu Yang
Zefan Yang
Zhicheng Yang
Zhikai Yang
Zhiwei Yang
Zhiwen Yang
Zixin Yang
Ziyuan Yang
Jiachen Yao
Linlin Yao
Tianyuan Yao
Xiaohui Yao
Xing Yao
Yongcheng Yao
Zhao Yao
Choon Hwai Yap
Ziya Ata Yazıcı
Chenfei Ye
Jian Ye
Jiayu Ye
Johan Ye
Kai Ye
Shuchang Ye
Xujiong Ye
Yiwen Ye
Chun-Hsiao Yeh
Srikar Yellapragada
Si Yong Yeo
Joe Yeong
Pak Hei Yeung
Murong Yi
Weixi Yi
Yugen Yi
Hanlong Yin
Yunting Yin
Zijin Yin
Haochao Ying
Wenhao Ying
Tatsuya Yokota
Jaejun Yoo
Seok Bong Yoo
Youngjin Yoo
Jee Seok Yoon
Jihun Yoon
Kyungho Yoon
Siyeop Yoon
Xin You
Alexandra Young
Boxiao Yu
Fanyang Yu
Hanyi Yu
Jianxun Yu
Jinhua Yu
Minhui Yu
Qinji Yu
Renping Yu
Sigang Yu
Weihao Yu
Weimin Yu
Yang Yu
Zhen Yu
Ziqi Yu
Cheng Yuan
Kun Yuan
Mingze Yuan
Xiaohan Yuan
Hailin Yue
Jiaxin Yue
Xiaodong Yue
Yutao Yue
Paul Yushkevich
Anna Zapaishchykova
Masoud Zarepisheh
Kamilia Zaripova
Tal Zeevi
Rachid Zeghlache
Ramy Zeineldin
Bolun Zeng
Dewen Zeng
Dong Zeng
Hongye Zeng
Qingjie Zeng
Tianyi Zeng
Yuwen Zeng
Hasnae Zerouaoui
Oliver Zettinig
Ruyi Zha
Baochang Zhang
Bodong Zhang
Chenxi Zhang
Chuyan Zhang
Daoqiang Zhang

Dong Zhang
Fan Zhang
Guoning Zhang
Hanxiao Zhang
Hanyun Zhang
Haochen Zhang
Haoyue Zhang
Hongliang Zhang
Jianhai Zhang
Jianjia Zhang
Jianping Zhang
Jiansong Zhang
Jianwei Zhang
Jiaqi Zhang
Jiaqing Zhang
Jiazhen Zhang
Jingyang Zhang
Jinwei Zhang
Kaiyu Zhang
Ke Zhang
Le Zhang
Lei Zhang
Li Zhang
Lu Zhang
Meng ze Zhang
Miaomiao Zhang
Minghui Zhang
Minqing Zhang
Molin Zhang
Pingping Zhang
Rongzhao Zhang
Ruisi Zhang
Shu Zhang
Shuai Zhang
Shuaitong Zhang
Song Zhang
Tian Tian Zhang
Tianjing Zhang
Tianyi Zhang
Tuo Zhang
Wei Zhang
Weihang Zhang
Weiyi Zhang
Wenhua Zhang
Xiaoming Zhang
Xiaoran Zhang
Xiaoshan Zhang
Xiaotong Zhang
Xin Zhang
Xuhong Zhang
Xukun Zhang
Xuzhe Zhang
Yanfu Zhang
Yanteng Zhang
Yi Zhang
Yichi Zhang
Yike Zhang
Yilan Zhang
Yizhe Zhang
You Zhang
Yuan Zhang
Yuanwang Zhang
Yue Zhang
Yuhan Zhang
Yundi Zhang
Yundong Zhang
Yuqi Zhang
Yuxin Zhang
Zeyu Zhang
Zheng Zhang
Zhenxuan Zhang
Zheyu Zhang
Zhicheng Zhang
An Zhao
Baoliang Zhao
Beidi Zhao
Chen Zhao
Cheng Zhao
Chongyue Zhao
Haoyu Zhao
He Zhao
Jichao Zhao
Jun Zhao
Kun Zhao
Mingyang Zhao
Puyang Zhao
Qingyu Zhao
Shen Zhao
Shensheng Zhao
Weibing Zhao
Weiqin Zhao
Weiren Zhao
Wenzhao Zhao
Xiangyu Zhao
Xingyue Zhao
Xinkai Zhao
Yidong Zhao
Yizhou Zhao
Yue Zhao
Zhihao Zhao
Zihao Zhao
Ziheng Zhao
Ziwei Zhao
Zixu Zhao
Xiantong Zhen
Boyun Zheng
Fudan Zheng
Hao Zheng
Haoxin Zheng
Qiaoyu Zheng
Qingqing Zheng
Shaoming Zheng
Shuai Zheng
Tianshu Zheng
Weihao Zheng
Yalin Zheng
Yefeng Zheng
Yijiang Zheng
Yushan Zheng
Zhou Zheng
Jiayang Zhong
Lanfeng Zhong
Yi Zhong
Yuan Zhong
Binggui Zhou
Congcong Zhou
Guangquan Zhou
Haoyin Zhou
Jun Zhou
Li Zhou
Meng Zhou
Qian Zhou
Rong Zhou
Shuo Zhou
Sijin Zhou

Wei Zhou
Wu Zhou
Xiaogen Zhou
Xinrui Zhou
Yaxuan Zhou
Yi Zhou
Yuan Zhou
Yuanpin Zhou
Yukun Zhou
Yuyue Zhou
Zhen Zhou
Zhengbo Zhou
Zongwei Zhou
Hancan Zhu
Jiayuan Zhu
Junchao Zhu
Lei Zhu
Lingting Zhu
Meilu Zhu
Qikui Zhu
Yanjie Zhu
Yazhou Zhu
Yitao Zhu
Yuemin Zhu
Zhanshi Zhu
Zhenyang Zhu
Jia-Xin Zhuang
Jun Zhuang
Yupeng Zhuo
Arif Hassan Zidan
Alexander Ziller
Veronika Zimmer
David Zimmerer
Fangrong Zong
Jing Zou
Xinrui Zou
Jian Zu
Wenqiang Zu
Maria A. Zuluaga
Lianrui Zuo
Gerald Zwettler

Contents – Part III

Learning and Augmented Reality for Surgical and Endoscopic Applications (I)

4D CardioSynth: Synthesising Dynamic Virtual Heart Populations Through Spatiotemporal Disentanglement 3
Haoran Dou, Jinghan Huang, Arezoo Zakeri, Zherui Zhou, Tingting Mu, Jinming Duan, and Alejandro F. Frangi

A Unified Continuous Staging Framework for Alzheimer's Disease and Lewy Body Dementia via Hierarchical Anatomical Features 13
Tong Chen, Minheng Chen, Jing Zhang, Yan Zhuang, Chao Cao, Xiaowei Yu, Yanjun Lyu, Lu Zhang, Li Su, Tianming Liu, and Dajiang Zhu

Accelerated Free-Breathing 5D Multi-Echo Respiratory Motion-Resolved R2*, PDFF, and QSM Using Novel Composite Total Variation 24
MungSoo Kang, Or Alus, and Youngwook Kee

AffinityUMamba: Uncertainty-Aware Medical Image Segmentation via Probabilistic Weak Supervision Beyond Gold-Standard Annotations 35
Yukun Zhang, Guisheng Wang, William Henry Nailon, and Kun Cheng

Anomaly Detection by Clustering DINO Embeddings Using a Dirichlet Process Mixture 46
Nico Schulthess and Ender Konukoglu

Aorta Multi-class Segmentation via Anatomically Constrained Plane Detection 57
Jonghoon An, Dong Hyun Lee, So Hyun Kim, Taejin Moon, and Minyoung Chung

Automated Characterization of Myocardial Scar Topological Patterns for Ventricular Tachycardia Screening 67
Xicheng Sheng, Yang Zhang, Lei Li, Bailiang Chen, Freddy Odille, and Xiahai Zhuang

Blind Restoration of High-Resolution Ultrasound Video 77
Chu Chen, Kangning Cui, Pasquale Cascarano, Wei Tang, Elena Loli Piccolomini, and Raymond H. Chan

Bridging Classical and Learning-Based Iterative Registration Through Deep Equilibrium Models 88
Yi Zhang, Yidong Zhao, and Qian Tao

Cascaded 3D Diffusion Models for Whole-Body 3D 18-F FDG PET/CT Synthesis from Demographics 99
Siyeop Yoon, Sifan Song, Pengfei Jin, Matthew Tivnan, Yujin Oh, Sekeun Kim, Dufan Wu, Xiang Li, and Quanzheng Li

Causality-Driven Spatio-Temporal Generator for Multi-phase Contrast-Enhanced CT Synthesis 110
Qikui Zhu, Hao Wu, Yanyan Zhang, and Shuo Li

Contrastive Disentanglement Learning Framework for Multi-lead Wearable ECG Denoising 120
Yue Zhang, Chenyu Zhao, Wen Zhang, Jinliang Wang, Jun Guo, Wei Yang, and Qianjin Feng

CS^{2}C: Collaborative Spatial and Spectral Neural Clustering for Organelle Segmentation from Volumetric Electron Microscopy 130
Jimao Jiang and Yuru Pei

CT-Based Hippocampus Segmentation with Dual-Decoder Network (HDD-Net) 140
Wonjun Son, Ji Young Lee, Sung Jun Ahn, and Hyunyeol Lee

DCT-Net: Dual-Branch CT Reconstruction from Orthogonal X-Rays with Diffusion Model and Contrastive Learning 150
Zhiyu Zhang, Cong Shen, Jijun Tang, and Zhijun Liao

EFMS-Net: Efficient Frequency-Enhanced Multi-scale Network for Ischemic Stroke Segmentation 161
Jie Yang, Shaowei Shen, Xuwei Fan, Ning Chen, Zhibin Gao, Lianfen Huang, and Yihong Zhan

Faster, Self-supervised Super-Resolution for Anisotropic Multi-view MRI Using a Sparse Coordinate Loss 172
Maja Schlereth, Moritz Schillinger, and Katharina Breininger

FDF-VQVAE: A Frequency Disentanglement and Fusion Learning Framework for Multi-sequence MRI Enhancement 183
Xinghe Xie, Luyi Han, Yue Sun, Chi Kin Lam, Jian Zheng, Tong Tong, Wei Ke, Chan-Tong Lam, and Tao Tan

Fit Pixels, Get Labels: Meta-learned Implicit Networks for Image Segmentation 194
Kushal Vyas, Ashok Veeraraghavan, and Guha Balakrishnan

Forget-MI: Machine Unlearning for Forgetting Multimodal Information in Healthcare Settings 204
Shahad Hardan, Darya Taratynova, Abdelmajid Essofi, Karthik Nandakumar, and Mohammad Yaqub

Generative Unsupervised Anomaly Detection with Coarse-Fine Ensemble for Workload Reduction in 3D Non-contrast Brain CT of Emergency Room 214
Jongjun Won, Jihwan Kim, Joonseo Oh, Yereen Yoo, Jieun Yum, Joonsang Lee, Joon Hyung Park, Wooyoung Jo, Yoojin Nam, Hyunki Lee, Gil-sun Hong, and Namkug Kim

GLM-SFNet: Global-Local Vision-Mamba with Semantic Fusion for Medical Image Segmentation 224
Jiahui Chen, Fei Qi, Chengyuan Chang, Qinjie Hu, Kaiwen Fu, Xiaotian Wang, and Kun Liu

Harnessing EHRs for Diffusion-Based Anomaly Detection on Chest X-Rays 235
Harim Kim, Yuhan Wang, Minkyu Ahn, Heeyoul Choi, Yuyin Zhou, and Charmgil Hong

Hierarchical Anatomy-Aware Guidance for Brain Tissue Microstructure Reconstruction from T1-Weighted MRI 246
Yuxing Li and Chuyang Ye

Hierarchical Part-Based Generative Model for Realistic 3D Blood Vessel 257
Siqi Chen, Guoqing Zhang, Jiahao Lai, Bingzhi Shen, Sihong Zhang, Caixia Dong, Xuejin Chen, and Yang Li

Hierarchical Spatio-Temporal Segmentation Network for Ejection Fraction Estimation in Echocardiography Videos 268
Dongfang Wang, Jian Yang, Yizhe Zhang, and Tao Zhou

HybridMamba: A Dual-Domain Mamba for 3D Medical Image Segmentation 279
Weitong Wu, Zhaohu Xing, Jing Gong, Qin Peng, and Lei Zhu

IFRFNet: Iterative Frequency Restoration-Fusion Network for Fast System Matrix Calibration on Magnetic Particle Image 290
Weixin Xu, Penghua Zhai, Jie Tian, and Wei Mu

Instrument-Splatting: Controllable Photorealistic Reconstruction of Surgical Instruments Using Gaussian Splatting 301
Shuojue Yang, Zijian Wu, Mingxuan Hong, Qian Li, Daiyun Shen, Septimiu E. Salcudean, and Yueming Jin

Is Hyperbolic Space All You Need for Medical Anomaly Detection? 312
Alvaro Gonzalez-Jimenez, Simone Lionetti, Ludovic Amruthalingam, Philippe Gottfrois, Fabian Gröger, Marc Pouly, and Alexander A. Navarini

LDDMEm: Large Deformation Diffeomorphic Metric Embedding: Decoupling Shape Analysis from Image Registration 323
Greg M. Fleishman and P. Thomas Fletcher

MDAA-Diff: CT-Guided Multi-dose Adaptive Attention Diffusion Model for PET Denoising 333
Xiaolong Niu, Zanting Ye, Xu Han, Yanchao Huang, Hao Sun, Hubing Wu, and Lijun Lu

Medical-Knowledge Driven Multiple Instance Learning for Classifying Severe Abdominal Anomalies on Prenatal Ultrasound 344
Huanwen Liang, Jingxian Xu, Yuanji Zhang, Yuhao Huang, Yuhan Zhang, Xin Yang, Ran Li, Xuedong Deng, Yanjun Liu, Guowei Tao, Yun Wu, Sheng Zhao, Xinru Gao, and Dong Ni

MGG-Net: A Multi-modal Feature Extraction and Global-Aware Feature Graph-Based Deep Learning Network for MGMT Status Classification in Glioma 355
Haoyang Liu, Yuwen Zeng, Xiaoyong Zhang, Wentong Zhou, Arata Nagai, Masayuki Kanamori, Hidenori Endo, and Noriyasu Homma

Minuscule Cell Detection in AS-OCT Images with Progressive Field-of-View Focusing 365
Boyu Chen, Ameenat Solebo, Daqian Shi, Jinge Wu, and Paul Taylor

MixStyleFlow: Domain Generalization in Medical Image Segmentation Using Normalizing Flows 376
Reza Safdari, Mohammad-Ali Nikouei Mahani, Mohamad Koohi-Moghadam, and Kyongtae Tyler Bae

MoDiff: A Morphology-Emphasized Diffusion Model for Ambiguous Medical Image Segmentation 386
Jung Su Ahn, Ki Hoon Kwak, Jung Woo Seo, and Young-Rae Cho

MOST: MR Reconstruction Optimization for Multiple DownStream Tasks via Continual Learning 396
Hwihun Jeong, Se Young Chun, and Jongho Lee

MTCNet: Motion and Topology Consistency Guided Learning for Mitral Valve Segmentation in 4D Ultrasound 407
Rusi Chen, Yuanting Yang, Jiezhi Yao, Hongning Song, Ji Zhang, Yongsong Zhou, Yuhao Huang, Ronghao Yang, Dan Jia, Yuhan Zhang, Xing Tao, Haoran Dou, Qing Zhou, Xin Yang, and Dong Ni

Multifrequency Neural Network-Based Wave Inversion in MR Elastography 418
Héloïse Bustin, Tom Meyer, Jakob Jordan, Lars Walczak, Heiko Tzschätzsch, Ingolf Sack, and Anja Hennemuth

Multi-level Gated U-Net for Denoising TMR Sensor-Based MCG Signals 428
Zeyu Xing, Hao Li, Hao Dou, Zhong Zheng, Jingguo Dai, Chen Wang, Jian Cui, Xin Zhang, and Tianzi Jiang

Multiscale Graph and Multi-step Cross-Frame Mamba for Myocarditis Lesion Segmentation 438
Chengjin Yu, Hao Zhang, Yuanting Yan, Dong Zhang, Sangyin Lv, and Cailing Pu

OsteoOpt: A Bayesian Optimization Framework for Enhancing Bone Union Likelihood in Mandibular Reconstruction Surgery 448
Hamidreza Aftabi, John E. Lloyd, Amanda Ding, Benedikt Sagl, Eitan Prisman, Antony Hodgson, and Sidney Fels

PD-INR: Prior-Driven Implicit Neural Representations for TOF-PET Reconstruction 459
Yuxuan Long, Yulin Zhang, Hong Wang, Xiaodong Kuang, Hailiang Huang, Fan Rao, Huafeng Liu, Yefeng Zheng, and Wentao Zhu

PFESA: FFT-Based Parameter-Free Edge and Structure Attention for Medical Image Segmentation 470
Mingqian Li, Zhiqian Yan, Miaoning Yan, Yaodong Liang, Qingmao Zhang, and Qiongxiong Ma

Physics-Informed Implicit Neural Representations for Joint B0 Estimation and Echo Planar Imaging 481
Wenqi Huang, Nan Wang, Congyu Liao, Yimeng Lin, Mengze Gao, Daniel Rueckert, and Kawin Setsompop

Physiological Neural Representation for Personalised Tracer Kinetic Parameter Estimation from Dynamic PET 491
Kartikay Tehlan and Thomas Wendler

Q-Space Guided Collaborative Attention Translation Network for Flexible Diffusion-Weighted Images Synthesis 501
Pengli Zhu, Yingji Fu, Nanguang Chen, and Anqi Qiu

RDMR: Recursive Inference and Representation Disentanglement for Multimodal Large Deformation Registration 512
Yibo Hu, Ziqi Zhao, Qi Zhang, Lisa X. Xu, and Jianqi Sun

Real Super-Resolution for Proximal Femur: Enhanced Computation of Structural Bone Metrics from Clinical CTs 523
Niklas C. Koser, Marten J. Finck, Felix N. von Brackel, Benjamin Ondruschka, Sören Pirk, and Claus-C. Glüer

Region-Based Text-Consistent Augmentation for Multimodal Medical Segmentation 533
Kunyan Cai, Chenggang Yan, Min He, Liangqiong Qu, Shuai Wang, and Tao Tan

ReSeg-UNet: A Reconstruction-Guided Optimization Framework for Enhanced Medical Image Segmentation 544
Lin Li, Dong Tang, Xiaowen Chu, Xiaofei Yang, and Fei Yu

Reverse Imaging for Wide-Spectrum Generalization of Cardiac MRI Segmentation 555
Yidong Zhao, Peter Kellman, Hui Xue, Tongyun Yang, Yi Zhang, Yuchi Han, Orlando Simonetti, and Qian Tao

Self is the Best Learner: CT-Free Ultra-low-Dose PET Organ Segmentation via Collaborating Denoising and Segmentation Learning 566
Zanting Ye, Xiaolong Niu, Xu Han, Xuanbin Wu, Wantong Lu, Yijun Lu, Hao Sun, Yanchao Huang, Hubing Wu, and Lijun Lu

Semi-Supervised Deformation-Free Image-to-Image Translation for Realistic CT Synthesis from CBCT 577
Ji Yong Han, Su Yang, Sujeong Kim, Sunjung Kim, Sang-Heon Lim, Heejin Yun, Dahee Kim, and Won-Jin Yi

Sim-to-Real Transformer-Based Shape Reconstruction for Automated Orthopedic Fracture Reduction Planning 587
Sutuke Yibulayimu, Yanzhen Liu, Yudi Sang, Gang Zhu, Hui Li, Hao Lu, Chunpeng Zhao, Xinbao Wu, and Yu Wang

SMF-Net: Unlocking Multimodal Insights for Enhanced Stroke Lesion Segmentation 598
Meklit Atlaw, Geng Chen, Haotian Jiang, Xuyun Wen, Hengfei Cui, and Yong Xia

Spatio-Temporal Pre-Trained Foundation Model for Neural Decoding with Fine-Grained Optimization 609
Ziyu Li, Zhiyuan Zhu, Yang Bai, Qing Li, and Xia Wu

Speech Audio Generation from Dynamic MRI via a Knowledge Enhanced Conditional Variational Autoencoder 619
Yaxuan Li, Han Jiang, Yifei Ma, Shihua Qin, Jonghye Woo, and Fangxu Xing

Steerable Anatomical Shape Synthesis with Implicit Neural Representations 630
Bram de Wilde, Max T. Rietberg, Guillaume Lajoinie, and Jelmer M. Wolterink

T2I-Diff: fMRI Signal Generation via Time-Frequency Image Transform and Classifier-Free Denoising Diffusion Models 640
Hwa Hui Tew, Junn Yong Loo, Yee-Fan Tan, Xinyu Tang, Hernando Ombao, Fuad Noman, Raphaël C.-W. Phan, and Chee-Ming Ting

Tied Prototype Model for Few-Shot Medical Image Segmentation 651
Hyeongji Kim, Stine Hansen, and Michael Kampffmeyer

Towards Accurate Tumor Budding Detection: A Benchmark Dataset and A Detection Approach Based on Implicit Annotation Standardization and Positive-Negative Feature Coupling 662
Rui-Qing Sun, Zeng Fan, Boyang Dai, Yiyan Su, Qun Hao, Chuyang Ye, and Shaohui Zhang

Unisyn: A Generative Foundation Model for Universal Medical Image Synthesis Across MRI, CT and PET 673
Yulin Wang, Honglin Xiong, Kaicong Sun, Jiameng Liu, Xin Lin, Ziyi Chen, Yuanzhe He, Qian Wang, and Dinggang Shen

Unpaired Multi-site Brain MRI Harmonization with Image Style-Guided Latent Diffusion 683
Mengqi Wu, Minhui Yu, Weili Lin, Pew-Thian Yap, and Mingxia Liu

Author Index 695

Learning and Augmented Reality for Surgical and Endoscopic Applications (I)

4D CardioSynth: Synthesising Dynamic Virtual Heart Populations Through Spatiotemporal Disentanglement

Haoran Dou[1,2,3], Jinghan Huang[1,2,3], Arezoo Zakeri[1,2,4], Zherui Zhou[1,2,5], Tingting Mu[1,2,3], Jinming Duan[1,2,4], and Alejandro F. Frangi[1,2,3,4,6,7(✉)]

[1] Centre for Computational Imaging and Modelling in Medicine (CIMIM), University of Manchester, Manchester, UK
[2] Christabel Pankhurst Institute, University of Manchester, Manchester, UK
[3] Department of Computer Science, University of Manchester, Manchester, UK
[4] Division of Informatics, Imaging and Data Sciences, University of Manchester, Manchester, UK
[5] Department of Electrical and Electronic Engineering, University of Manchester, Manchester, UK
[6] NIHR Manchester Biomedical Research Centre, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, UK
alejandro.frangi@manchester.ac.uk
[7] Medical Imaging Research Centre (MIRC), Department of Cardiovascular Sciences and Department of Electrical Engineering, KU Leuven, Leuven, Belgium

Abstract. Dynamic virtual populations are critical for realistic in-silico cardiovascular trials, yet current approaches primarily generate static anatomies, limiting their clinical and computational value. In this study, we present 4D CardioSynth, a generative framework for constructing dynamic 3D virtual populations of cardiovascular structures that change over time (3D+t). To model the complex interplay between cardiac structure and motion, we develop a factorised variational approach that disentangles spatial and temporal information in latent space, enabling independent control over anatomical variations and motion patterns. We demonstrate 4D CardioSynth's performance using a diverse dataset of bi-ventricle shapes acquired from 6,500 patients across complete cardiac cycles. Our results illustrate the superiority of 4D CardioSynth over state-of-the-art methods with respect to anatomical specificity, diversity, and generalisability, as well as motion plausibility. This approach enables more accurate virtual trials for cardiovascular interventions.

Keywords: Generative Model · Virtual Population · In-silico Trials

1 Introduction

In-silico trials have emerged as powerful tools that offer a cost-effective and ethical alternative to traditional clinical trials. Virtual populations are a fundamental component of in-silico trials, aiming to provide a more diverse range of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 3–12, 2026.
https://doi.org/10.1007/978-3-032-04947-6_1

anatomical and physiological variations than the restricted real cohorts in clinical trials. These virtual populations are typically represented parametrically as a set of anatomical structures sampled from generative models [5]. In the cardiovascular domain, the generation of realistic patient-like cardiac anatomies that capture both spatial and temporal dynamics is crucial for ensuring the validity and generalisability of in-silico trials. However, creating such virtual populations of dynamic anatomy remains challenging due to the complex interplay between cardiac structure and motion, as well as the potential missing time point across the sequence in the dataset.

Statistical shape models have served as the traditional baseline for generating virtual populations [6,8], but they are limited in their capacity to model complex spatiotemporal variations of cardiac shape. With the advent of deep learning, generative models have emerged as promising alternatives for capturing anatomical variability more accurately. Recent approaches focused on static shape modelling, where Beetz *et al.* proposed a variational autoencoder (VAE) to learn left ventricular shape variations [1], and Dou *et al.* enhanced this framework by incorporating normalising flows in the latent space to improve the model flexibility [4]. Whilst these methods demonstrated improved shape modelling capabilities, their limitation to fixed time points (end-diastole, end-systole or both) restricted their utility in motion-related in-silico studies. To address this limitation, a few studies explored dynamic cardiac anatomy generation. Qiao *et al.* introduced a VAE-based framework to model voxel-based cardiac anatomy sequences, employing recurrent neural networks (RNNs) to capture temporal dynamics [12], and later advancing to transformer architectures for improved spatiotemporal modelling [11]. However, the entanglement of spatial and temporal features in the latent space may constrain the model's flexibility, potentially reducing the diversity of virtual populations. On the other hand, they require the full sequence data to be available for each training iteration so that the latent RNN can capture the temporal information.

We introduce a novel generative framework, namely 4D CardioSynth, for synthesising dynamic 3D+t virtual heart populations with disentangled spatial and temporal representations. Our approach employs a specialised VAE architecture that explicitly decomposes the latent space into independent spatial and temporal subspaces, each processed by dedicated decoders for shape and motion reconstruction. We incorporate a temporal recurrent module to capture complex temporal dependencies, ensuring physiologically plausible cardiac motion patterns. We evaluate our framework comprehensively, assessing anatomical fidelity, diversity and generalisability, as well as motion plausibility through established metrics. Our experimental results demonstrate that the proposed framework generates more realistic cardiac anatomies with superior motion characteristics over the current state-of-the-art method, CHeart [12]. The disentangled representation enables independent control over anatomical variations and motion patterns, facilitating more flexible virtual population generation for in-silico trials.

2 Methodology

Our framework aims to generate dynamic 3D cardiac anatomies by explicitly disentangling spatial and temporal representations through a specialised variational autoencoder (VAE) architecture. As illustrated in Fig. 1, the framework consists of four main components: an encoder that extracts latent representations from input cardiac meshes, a decomposed latent space that separates spatial and temporal information, a temporal recurrent module (TRM) to capture the motion features and dual decoders that reconstruct anatomical shapes and motion patterns, respectively.

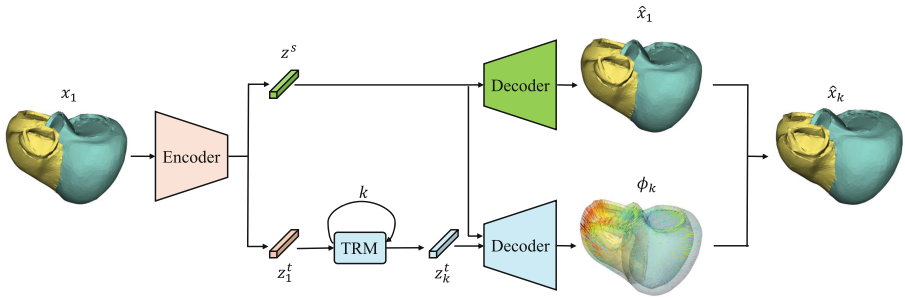


Fig. 1. Schematic illustration of our proposed 4D CardioSynth architecture.

2.1 Dynamic Virtual Heart Populations Synthesis

In our approach, a dynamic virtual heart sequence is represented as a sequence of triangular surface meshes. These meshes correspond to the cardiac structures of interest, i.e., Bi-Ventricle used in our study containing Left and Right Ventricles (BiV, LV and RV). Each mesh is defined by a set of 3D vertex coordinates along with an adjacency matrix that encodes vertex connectivity (i.e., the edges forming triangular faces). By performing template-based registration [14], we ensure that all meshes (regardless of the sample or time point) share a consistent topology and vertex ordering.

Within this framework, we denote the sequence of meshes as $X = \{x_k\}_{k=1}^{T}$, where x_k represents the 3D coordinates of all vertices at time point k. The proposed framework is designed to estimate the conditional distribution $p(x_k \mid x_1)$ of future heart shape x_k given the shape at end-diastole (ED) time point x_1. This is achieved by modeling the following joint distribution:

$$p(x_k, z_k^{\mathrm{t}}, z_1^{\mathrm{t}}, z^{\mathrm{s}}, x_1) = p_\theta(x_k \mid z_k^{\mathrm{t}}, z^{\mathrm{s}}, x_1) p_\gamma(z_k^{\mathrm{t}} \mid z_1^{\mathrm{t}}) q_\psi(z^{\mathrm{s}} \mid x_1) q_\psi(z_1^{\mathrm{t}} \mid x_1), \quad (1)$$

where the latent temporal and spatial codes, i.e., z_1^{t} and z^{s} computed from the static input by an encoder ψ as well as the temporal code z_k^{t} predicted at each time step by a neural network γ, are explained below.

Spatiotemporal Disentanglement: We assume that a heart shape at a fixed time point (i.e., ED in this study) can be explicitly decomposed into two distinct latent components: a spatial code z^{s} and a temporal code z_1^{t}, modelled by the conditional distribution $q_\psi(z^{\mathrm{s}} \mid x_1^{\mathrm{t}})$ and $q_\psi(z_1^{\mathrm{t}} \mid x_1^{\mathrm{t}})$, respectively, parameterised by an encoder ψ. The spatial code z^{s} is intended to capture patient-specific anatomical features that remain constant throughout the cardiac cycle, while the temporal code z_1^{t} encodes the time-dependent motion state for ED.

To enforce the disentanglement, we decode z^{s} to reconstruct the input mesh at ED (producing $\hat{x}_1$) so that z^{s} retains detailed static shape information. In contrast, z_1^{t} is learned indirectly through the TRM rather than via direct decoding, which encourages z_1^{t} to encode only the dynamic aspects of cardiac motion.

Temporal Recurrent Module: The temporal code of the heart changes over time, which is modelled as a first-order Markov process, meaning each latent state depends only on the previous state. As a result, the k-step transition from z_1^{t} to z_k^{t} can be expressed as:

$$p_\gamma(z_k^{\mathrm{t}} \mid z_1^{\mathrm{t}}) = p(z_1^{\mathrm{t}}) \int \prod_{k'=1}^{k-1} p_\gamma(z_{k'+1}^{\mathrm{t}} \mid z_{k'}^{\mathrm{t}}) \, \mathrm{d}z_{1:k-1}^{\mathrm{t}}. \tag{2}$$

Starting from z_1^{t}, we iteratively apply the one-step transition $p_\gamma(z_{k'+1}^{\mathrm{t}} \mid z_{k'}^{\mathrm{t}})$ (parameterized by a multi-layer perceptron γ) to reach z_k^{t}, instead of directly predicting z_k^{t} in a single leap. This yields a smoother temporal transition in the latent space as compared to a direct multi-step prediction.

Vertex-Wise Motion Prediction: We obtain the future mesh at different time step by deforming the current mesh x_1. Applying a deformation field (the displacement of each vertex) from time 1 to time k, denoted by ϕ_k, the predicted mesh coordinates is given by

$$x_k = x_1 \circ \phi_k. \tag{3}$$

As a result, instead of directly estimating the distribution of the absolute coordinates of a heart $p_\theta(x_k \mid z_k^{\mathrm{t}}, z^{\mathrm{s}})$, we learn the distribution of the deformation field $p_\theta(\phi_k \mid z_k^{\mathrm{t}}, z^{\mathrm{s}})$ conditioned on the latent codes, through learning a decoder θ. By focusing on vertex-wise deformations, 4D CardioSynth ensures that the predicted motion remains physiologically coherent, resulting in more realistic virtual cardiac anatomy sequences.

2.2 Network Architecture

Both the encoder ψ and decoder θ contain five residual graph convolutional blocks. Each block comprises two Chebyshev graph convolutions [3], each followed by instance normalisation and SiLU activation [7]. A residual connection is added between the input and output of each graph-convolutional block. The number of feature maps for each block is 16, 32, 32, and 64 in the encoder and inverted in the decoder. We employ hierarchical mesh down/up-sampling

operations as proposed in CoMA [13]. The MLP γ for learning temporal representations contains two fully connected layers with hidden dimensions of 64 and SiLU activation.

2.3 Loss Functions

The encoder ψ for computing static latent codes, the multi-layer perceptron γ for predicting dynamic latent code, and the decoder θ for predicting the deformation are trained in a supervised fashion. An objective function modified from the original evidence lower bound in [9] is used. It consists of three components: reconstruction loss, motion loss, and KullbackLeibler (KL) divergence loss. A reconstruction loss is used to measure the difference between the reconstructed mesh and the input ground-truth mesh:

$$\mathcal{L}_{\text{recon}} = \|x_1 - \hat{x}_1\|_1. \tag{4}$$

To encourage accurate estimation of motion pattern, we use L1-loss to penalise the difference between the predicted mesh and its ground truth:

$$\mathcal{L}_{\text{motion}} = \|x_k - \hat{x}_1 \circ \phi_k\|_1. \tag{5}$$

The KL loss is leveraged to measure the divergence between the approximate posterior and the prior distribution (i.e., standard Gaussian distribution):

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}(q_\psi(z^{\text{s}}|x_0)\|p(z^{\text{s}})) + D_{\text{KL}}(q_\psi(z_0^{\text{t}}|x_0)\|p(z_0^{\text{t}})). \tag{6}$$

Finally, the total loss is a weighted sum of these components:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{recon}} + \lambda_2\mathcal{L}_{\text{motion}} + \lambda_3\mathcal{L}_{\text{KL}}. \tag{7}$$

The setting of $\lambda_1 = 1$, $\lambda_2 = 2e^{-3}$, and $\lambda_3 = 1$ for weighting coefficients is employed, balancing the contribution of each term.

3 Experiments and Results

3.1 Dataset

In this study, we created a cohort of 6500 triangular meshes of the bi-ventricle of the heart based on a subset of cardiac cine-MR imaging data available from the UK Biobank (UKBB). The patient-specific meshes were reconstructed using [15]. Each patient contains 50 time points that cover the full cardiac cycle. We randomly split the dataset into 5000/500/1000 for training, validation, and testing, respectively.

3.2 Implementation Details

The framework was implemented using PyTorch on a standard PC with an NVIDIA RTX 4090 GPU with 24 GB memory. We trained our model using the AdamW optimiser with an initial learning rate of 5e−4 and batch size of 128 for 500 epochs. The spatial and temporal latent dimensions were set at 16. The down/up-sampling factor was four.

3.3 Evaluation Metrics

We compared 4D CardioSynth with the state-of-the-art method, CHeart [12], which uses recurrent neural networks to model joint spatiotemporal representations without disentanglement. We also compared our method to a vanilla VAE [9] trained exclusively on the same datasets with solely single time points (ED) to benchmark spatial representations without input from motion patterns. Three metrics were used in the evaluation: (1) specificity [2], measured by the distance between generated meshes and their nearest neighbours in the real population; (2) diversity [10], measured by the fraction of real samples whose neighbourhoods contain at least one virtual sample; and (3) generalisability [5], formulated as the reconstruction error between the reconstructed and unseen shapes. The Euclidean distance metric was used in all metric calculations.

3.4 Results

Quantitative Analysis of Spatial Features: Our quantitative analysis of the spatial features of virtual populations generated by investigated methods is summarised in Table 1, focusing on the ED time point to isolate spatial aspects of cardiac morphology. It can be observed that 4D CardioSynth demonstrated superior performance compared to CHeart across all metrics and structures. With a biventricular specificity of 2.27 ± 0.49 mm compared to 3.12 ± 1.61 mm of CHeart, our method achieved better fidelity, indicating that the shapes generated by our model more closely resemble the anatomical characteristics observed in real populations. This improvement was consistent across individual chambers. Regarding diversity, 4D CardioSynth achieved 48.3% for BiV compared to CHeart (41.3%), with a 7% improvement in the anatomical variability that models capture from the real population. Similar improvements were observed for individual ventricles as well. Such results suggest that our disentangled latent space approach allows more flexible modelling of diverse cardiac anatomical structures. The results of generalisability evaluation also showed improvements with our method, achieving 1.58 ± 0.43 mm compared to 1.73 ± 0.58 mm for CHeart. This consistent improvement across all three metrics highlights the advantage of our spatiotemporal disentanglement approach. Notably, the conventional VAE model, which was trained exclusively on ED time point data, demonstrated comparable performance in all metrics compared to our method, indicating that 4D CardioSynth can avoid the degradation typically caused by the spatiotemporal entanglement in the latent representation.

Temporal Prediction Performance: Further temporal analysis revealed consistent performance advantages of our method throughout the cardiac cycle. Figure 2 illustrates the predictive error across all future time points given the input of the ED time point for CHeart and 4D CardioSynth. The prediction referred to here was formulated as a sequence completion task where the model takes shape at the first time point (ED) as input and predicts the shapes of the full cardiac cycle. Our model consistently achieved lower predictive errors

Table 1. Quantitative analysis on the spatial features of virtual populations generated by the investigated methods. The specificity and generalisability error are illustrated as mean±std

Methods	Specificity (mm, ↓)			Diversity (%, ↑)			Generalisability (mm, ↓)		
	LV	RV	BiV	LV	RV	BiV	LV	RV	BiV
VAE [9]	1.88 ± 0.45	2.08 ± 0.40	2.17 ± 0.45	49.7	43.7	45.2	1.58 ± 0.53	1.82 ± 0.54	1.70 ± 0.50
CHeart [12]	2.84 ± 1.61	3.01 ± 1.66	3.12 ± 1.61	43.5	40.4	41.3	1.56 ± 0.62	1.88 ± 0.59	1.73 ± 0.58
4D CardioSynth	1.98 ± 0.48	2.17 ± 0.45	2.27 ± 0.49	49.2	44.7	48.3	1.47 ± 0.44	1.68 ± 0.48	1.58 ± 0.43

than the CHeart throughout the cardiac cycle. This performance gap is particularly pronounced during the early time points and late time points, corresponding to diastole phases where anatomical precision is clinically essential. Both methods exhibited increased predictive errors during the mid-cycle time points, which coincide with the rapid deformation phases during systole. However, our method maintained a more stable performance with a narrower error variance. This stability can be attributed to our spatiotemporal disentanglement strategy, which effectively disentangles the complex cardiac motion patterns from spatial anatomical features.

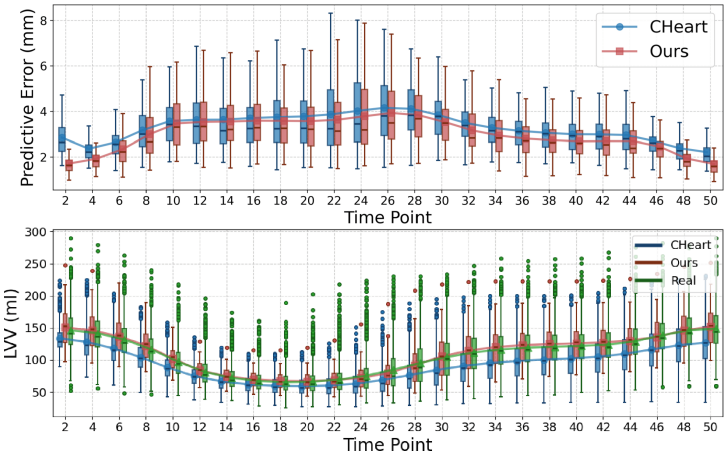


Fig. 2. Temporal analysis of the investigated methods. The top panel displays the distributional comparison of predictive error across the cardiac cycle between CHeart and our method. The bottom panel shows the LVV changes across the cardiac cycle of the real distribution and virtual populations generated by our method and CHeart.

Clinical Relevance of Virtual Populations: To assess the clinical relevance of the generated dynamic 3D+t virtual heart population, we measured each virtual patient's cardiac parameters (i.e., left ventricle volume, LVV). Figure 2

presents the LVV measurements across all 50 time points for the real cohort, virtual populations generated by 4D CardioSynth and CHeart. Both generative models captured the pattern of ventricular volume changes during the cardiac cycle, with volume reduction during systole and subsequent filling during diastole. During the diastole phase, virtual hearts generated by 4D CardioSynth more accurately represented the higher LVV and their associated variability across the population. The median LVV values from our synthetic population closely tracked the real data distribution, whereas the virtual population from CHeart slightly underestimated volumes during these phases. Similarly, during the systole phase, 4D CardioSynth better captured the reduced variability observed in the real data. The temporal coherence of volume changes in our virtual population also more closely resembled the pattern observed in real cardiac cycles, with smoother transitions between phases. This improved temporal modelling can be attributed to our spatiotemporal disentanglement approach, which effectively learns the underlying dynamics of cardiac motion for generating realistic virtual hearts.

Visual Assessment of Generated Virtual Hearts: To demonstrate the quality of our generated virtual patient of the heart anatomy, Fig. 3 visualises a representative heart shape generated by our method across different time points of the cardiac cycle. The visualisation shows the plausibility of the generated cardiac shape and the myocardium smoothly deforming through systole and diastole.

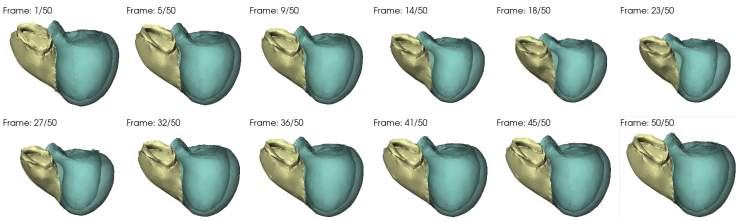


Fig. 3. Visualisation of a representative heart shape generated by our method at selected time points throughout the cardiac cycle.

4 Conclusion

In this study, we presented 4D CardioSynth for synthesising dynamic 3D+t virtual heart populations for in-silico trials. To learn the complex interplay between cardiac structure and motion, we introduced a disentangled latent space within a specialised VAE that independently captures spatial and temporal representations. This design enables separate control over anatomical variations and motion patterns, facilitating more flexible and targeted virtual population generation.

We compared our method to CHeart, with results demonstrating that our approach achieves better fidelity, diversity, and generalisability in terms of spatial patterns, as well as the plausibility of the cardiac motion. Future work will focus on extending the current anatomical representation to encompass the full heart and more detailed structures, such as heart valves.

Acknowledgments. This research was carried out using data from the UK Biobank (access application 11350). This work was supported by the Royal Academy of Engineering, United Kingdom under the RAEng Chair in Emerging Technologies (INSILEX CiET1919/19), the ERC Advanced Grant UKRI Frontier Research Guarantee (INSILICO EP/Y030494/1), the UK Centre of Excellence on in-silico Regulatory Science and Innovation (UK CEiRSI) (10139527), the National Institute for Health and Care Research (NIHR) Manchester Biomedical Research Centre (BRC) (NIHR203308), the BHF Manchester Centre of Research Excellence (RE/24/130017), and the CRUK RadNet Manchester (C1994/A28701).

Disclosure of Interests. AFF is a Founder, Board Member and Shareholder of adsilico ltd and OculomeX Health Ltd and consults with both companies. AFF has non-financial collaborative agreements with several medical device manufacturers. AFF and HD have patents licensed to adsilico or OculomeX and receive royalties from these arrangements. AFF also receives royalties from Elsevier associated with a medical image computing textbook. All other authors declare no competing interests.

References

1. Beetz, M., et al.: Interpretable cardiac anatomy modeling using variational mesh autoencoders. Front. Cardiovas. Med. **9**, 983868 (2022)
2. Davies, R.H., Twining, C.J., Cootes, T.F., Taylor, C.J.: Building 3-D statistical shape models by direct optimization. IEEE Trans. Med. Imaging **29**(4), 961–981 (2009)
3. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
4. Dou, H., Ravikumar, N., Frangi, A.F.: A conditional flow variational autoencoder for controllable synthesis of virtual populations of anatomy. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14226, pp. 143–152. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43990-2_14
5. Dou, H., Virtanen, S., Ravikumar, N., Frangi, A.F.: A generative shape compositional framework to synthesize populations of virtual chimeras. IEEE Trans. Neural Netw. Learn. Syst. (2024)
6. Frangi, A.F., Rueckert, D., Schnabel, J.A., Niessen, W.J.: Automatic construction of multiple-object three-dimensional statistical shape models: application to cardiac modeling. IEEE Trans. Med. Imaging **21**(9), 1151–1166 (2002)
7. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415 (2016)

8. Hoogendoorn, C., Sukno, F.M., Ordás, S., Frangi, A.F.: Bilinear models for spatio-temporal point distribution analysis: application to extrapolation of left ventricular, biventricular and whole heart cardiac dynamics. Int. J. Comput. Vision **85**(3), 237–252 (2009)
9. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)
10. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International Conference on Machine Learning, pp. 7176–7185. PMLR (2020)
11. Qiao, M., McGurk, K.A., Wang, S., Matthews, P.M., Regan, D.P., Bai, W.: A personalised 3D+ t mesh generative model for unveiling normal heart dynamics. arXiv preprint arXiv:2409.13825 (2024)
12. Qiao, M., et al.: CHeart: a conditional spatio-temporal generative model for cardiac anatomy. IEEE Trans. Med. Imaging **43**(3), 1259–1269 (2023)
13. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3D faces using convolutional mesh autoencoders. In: Proceedings of the European conference on computer vision (ECCV), pp. 704–720 (2018)
14. Ravikumar, N., Gooya, A., Frangi, A.F., Taylor, Z.A.: Generalised coherent point drift for group-wise registration of multi-dimensional point sets. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 309–316. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_36
15. Xia, Y., et al.: Automatic 3D+ t four-chamber CMR quantification of the UK biobank: integrating imaging and non-imaging data priors at scale. Med. Image Anal. **80**, 102498 (2022)

A Unified Continuous Staging Framework for Alzheimer's Disease and Lewy Body Dementia via Hierarchical Anatomical Features

Tong Chen[1], Minheng Chen[1], Jing Zhang[1], Yan Zhuang[1], Chao Cao[1], Xiaowei Yu[1], Yanjun Lyu[1], Lu Zhang[2], Li Su[3,4], Tianming Liu[5], and Dajiang Zhu[1(✉)]

[1] Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, USA
dajiang.zhu@uta.edu
[2] Department of Computer Science, Indiana University Indianapolis, Indianapolis, USA
[3] Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, UK
[4] Department of Psychiatry, University of Cambridge, Cambridge, UK
[5] School of Computing, University of Georgia, Athens, USA

Abstract. Alzheimer's Disease (AD) and Lewy Body Dementia (LBD) often exhibit overlapping pathologies, leading to common symptoms that make diagnosis challenging and protracted in clinical settings. While many studies achieve promising accuracy in identifying AD and LBD at earlier stages, they often focus on discrete classification rather than capturing the gradual nature of disease progression. Since dementia develops progressively, understanding the continuous trajectory of dementia is crucial, as it allows us to uncover hidden patterns in cognitive decline and provides critical insights into the underlying mechanisms of disease progression. To address this gap, we propose a novel multi-scale learning framework that leverages hierarchical anatomical features to model the continuous relationships across various neurodegenerative conditions, including Mild Cognitive Impairment, AD, and LBD. Our approach employs the proposed hierarchical graph embedding fusion technique, integrating anatomical features, cortical folding patterns, and structural connectivity at multiple scales. This integration captures both fine-grained and coarse anatomical details, enabling the identification of subtle patterns that enhance differentiation between dementia types. Additionally, our framework projects each subject onto continuous tree structures, providing intuitive visualizations of disease trajectories and offering a more interpretable way to track cognitive decline. To validate our approach, we conduct extensive experiments on our in-house dataset of 308 subjects spanning multiple groups. Our results demonstrate that the proposed tree-based model effectively represents dementia progression, achieves promising performance in intricate classification task of AD and LBD, and highlights discriminative brain regions that

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 13–23, 2026.
https://doi.org/10.1007/978-3-032-04947-6_2

contribute to the differentiation between dementia types. Our code is available at https://github.com/tongchen2010/haff.

Keywords: Alzheimer's Disease · Lewy Body Dementia · Dementia Progression · Cortical folding pattern · Multi-scale fusion · 3-Hinge Gyrus

1 Introduction

Alzheimer's Disease (AD) and Lewy Body Dementia (LBD) are two of the most common neurodegenerative disorders, with AD accounting for 60–80% of cases and LBD contributing approximately 4.2% [1,18]. Both lead to irreversible brain changes, severe cognitive decline, and ultimately death. Early diagnosis and studying disease progression are crucial, as they enable patients and families to plan care [1]. However, differentiating AD and LBD is challenging due to overlapping neuropathological and symptomatic features. For example, LBD patients may also exhibit beta-amyloid plaques typical of AD, while AD patients can show parkinsonian symptoms resembling Parkinson Disease Dementia, a subtype of LBD [8]. This overlap often results in misdiagnosis and delayed diagnosis, approximately 18 months for LBD and 14.8 months for young AD patients after initial hospital visit [10]. Although current AD biomarkers-such as CSF Aβ 42/Aβ40 ratios, tau tests, and plasma ptau181-are effective, they are invasive, expensive, and not widely accessible, prompting interest in non-invasive alternatives like structural magnetic resonance imaging (sMRI) and diffusion tensor imaging (DTI) for differential diagnosis [19,24].

Recent advancements in deep learning have demonstrated promising performance in the early diagnosis of AD [21,22]. However, many of these approaches are primarily focused on classification tasks, which limits their ability to capture the continuous progression of dementia. A recent study introduced a tree embedding framework that models the continuous progression of AD by encoding anatomical features across different disease stages [23]. This framework not only achieves state-of-the-art performance in multi-class classification but also projects patient data onto a tree-like structure, facilitating a more intuitive visualization of AD progression. Despite these strengths, the approach relies on coarse anatomical features, potentially overlooking small brain regions that are critical for distinguishing between overlapping dementia syndromes. For instance, previous research has noted that only limited areas within the temporal lobe exhibit significantly greater cortical thinning in AD compared to LBD [13,28,29], therefore, it's necessary to include fine-scale measures to differentiate AD and LBD. Furthermore, its application is limited to AD and does not demonstrate the capacity to address the complexities of multi-dementia progression.

To overcome the limitations of prior work, we propose a multi-scale feature fusion framework that captures hierarchical anatomical characteristics to model the continuous transitions of various dementia stages, CN, MCI, AD and LBD.

Our method employs a graph neural network-based model to integrate structural connectivity (SC) and cortical folding pattern features at different scales so that it captures the fine-grained regional differences that's able to differentiate AD and LBD. Coarse regions are derived from the neuroanatomical atlas, ensuring consistency with well-established cortical parcellation framework. In contrast, hub-based patches are obtained using an innovative cortical folding pattern representation, namely 3-Hinge Gyrus (3HG) [2,3,26] where 3HG is defined by the conjunction of gyri from three different directions. 3HG can serve as cortical hubs [27] and 3HG connectomes has outperforms traditional region-based connectomes in classifying AD stages, underscoring the benefits of finer-scale brain region [11]. The Multi-scale Graph Embedding Fusion (MGEF) block of our method effectively combines the graph emebddings of both atlas-based and hub-based anatomical features, enabling improved differentiation between AD and LBD. To validate our approach, we conducted extensive experiments on our in-house dataset, demonstrating its effectiveness in addressing the intricate task of differentiating four brain conditions and visualizing the dementia progression.

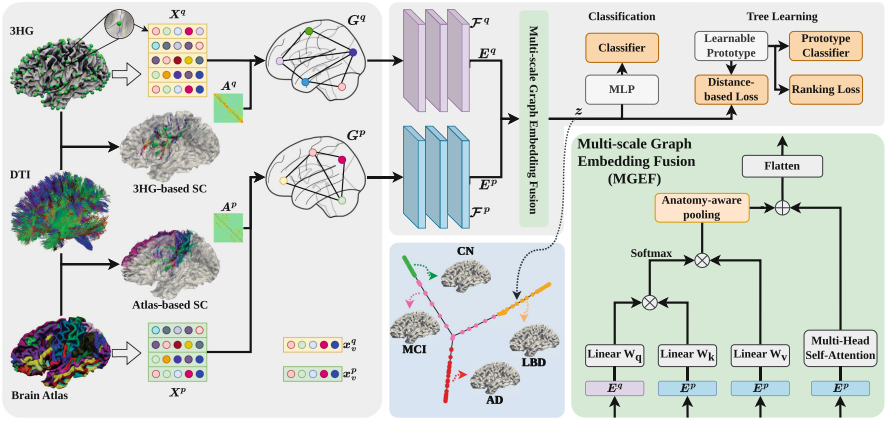


Fig. 1. illustrates our overall framework. For each subject, we generate two graphs: one for atlas-based and one for 3HG-based anatomical features. We feed the graph data into our proposed framework, which learns meaningful representations for classification and subsequently generates a tree structure to visualize dementia progression.

2 Methods

2.1 Anatomical Feature Selection

We leverage vertex-based cortical features (CF) to characterize cortical morphology, including cortical thickness (CTh), curvature, sulcal depth (SD), local gyrification index (LGI), and fractal dimension (FD). CTh, SD, curvature, and

LGI are derived using Freesurfer tools [5], while FD is computed with our in-house tool using the Minkowski–Bouligand dimension. Notably, measures such as CTh have been shown to correlate strongly with dementia severity and are widely used to differentiate clinical stages such as AD, CN, and MCI [12,20]. Additionally, FD quantifies the complexity of cortical patterns by examining changes across different scales, and has recently emerged as a valuable metric for assessing cortical atrophy in dementia [17,30].

2.2 Modal Architecture

Graph Convolution Network. Figure 1 illustrates the overview of our proposed framework. Each subject is characterized by two types of graph inputs, namely $G^p = (\mathcal{V}^p, \mathcal{E}^p)$ for atlas-based data and $G^q = (\mathcal{V}^q, \mathcal{E}^q)$ for 3HG-based data, where $\mathcal{V}$ denotes a set of nodes and $\mathcal{E}$ denotes a set of edges connecting those nodes. For any node v, its feature is denoted by x_v and the matrix X represents the features of all nodes. The edge between nodes u and v is denoted by e_{vu}, and the adjacency matrix A is constructed from these edges, representing the SC. The label for each subject is given by y. For the model, we first employ two dual graph convolutional networks (GCN) [9], defined by functions $\mathcal{F}(\cdot)$ that operate on the respective graph data. The node embeddings for the two modalities are computed as $E^p = \mathcal{F}^p(X^p, A^p)$ and $E^q = \mathcal{F}^q(X^q, A^q)$, where $E^p \in \mathbb{R}^{r\times k}$ and $E^q \in \mathbb{R}^{s\times k}$ denote the node embeddings for each graph input projected into a k-dimensional latent space, with r and s being the numbers of atlas regions and 3HG, respectively.

Multi-scale Anatomy-Aware Fusion. After $\mathcal{F}(\cdot)$, E^p are fed into a multi-head self-attention block to capture the global contextual information of X^p. In contrast, instead of applying self-attention to X^q, we introduce a cross attention mechanism, termed *Anatomy-aware Cross Attention* (ACA) (Eq. 1), that enables each x_v^q, identified by its label m_v to query the corresponding X^p. This mechanism allows the 3HG nodes to extract relevant information from the atlas regions. The cross attention is formulated as follows:

$$\begin{gathered} Q^q = E^q W_Q, \quad K^p = E^p W_K, \quad V^p = E^p W_V, \\ \text{Attention}(Q^q, K^p, V^p) = \text{softmax}\left(\frac{Q^q {K^p}^\top}{\sqrt{d_k}}\right) V^p, \end{gathered} \tag{1}$$

where W_Q, W_K, and W_V are learnable projection matrices, and d_k is the dimensionality of the node embeddings. Following the cross-attention mechanism, we perform our proposed *Anatomy-aware Mean Pooling* (AMP) (Eq. 2). In the equation, R_j represents a set of x^q located in atlas region j and $e_i^q \in \mathbb{R}^d$ is the embedding of node x_i^q. AMP aggregates the x^q in each atlas region, and if a region has no 3HG (i.e., $|R_j| = 0$), its pooled embedding is defined as the zero vector in $\mathbb{R}^d$.

$$m_j = \begin{cases} \dfrac{1}{|R_j|} \displaystyle\sum_{i\in R_j} e_i^q, & \text{if } |R_j| > 0, \\ \mathbf{0} \in \mathbb{R}^d, & \text{if } |R_j| = 0. \end{cases} \tag{2}$$

Finally, the pooled 3HG embeddings are combined with the atlas region embeddings by a simple element-wise sum to form the final feature representation $Z = E^p + E^q$, which is then processed for classification using cross-entropy loss and tree structure learning.

2.3 Tree Structure Learning

Prototype Model. Firstly a set of a learnable prototype embeddings are defined as $\mathscr{E}_{\text{learn}} = \{e_i \in \mathbb{R}^k \mid i = 1, 2, \ldots, C\}$, where $\mathscr{E}_{\text{learn}}$ represents the learnable parameters for C classes in the k-dimensional latent space where $k = r \times d$. Each prototype embedding e_i represents a class y_i in the latent space. Then we minimize the distance between the z_i and its corresponding prototype embeddings e_i using l2 distance and the distance-based loss of each sample towards its corresponding prototype center is formulated as $\mathcal{L}_{\text{distance}} = \|z_i - e_y\|_2^2$ where e_y is the corresponding class embedding of the input x_i. To enhance generalization, a batch-wise mean squared error (MSE) loss over b samples was incorporated into the overall loss function: $\mathcal{L}_{\text{MSE}} = \frac{1}{b}\sum_{i=1}^{b} \|h(x_b, \theta) - e_y\|^2$. The tree loss combines the distance loss and MSE loss where α is a hyperparameter:

$$\mathcal{L}_{\text{tree}} = \mathcal{L}_{\text{distance}} + \alpha \mathcal{L}_{\text{MSE}} \tag{3}$$

Learnable Prototype Embedding. A set prototype embeddings $\mathcal{E}_{\text{init}} \in \mathbb{R}^{C \times k}$ is firstly initialized such that all prototype embeddings are equally spaced in terms of euclidean distance. The learnable embedding matrix $\mathcal{E}_{\text{learn}} \in \mathbb{R}^{C \times k}$ is defined by $\mathcal{E}_{\text{learn}} = \mathcal{E}_{\text{init}} \cdot \ \mathrm{W} + b$ where W and b are learnable parameters. Prototype learning consists of two loss functions:

a. Classification loss ensures the learnable prototype embeddings are aligned with their class labels using a cross-entropy loss. $\mathcal{L}_{\text{classification}}$ is the classification loss of the prototype embedding to its corresponding label.
b. Ranking Loss [4] is adopted as a regularization term and also to capture the prior knowledge of known orders between classes (e.g., CN $\rightarrow$ MCI $\rightarrow$ AD). The similarity-based adjacency matrix S is constructed based on the following:

$$S_{i,j} = \begin{cases} s_0 & \text{if } \text{label}_i = \text{label}_j, \\ s_p & \text{if } |\text{label}_i - \text{label}_j| = d_p, \quad \text{for } p = 1, 2, \ldots, k, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Here, i and j denote the row and column indices of the matrix. The term s_0 represents the self-similarity values when labels are identical. The values $s_1, s_2, \ldots, s_k$ are predefined similarity values corresponding to neighboring classes at increasing distances. The thresholds $d_1, d_2, \ldots, d_k$ define specific distances between labels that establish adjacency, where each d_p corresponds to similarity level s_p. This setup ensures that similarity values are applied based on both label proximity and predefined distance thresholds. To normalize the adjacency matrix, we compute the Laplacian matrix $\mathbb{L}_y = S - D_s$, where D_s is

the degree matrix of S. Using this Laplacian matrix, the order constraint loss is defined as:

$$\mathcal{L}_{\text{ranking}} = \text{trace}(O_{\text{logit}} \cdot \mathbb{L}_y \cdot O_{\text{logit}}^T) \tag{5}$$

The final loss for prototype embedding learning combines the classification loss and the order constraint loss where β is a hyperparameter(Eq. 5) as follows:

$$\mathcal{L}_{\text{prototype}} = \mathcal{L}_{\text{classification}} + \beta \mathcal{L}_{\text{ranking}} \tag{6}$$

3 Experiments

3.1 Dataset

Data Preprocessing. In our study, we analyzed a dataset comprising T1-weighted sMRI and DTI scans from 308 subjects, including 77 cognitively normal (CN), 77 MCI, 77 AD, and 77 LBD subjects. All T1-weighted MRI scans underwent standardized preprocessing procedures as described in [25]. In brief, the preprocessing included brain extraction using the FMRIB Software Library (FSL), cortical surface reconstruction for both white matter and pial surfaces, and subsequent parcellation based on the Destrieux Atlas using the FreeSurfer toolkit [5]. DTI data were similarly preprocessed with fiber tracking conducted using DSI Studio.

3.2 Implementation Details

All experiments were conducted using 5-fold cross-validation. The GCN model was implemented with two GCN layers, and each multi-head attention block consisted of four attention heads. For training, all models were run with 100 epochs with a batch size of 16 and a dropout rate of 0.3. Regarding optimization, we employed the AdamW optimizer together with the ReduceLROnPlateau scheduler. Specific hyperparameters included a learning rate of 0.001 for the classification model and $\alpha = 0.1$ and $\beta = 0.3$ for the prototype model.

4 Results

4.1 Classification

Table 1 summarizes the classification performance for CN, MCI, AD, and LBD. Because of the pathological and symptomatic overlap, differentiating AD from LBD is particularly challenging. Previous studies have primarily focused on simpler binary comparisons, such as AD versus LBD using voxel-based (VB) method [7,15,16], rather than tackling the complexity of a four-class classification task. Traditional classifiers using only SC achieved moderate accuracies between 54.85% and 57.42%, reflecting the inherent difficulty of the problem. In contrast, our innovative approach, which seamlessly integrates CF with SC, achieved a notable accuracy of 61.68%, demonstrating the effectiveness of multimodal feature integration.

Table 1. Comparison of classification performance metrics for distinguishing among CN, MCI, AD, and LBD using various methods and feature combinations.

Method	Groups	Feature	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 (%)
Nemoto *et al.* [16]	AD vs LBD	VB	79.15 ± 5.22	81.54 ± 10.43	76.77	N/A
Nakatsuka *et al.* [15]	AD vs LBD	VB	75.00	80.00	64.00	N/A
Goto *et al.* [7]	AD vs LBD	VB	70.00	N/A	N/A	N/A
MLP	CN vs MCI vs AD vs LBD	SC	57.10 ± 6.19	57.10 ± 5.95	85.7 ± 2.03	56.79 ± 5.56
SVM	CN vs MCI vs AD vs LBD	SC	55.79 ± 8.14	55.79 ± 8.14	85.26 ± 2.69	55.46 ± 8.02
Random Forest	CN vs MCI vs AD vs LBD	SC	54.85 ± 5.02	54.85 ± 5.02	84.94 ± 1.65	54.07 ± 5.15
Logistic Regression	CN vs MCI vs AD vs LBD	SC	55.48 ± 7.94	55.54 ± 7.85	85.17 ± 2.63	54.76 ± 7.65
Our method	CN vs MCI vs AD vs LBD	CF+SC	**61.68 ± 3.12**	**61.35 ± 4.80**	**87.19 ± 1.07**	**57.42 ± 5.76**

4.2 Top Discriminative Atlas Region and 3HG

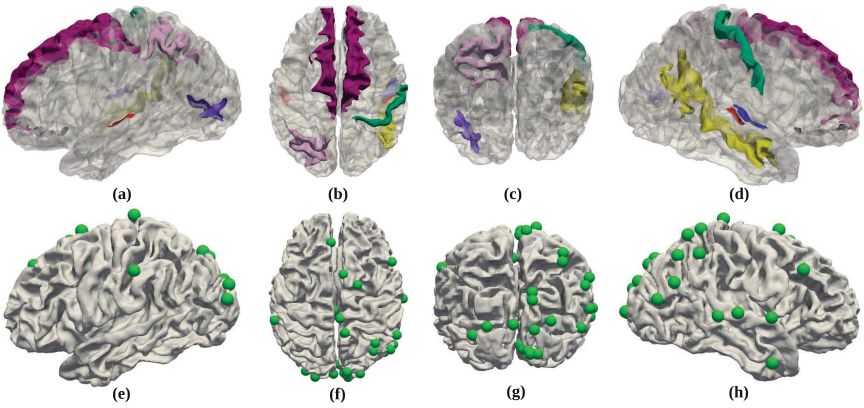


Fig. 2. (a)–(d) present the 10 most discriminative atlas regions, while (e)–(h) showcase the top 30 discriminative 3HG features for classification.

Figure 2 shows the top 10 Destrieux regions contributing to the classification task. In particular, Fig. 2(b) and (d) highlight the superior frontal gyri in both hemispheres, consistent with studies linking impairments in working memory and attention to AD [14]. Additionally, Fig. 2(d) underscores the superior temporal sulcus (STS), with prior work reporting an approximate 50% loss in AD, which may relate to social and cognitive deficits [6]. Figure 2(c) draws attention to occipital regions, potentially linked to visual hallucinations in LBD. Figure 2(e)–(h) present the top 30 discriminative 3HG features for classification. This approach allows for the detection of microstructural changes that may be missed by traditional atlas-based methods. By complementing the atlas-based analysis, the 3HG features enhance the sensitivity of the classification task, potentially leading to improved diagnostic accuracy.

4.3 Embedding Tree

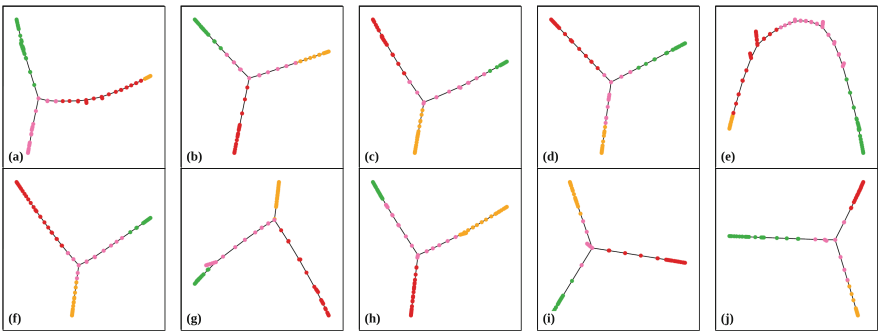


Fig. 3. Each panel (a–j) presents a disease embedding tree learned by the proposed framework, showing how individual subjects (represented by circles) cluster according to their disease stage. The four disease stages are color-coded: CN in **green**, MCI in pink, AD in **red**, and LBD in orange. (Color figure online)

For the tree structure generation, we adopted a distance-based method proposed in [23]. In Fig. 3, we present 10 learned trees to demonstrate the stability of the tree structures. Although each tree's structure shows some variability, they consistently reveal a key pattern: LBD subjects tend to cluster in close proximity to MCI and AD subjects. This clustering on the tree structure suggests that the anatomical and connectivity features captured by our model reflect underlying similarities in the disease progression of LBD relative to MCI and AD. Figure 3(b–d), (f–j) predominantly indicate that LBD subjects are positioned near MCI subjects while panels Fig. 3(a) and (e) show LBD subjects clustering closer to AD subjects. The tree structures provide a continuous representation of patient status. By observing patients onto the disease tree, we can learn their brain state within the cohort. These tree structures offer a continuous representation of patient status, allowing us to map each patient's brain state within the cohort. By situating patients on the disease tree, we can compare their progression relative to others and gain insights into the continuum of dementia progression.

5 Conclusion

In this study, we introduce a novel multi-scale learning framework that captures hierarchical anatomical features and effectively models continuous relationships across various dementia conditions. Our approach not only outperforms both traditional and deep learning models in a multi-dementia classification task but also provides a comparative and interpretable visualization of patient progression through a tree-structured representation. This tree structure offers valuable insights into disease trajectories, enabling a more nuanced understanding of

neurodegenerative progression. Furthermore, our framework identifies the most discriminative atlas regions and 3HG features, highlighting key regions that distinguish different forms of dementia.

Acknowledgments. This study was supported by the National Institutes of Health (R01AG075582 and RF1NS128534) and funded by the Sir Jules Thorn Charitable Trust (05/JTA), Wellcome Trust Intermediate Clinical Fellowship (WT088441MA), Alzheimer's Research UK Senior Research Fellowship (ARUK-SRF2017B-1), The Lewy body Society (LBS/002/2019) and NIHR Biomedical Research Centres in Sheffield, Cambridge and Newcastle. We also thank Andrew M. Blamire, John-Paul Taylor, Li Su and John T. O'Brien for sharing the data.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 2024 Alzheimer's disease facts and figures. Alzheimer's Dementia **20**(5), 3708–3821 (2024)
2. Chen, H., Li, Y., Ge, F., Li, G., Shen, D., Liu, T.: Gyral net: a new representation of cortical folding organization. Med. Image Anal. **42**, 14–25 (2017)
3. Chen, M., et al.: Using structural similarity and Kolmogorov-Arnold networks for anatomical embedding of cortical folding patterns. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2025)
4. Chen, W., Liu, T.Y., Lan, Y., Ma, Z.M., Li, H.: Ranking measures and loss functions in learning to rank. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 22. Curran Associates, Inc. (2009)
5. Fischl, B.: Freesurfer. Neuroimage **62**(2), 774–781 (2012)
6. Gómez-Isla, T., et al.: Neuronal loss correlates with but exceeds neurofibrillary tangles in Alzheimer's disease. Ann. of Neurol.: Off. J. Am. Neurol. Assoc. Child Neurol. Soc. **41**(1), 17–24 (1997)
7. Goto, H., et al.: Differential diagnosis of dementia with Lewy bodies and Alzheimer disease using combined MR imaging and brain perfusion single-photon emission tomography. Am. J. Neuroradiol. **31**(4), 720–725 (2010)
8. Gurnani, A.S., Gavett, B.E.: The differential effects of Alzheimer's disease and Lewy body pathology on cognitive performance: a meta-analysis. Neuropsychol. Rev. **27**(1), 1–17 (2017)
9. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the 5th International Conference on Learning Representations, ICLR 2017 (2017)
10. Kvello-Alme, M., Bråthen, G., White, L.R., Sando, S.B., Knopman, D.: Time to diagnosis in young onset Alzheimer's disease: a population-based study from central Norway. J. Alzheimers Dis. **82**(3), 965–974 (2021)
11. Lyu, Y., Zhang, L., Yu, X., Cao, C., Liu, T., Zhu, D.: Mild cognitive impairment classification using a novel finer-scale brain connectome. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2024)

12. Mak, E., et al.: Longitudinal assessment of global and regional atrophy rates in Alzheimer's disease and dementia with Lewy bodies. NeuroImage: Clin. **7**, 456–462 (2015)
13. Mak, E., et al.: Progressive cortical thinning and subcortical atrophy in dementia with Lewy bodies and Alzheimer's disease. Neurobiol. Aging **36**(4), 1743–1750 (2015)
14. Miller, E.K., Cohen, J.D.: An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. **24**(1), 167–202 (2001)
15. Nakatsuka, T., Imabayashi, E., Matsuda, H., Sakakibara, R., Inaoka, T., Terada, H.: Discrimination of dementia with Lewy bodies from Alzheimer's disease using voxel-based morphometry of white matter by statistical parametric mapping 8 plus diffeomorphic anatomic registration through exponentiated lie algebra. Neuroradiology **55**(5), 559–566 (2013)
16. Nemoto, K., et al.: Differentiating dementia with Lewy bodies and Alzheimer's disease by deep learning to structural MRI. J. Neuroimaging **31**(3), 579–587 (2021)
17. Núñez, C., Callén, A., Lombardini, F., Compta, Y., Stephan-Otto, C.: for the Alzheimer's Disease Neuroimaging Initiative: Different cortical gyrification patterns in Alzheimer's disease and impact on memory performance. Ann. Neurol. **88**(1), 67–80 (2020)
18. Vann Jones, S.A., O'Brien, J.T.: The prevalence and incidence of dementia with Lewy bodies: a systematic review of population and clinical studies. Psychol. Med. **44**(4), 673–683 (2014)
19. Wang, L., Thompson, P.M., Zhu, D.: Analyzing mild cognitive impairment progression via multi-view structural learning. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 656–668. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_51
20. Wee, C.Y., Yap, P.T., Shen, D.: for the Alzheimer's disease neuroimaging initiative: prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns. Hum. Brain Mapp. **34**(12), 3411–3425 (2013)
21. Zhang, J., et al.: Classification of mild cognitive impairment based on dynamic functional connectivity using spatio-temporal transformer. arXiv preprint arXiv:2501.16409 (2025)
22. Zhang, J., et al.: Brain-adapter: enhancing neurological disorder analysis with adapter-tuning multimodal large language models. arXiv preprint arXiv:2501.16282 (2025)
23. Zhang, L., Wang, L., Liu, T., Zhu, D.: Disease2vec: encoding Alzheimer's progression via disease embedding tree. Pharmacol. Res. **199**, 107038 (2024)
24. Zhang, L., Wang, L., Zhu, D.: Recovering brain structural connectivity from functional connectivity via multi-GCN based generative adversarial network. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12267, pp. 53–61. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59728-3_6
25. Zhang, L., Wang, L., Zhu, D.: Predicting brain structural network using functional connectivity. Med. Image Anal. **79**, 102463 (2022)
26. Zhang, L., et al.: Cortex2Vector: anatomical embedding of cortical folding patterns. Cereb. Cortex **33**(10), 5851–5862 (2022)
27. Zhang, T., et al.: Cortical 3-hinges could serve as hubs in cortico-cortical connective network. Brain Imaging Behav. **14**(6), 2512–2529 (2020)
28. Zhu, D., et al.: DICCCOL: dense individualized and common connectivity-based cortical landmarks. Cereb. Cortex **23**(4), 786–800 (2013)

29. Zhu, D., et al.: Connectome-scale assessments of structural and functional connectivity in mci. Hum. Brain Mapp. **35**(7), 2911–2923 (2014)
30. Ziukelis, E.T., Mak, E., Dounavi, M.E., Su, L., T O'Brien, J.: Fractal dimension of the brain in neurodegenerative disease and dementia: a systematic review. Ageing Res. Rev. **79**, 101651 (2022)



Accelerated Free-Breathing 5D Multi-Echo Respiratory Motion-Resolved R2*, PDFF, and QSM Using Novel Composite Total Variation

MungSoo Kang[1,2], Or Alus[1,2], and Youngwook Kee[1,2(✉)]

[1] Stony Brook University, Stony Brook, NY 11794, USA
youngwook.kee@stonybrook.edu
[2] Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

Abstract. We introduce a novel composite total variation (TV) and its solution algorithm with their application to multi-echo, respiratory motion-resolved 5D (3D space + 1D respiratory motion + 1D echo signal evolution) compressed sensing (CS) abdominal MR image reconstruction. The proposed formalism ensures a sparse representation between multi-echo images with varying contrast—a vital feature that needs to be preserved—making it highly suitable for applications in multi-dimensional computational/quantitative imaging. The key idea of the proposed composite TV and its formal definition were inspired by the observation that the spatial gradient of difference images in multi-echo MRI appears sparse. Throughout extensive experiments on a small number of healthy volunteers, we have demonstrated improved performance of the proposed method in 5D motion-resolved CS reconstruction of multi-echo MRI data compared to the state-of-the-art method. We have also demonstrated improved performance of the proposed method in quantitative tissue parameter mapping (such as R2*, proton density fat fraction, and quantitative susceptibility mapping) across a wide range of undersampling factors. In conclusion, the proposed method enables vastly accelerated motion-resolved multi-echo CS-MRI minimally impacting the quantification of downstream tissue parameters.

Keywords: Compressed sensing · Model-based MR image reconstruction · Non-Cartesian multi-echo MRI · Quantitative imaging

1 Introduction

Free-breathing multi-echo gradient echo (mGRE) MRI has recently emerged as a new imaging technique [1,14,17,21,22]. This technique acquires 3D volumetric images at multiple time points, encoding a 1D echo signal evolution at each voxel over time (on the order of milliseconds). The slower, second-scale respiratory motion embedded in the 4D spatiotemporal mGRE data can then be retrospectively resolved to mitigate motion-induced errors [8]. Free-breathing mGRE

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 24–34, 2026.
https://doi.org/10.1007/978-3-032-04947-6_3

MRI with respiratory motion-resolved reconstruction has enabled voxel-wise quantification of R2*, proton density fat fraction (PDFF), and tissue magnetic susceptibility [14,21,22], offering great potential for use in patient populations—such as children or certain adults—where maintaining a breath-hold is impractical.

Despite its free-running and/or free-breathing capability, a major challenge in its clinical adoption is the lengthy acquisition time. This increases the likelihood of subject bulk motion that may not be adequately corrected or resolved under the typical assumption of strictly periodic motion (e.g., respiration). Moreover, there is a trade-off between high isotropic resolution and maintaining a shorter acquisition time. Accurate tissue parameter mapping often requires high isotropic resolution to avoid undesirable bias [16,26]. Ideally, such a 5D imaging dataset would require multi-way array or tensor-based reconstruction/processing, but this area has not been extensively explored. One approach is to reconstruct all echoes simultaneously using ℓ_1 coupling [14], which results in a 4D echo-by-echo reconstruction that neglects the correlations between echoes. We note that several MRI reconstruction studies have explored TV-like joint smoothness constraints across images with significant contrast differences [2,10,11,23]. Related TV-based regularizers for multi-contrast images have also been proposed in the image processing literature [4,5,18,25]. Finally, low-rank methods [3,6,24] have shown promise. These studies were primarily designed for 2D/3D Cartesian imaging of static organs like the brain with little attention to tissue parameter mapping and have not been extended to 5D non-Cartesian motion-resolved mGRE MRI.

We hypothesize that true 5D processing enables further acceleration by introducing a novel generalization of TV tailored for free-breathing mGRE data. The proposed reconstruction approach exploits data sparsity among the multi-echo images through this generalized TV—referred to as *composite TV* throughout the paper. Integrating composite TV along the echo dimension and temporal TV along the motion-state dimension, our method performs multi-way array (5D) reconstruction, substantially accelerating free-breathing mGRE MRI. Note that classical TV and existing generalizations do not account for the notion of sparsity among multi-echo images because the contrast differences between echoes are not inherently sparse and must be preserved. The challenge is thus to devise a new regularizer such that it preserves image contrast while sparsifying relevant features across echoes. Our contributions/findings are as follows: (a) We introduce a novel composite TV that imposes sparsity between echoes; (b) We derive a solution algorithm compactly expressed via the standard primal-dual hybrid gradient (PDHG) method; and (c) We demonstrate the feasibility of highly accelerated free-breathing, motion-resolved mGRE MRI for liver R2*, PDFF, and quantitative susceptibility mapping (QSM) with high isotropic resolution.

2 Theory

2.1 Problem Formulation and Prior Works

Let Q be the number of fully-sampled k-space samples that meet the Nyquist sampling criterion. Let T be the number of latent respiratory phases (from the end-expiratory to end-inspiratory states) embedded in the incoherently sampled k-space samples $g \in \mathbb{C}^Q$ during free breathing. Let $P_\pi : \mathbb{C}^Q \to \mathbb{C}^Q$ be some permutation matrix that reorders $g = (g^1, \ldots, g^Q)$ to $P_\pi g = \tilde{g} = (\tilde{g}^1, \ldots, \tilde{g}^Q)$ such that the $\tilde{g}^{(n-1)\lfloor Q/T \rfloor + 1}, \ldots, \tilde{g}^{n\lfloor Q/T \rfloor}$ belong to the n-th motion state where $n = 1, \ldots, T$. Notice that the k-space samples that fall into each motion state n is retrospectively undersampled by a factor of T, i.e., the number of k-space measurements for each motion state is approximately $\lfloor Q/T \rfloor$. As g is incoherently sampled uniformly covering the entire k-space (so is $\tilde{g}$), one can utilize CS reconstruction.

Let N be the image size, and let $u \in \mathbb{C}^{N \cdot T}$ be a set of vectorized motion-resolved images to be reconstructed from $\tilde{g}$. Feng et al. [8] cast the problem by incorporating the ℓ_1-norm of the forward difference between motion states into the cost function as follows:

$$\Psi_{\text{Feng}}(u; \tilde{g}) \triangleq ||Au - \tilde{g}||_W^2 + \lambda_m ||\Delta_m^+ u||_1, \tag{1}$$

where $A : \mathbb{C}^{N \cdot T} \to \mathbb{C}^Q$ is the MRI encoding operator, $\Delta_m^+ : \mathbb{C}^{N \cdot T} \to \mathbb{C}^{N \cdot T}$ is the forward difference operator along the motion state, and $\lambda_m > 0$. The weighted least-squares fidelity $||Au - \tilde{g}||_W^2$ is given as $(Au - \tilde{g})^{\mathsf{H}} W (Au - \tilde{g})$ where the density compensation factor is commonly chosen for W and A^{H} is the adjoint of A (the conjugate transpose $A^{\mathsf{H}} = \overline{A}^\top$).

Suppose that we are now given an mGRE k-space data $g_1, \ldots, g_E$ where $g_j \in \mathbb{C}^Q$ for $j = 1, \ldots, E$, and some permutation matrix P_π. Then, $\tilde{g}_j$ is a reordering of g_j by P_π, i.e., $\tilde{g}_j = P_\pi g_j$. Let $u_1, \ldots, u_E \in \mathbb{C}^{N \cdot T}$ be a set of motion-resolved mGRE images. Kang et al. [14] extended (1) as follows:

$$\Psi_{\text{Kang}}(u_1, \ldots, u_E; \tilde{g}_1, \ldots, \tilde{g}_E) \triangleq \sum_{j=1}^{E} ||Au_j - \tilde{g}_j||_W^2 + \lambda_m \sum_{j=1}^{E} ||\Delta_m^+ u_j||_1. \tag{2}$$

Note that an extension of (2) to multi-coil reconstruction is straightforward.

2.2 Proposed Method and Numerical Optimization

Denoting the concatenation of $u_1, \ldots, u_E$ as $\underline{u} \in \mathbb{C}^{N \cdot T \cdot E}$ and the concatenation of $\tilde{g}_1, \ldots, \tilde{g}_E$ as $\underline{\tilde{g}} \in \mathbb{C}^{Q \cdot E}$, we propose to minimize the following cost function

$$\Psi_{\text{Ours}}(\underline{u}; \underline{\tilde{g}}) \triangleq ||A\underline{u} - \underline{\tilde{g}}||_W^2 + \lambda_m ||\Delta_m^+ \underline{u}||_1 + \lambda_e ||(\nabla_{\mathbf{x}} \circ \Delta_e^+)\underline{u}||_{2,1} \tag{3}$$

where $\nabla_{\mathbf{x}} : \mathbb{C}^{N \cdot T \cdot E} \to \mathbb{C}^{(N \cdot T \cdot E) \times 3}$ is the spatial gradient operator and $||\cdot||_{2,1}$ is a matrix norm that maps $\underline{v} \in \mathbb{C}^{(N \cdot T \cdot E) \times 3}$ to $\sum_i \sqrt{\sum_j |\underline{v}_{i,j}|^2}$ for $i = 1, \ldots, N \cdot T \cdot E$

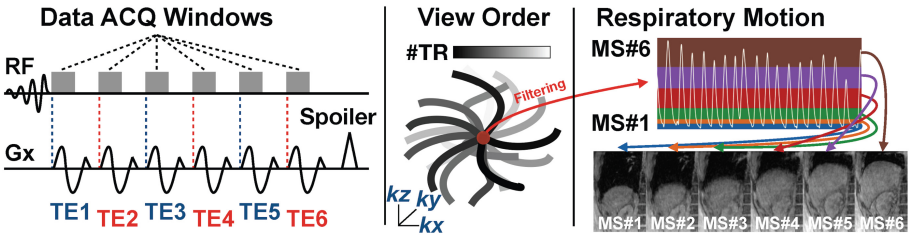


Fig. 1. Illustration of data acquisition and respiratory motion-resolved image reconstruction. Six echoes (TE1 to TE6) are sequentially acquired within a single TR as schematized in the pulse timing diagram. Cones readouts are uniformly distributed in k-space using a pseudorandom view order. The center of k-space (DC) is used to estimate the underlying respiratory motion. The number of motion states to be resolved is set to 6, and the corresponding respiratory motion-resolved images are shown.

and $j = 1, 2, 3$. Extending the operator A and the diagonal matrix W to accommodate the dimensions of the vectorized $\underline{u}$ and $\underline{\tilde{g}}$ is straightforward, and $\lambda_e > 0$ is a regularization parameter. The composite operator inside the $\ell_{2,1}$ norm extracts a contrast-invariant sparse representation—while contrast changes across echoes are not sparse, image edges remain largely sparse. Noise-like undersampling artifacts (as the difference of two Gaussians is also Gaussian), extracted with sparse edges, are selectively suppressed by the $\ell_{2,1}$ norm, effectively preserving sharp edges.

We provide implementation details of a numerical method that solves the cost function (3) in a multi-coil setting. Let C be the number of coil elements; and let $I_\otimes : \mathbb{C}^{N\cdot T\cdot E} \to \mathbb{C}^{(C+3)\cdot N\cdot T\cdot E}$ be a linear operator such that $\underline{u} \mapsto I_\otimes \underline{u} = [\underbrace{\underline{u}^\top, \ldots, \underline{u}^\top}_{C \text{ times}}, \underline{u}^\top, \underline{u}^\top, \underline{u}^\top]^\top$. Then, the minimization problem (3) can be reformulated as the following saddle-point problem [7]:

$$\text{Find } \underline{u}, v \text{ such that } \min_{\underline{u}} \max_{v} \operatorname{Re} \langle (K \circ I_\otimes)\underline{u}, v\rangle - Hv, \tag{4}$$

where the linear operator $K : \mathbb{C}^{(C+3)\cdot N\cdot T\cdot E} \to \underbrace{\mathbb{C}^{Q\cdot E} \times \cdots \times \mathbb{C}^{Q\cdot E}}_{C \text{ times}} \times \mathbb{C}^{N\cdot T\cdot E} \times \mathbb{C}^{(N\cdot T\cdot E)\times 3}$ is defined as

$$K = \begin{bmatrix} W^{1/2}FS_1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & W^{1/2}FS_C & 0 & 0 \\ 0 & \cdots & 0 & \Delta_m^+ & 0 \\ 0 & \cdots & 0 & 0 & \nabla_{\mathbf{x}} \circ \Delta_e^+ \end{bmatrix}; \tag{5}$$

where F is the nonuniform Fourier transform operator, $S_1, \ldots, S_C$ are coil sensitivity maps, and $H : \underbrace{\mathbb{C}^{Q\cdot E} \times \cdots \times \mathbb{C}^{Q\cdot E}}_{C \text{ times}} \times \mathbb{C}^{N\cdot T\cdot E} \times \mathbb{C}^{(N\cdot T\cdot E)\times 3} \to \overline{\mathbb{R}}$,

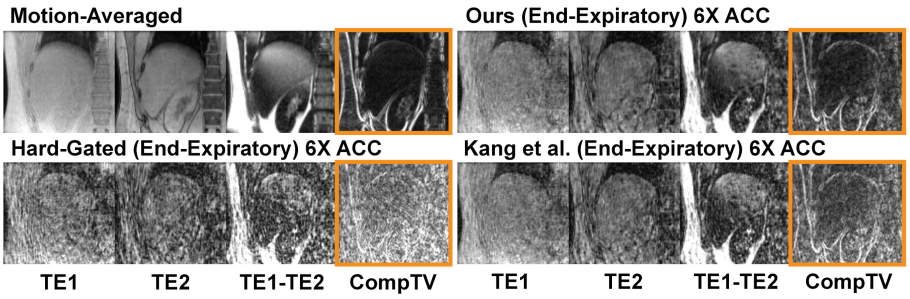


Fig. 2. Desired effect of incorporating $||\nabla_{\mathbf{x}} \circ \Delta_e^+ \cdot ||_{2,1}$; see Sect. 4.

$v := (p_1, \ldots, p_C, q_m, q_{\mathbf{x}}) \mapsto H(p_1, \ldots, p_C, q_m, q_{\mathbf{x}}) = \sum_{c=1}^{C} \mathrm{Re}\langle W^{1/2}\tilde{g}_c, p_c\rangle + (1/2)||p_c||_2^2 + \delta_{Q_m}(q_m) + \delta_{Q_{\mathbf{x}}}(q_{\mathbf{x}})$. Here, $\delta_{Q_m}(q_m) = 0$ if $q_m \in Q_m$ and $+\infty$ elsewhere; $\delta_{Q_{\mathbf{x}}}(\cdot)$ is similarly defined.

The convex sets Q_m and $Q_{\mathbf{x}}$ are given as $Q_m = \{q_m \in \mathbb{C}^{N\cdot T\cdot E} : ||q_m||_\infty \leq \lambda_m\}$ and $Q_{\mathbf{x}} = \{q_{\mathbf{x}} \in \mathbb{C}^{(N\cdot T\cdot E)\times 3} : ||q_{\mathbf{x}}||_\infty \leq \lambda_e\}$, respectively. Then, a saddle point $(\underline{u}, v)$ that satisfies $0 \in I_{\otimes}^H(K^H\hat{v})$ and $\partial H(\hat{v}) \in K(I_{\otimes}\underline{\hat{u}})$ can be found by the primal-dual hybrid gradient (PDHG) method [7] as follows.

$$v^{k+1} = (I + \sigma\partial H)^{-1}(v^k + \sigma K I_{\otimes}\tilde{\underline{u}}^k); \tag{6a}$$

$$\underline{u}^{k+1} = \underline{u}^k - \tau I_{\otimes}^H K^H v^{k+1}; \tag{6b}$$

$$\tilde{\underline{u}}^{k+1} = 2\underline{u}^{k+1} - \underline{u}^k. \tag{6c}$$

The resolvent (proximal) operator $(I+\sigma\partial H)^{-1}$ associated with $\mathrm{Re}\,\langle W^{1/2}\tilde{g}_c, p_c\rangle + (1/2)||p_c||_2^2$w can be derived as

$$\hat{r}_c = \underset{r_c}{\mathrm{argmin}} \frac{||r_c - p_c||_2^2}{2\sigma} + \mathrm{Re}\,\langle W^{1/2}\tilde{g}_c, r_c\rangle + \frac{1}{2}||r_c||_2^2 \iff \hat{r}_c = \frac{p_c - \sigma W^{1/2}\tilde{g}_c}{1+\sigma},$$

for $c = 1, \ldots, C$. The derivation of the projection operators associated with the indicator functions $\delta_{Q_m}(\cdot)$ and $\delta_{Q_{\mathbf{x}}}(\cdot)$ is straightforward using the Moreau decomposition [20] and the soft-thresholding operator.

3 Experimental Methods

MRI Data Acquisition and Imaging Parameters. Upon IRB approval, 3D mGRE cones MRI implemented based on [9,14,17] was performed on 3 healthy subjects using a 3T clinical scanner (GE Healthcare, Waukesha, WI). All subjects underwent the imaging procedure with free breathing and one of the subjects was instructed to perform deep breathing throughout the scan. Imaging parameters for the 3D mGRE cones MRI were: initial TE/ΔTE/TR = 0.032/1.4–1.5/11.4–11.5 ms, #TEs = 6 (#shots = 1, ETL = 6), FA = 3°, in-plane resolution = 2 × 2 mm^2, slice thickness = 2 mm, rBW = 1,106–1,315

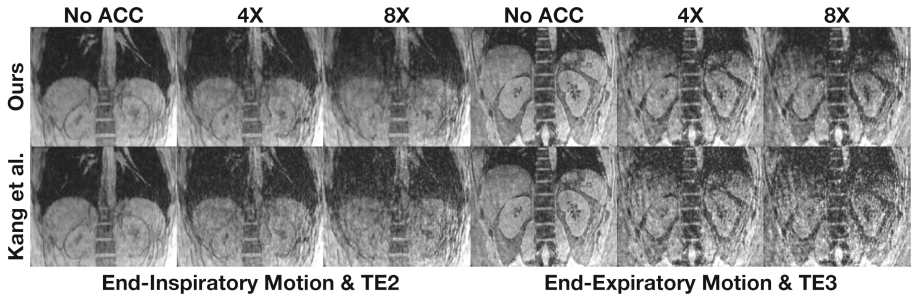


Fig. 3. Reconstructed images of a healthy subject instructed to perform deep breathing with *effective* undersampling factors 6X (i.e., no ACC and 6 motion states), 24X (i.e., 4X ACC and 6 motion states), and 48X (i.e., 8X ACC and 6 motion states) are shown.

Hz/Px, scan time = 5 min 15 s to 7 min 37 s (without acceleration), readout duration = ∼1 ms, and #interleaves = 27,651 to 39,699. Echo spacing for cone acquisition was set at 1.4–1.5 ms to facilitate in- and out-of-phase data acquisition. This was combined with an ultrashort initial TE to capture the rapid T2* signal decay. The data acquisition scheme and reconstruction strategy are illustrated in Fig. 1.

Image Reconstruction and Tissue Parameter Mapping. In our context, *acceleration* (ACC) refers to retrospective undersampling applied prior to motion binning, with the goal of reducing scan time. The fully-sampled k-space raw data was retrospectively undersampled by factors ranging from 2X to 10X, achieved by discarding the corresponding number of readouts. This was done in a manner that allows for straightforward implementation of prospective undersampling. Then, *motion binning* which is another form of retrospective undersampling was carried out subsequent to 2X to 10X ACC. In our paper, the number of motion states was set to 6; therefore, the overall *effective* undersampling factors were from 6X (No ACC) to 60X (10X ACC with 6 motion states).

Implemented reconstruction methods were: 1) motion-averaged gridding, 2) hard-gated gridding applied to each motion state, 3) motion-resolved (2), and 4) the proposed (3) reconstructions. For 2), 3), and 4), the center of k-space of each cones readout was extracted to estimate the respiratory motion [12,13] as illustrated in Fig. 1. For 3) and 4), the PDHG algorithm described in Sect. 2.2 was implemented in Python using the nuFFT routine in SigPy [19]. All reconstructions were run on a 4-way Nvidia A100 GPU machine (80GB of video memory per GPU). These GPUs were utilized concurrently. We considered the following combinations of regularization parameters: $10^{-7} \leq \lambda_m \leq 10^{-4}$ and $10^{-7} \leq \lambda_e \leq 10^{-4}$ for all *effective* undersampling factors = {6X, 12X, 24X, 36X, 48X, 60X} per subject. From our subjective evaluation, we sequentially chose the best $\lambda_m = 7 \times 10^{-6}$ and $\lambda_e = 1 \times 10^{-6}$ for all subjects. Using the standard post-processing pipeline [14,17], R2*, PDFF, and QSM were

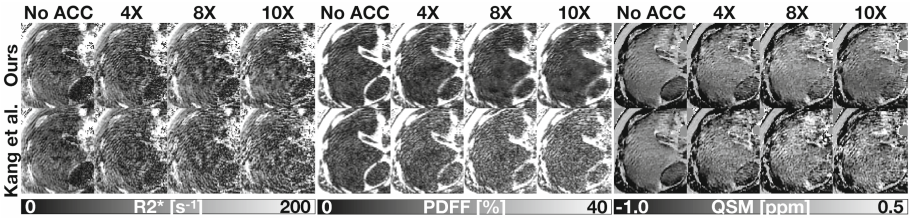


Fig. 4. Quantitative parameter maps of a healthy subject with *effective* undersampling factors 6X (i.e., no ACC and 6 motion states), 24X (4X ACC and 6 motion states), 48X (8X ACC and 6 motion states), and 60X (10X ACC and 6 motion states) are shown.

generated from the reconstructed complex-valued mGRE images. The model $s(\mathbf{x}, t) = (\rho_w(\mathbf{x}) + \rho_f(\mathbf{x})e^{-i2\pi f_{cs}t})e^{-R2^*(\mathbf{x})t - i2\pi f_b(\mathbf{x})t}$ for $t = \mathrm{TE}_1, \ldots, \mathrm{TE}_6$ was fitted to the reconstructed mGRE images with respect to ρ_w (water), ρ_f (fat), $R2^*$, and f_b (B0 field) voxel by voxel, i.e., for all $\mathbf{x}$. Then, PDFF was computed as $|\rho_f(\mathbf{x})|/(|\rho_w(\mathbf{x})| + |\rho_w(\mathbf{x})|)$. For QSM, the fat-referenced nonlinear dipole inversion was performed with the same set of reconstruction parameters and background field removal presented in [14].

Image Quality Assessment and ROI-Based Quantification. Due to the nature of the study, acquiring fully sampled motion-resolved MRI data is infeasible as subjects breathe freely during the 6–8 minute imaging period. The concept of "ground truth" is not applicable in this context; therefore, we used the motion-resolved reconstruction (2), i.e., [14] from 6X undersampled k-space data (no ACC and 6 motion states) was considered as the reference to calculate the relative error, PSNR, SSIM, and MSE. To assess contrast fidelity, a region of interest (ROI) with a radius of 10 voxels was placed in the tissue of interest (liver parenchyma) on three consecutive slices of the computed tissue parameter maps (R2*, PDFF, and QSM), while avoiding large vessels. Then, mean ± SD was calculated for 6–60X undersampling factors and compared between the methods.

4 Experimental Results

The desired effect of incorporating $||\nabla_{\mathbf{x}} \circ \Delta_e^+ \cdot||_{2,1}$ is demonstrated in Fig. 2. The key observation is that the motion-averaged mGRE images (neither accelerated nor retrospectively undersampled for motion binning) exhibit a sparse representation after the $\nabla \circ \Delta_e^+$ operation (image in orange bounding box). The operator Δ_e^+ alone does not provide such a desired representation (third column from the left) due to the difference in contrast between TE1 and TE2 images, which encodes important tissue information. Hard-gated reconstruction of the same subject (6X ACC and 6 motion states, i.e., overall 36X undersampling), exhibits a noisy representation after the $\nabla \circ \Delta_e^+$ operation. This noisy representation is

Table 1. Image quality assessment and ROI-based measurements. PSNR/SSIM (higher the better); MSE (lower the better). R2*/PDFF/QSM are reported as mean ± SD.

Subj #1	Metric	No ACC	2X	4X	6X	8X	10X
Ours	PSNR	53.9393	35.2265	32.8812	32.0261	30.7820	29.0965
	SSIM	0.9951	0.9003	0.8265	0.8021	0.7577	0.7596
	MSE	0.0040	0.3002	0.5151	0.6272	0.8352	1.2313
[14]	PSNR	Inf	35.0066	32.1682	31.1111	29.6637	28.4149
	SSIM	1	0.8946	0.7850	0.7389	0.6738	0.6856
	MSE	0	0.3157	0.6070	0.7743	1.0805	1.4405
Subj #1	Metric	No ACC	4X	8X	10X		
Ours	R2*	44.01 ± 24.87	55.00 ± 28.20	63.80 ± 33.81	72.51 ± 37.31		
	PDFF	6.79 ± 3.00	5.61 ± 3.06	6.91 ± 3.15	5.71 ± 2.96		
	QSM	-0.19 ± 0.09	-0.21 ± 0.13	-0.25 ± 0.18	-0.28 ± 0.17		
[14]	R2*	44.01 ± 25.55	53.26 ± 30.83	64.95 ± 41.61	72.24 ± 46.38		
	PDFF	7.32 ± 3.45	8.59 ± 4.46	10.69 ± 6.14	10.89 ± 5.54		
	QSM	-0.19 ± 0.10	-0.18 ± 0.17	-0.33 ± 0.27	-0.33 ± 0.25		
Subj #2	Metric	No ACC	2X	4X	6X	8X	10X
Ours	PSNR	55.3231	39.2684	37.3378	35.8115	34.8276	34.2043
	SSIM	0.9936	0.9347	0.9055	0.8798	0.8551	0.8477
	MSE	0.0029	0.1183	0.1846	0.2623	0.3290	0.3798
[14]	PSNR	Inf	38.7403	36.3399	34.5567	33.3129	32.7903
	SSIM	1	0.9201	0.8611	0.8086	0.7581	0.7418
	MSE	0	0.1337	0.2323	0.3502	0.4663	0.5260
Subj #2	Metric	No ACC	4X	8X	10X		
Ours	R2*	34.00 ± 19.52	33.69 ± 23.86	43.02 ± 23.79	47.67 ± 24.62		
	PDFF	7.85 ± 3.39	7.98 ± 3.64	7.58 ± 2.97	8.68 ± 3.10		
	QSM	-0.41 ± 0.10	-0.32 ± 0.16	-0.27 ± 0.14	-0.43 ± 0.13		
[14]	R2*	34.38 ± 20.78	37.33 ± 27.67	49.83 ± 33.67	56.97 ± 37.34		
	PDFF	8.75 ± 4.12	11.29 ± 5.69	12.88 ± 6.41	14.14 ± 7.41		
	QSM	-0.38 ± 0.12	-0.28 ± 0.22	-0.20 ± 0.26	-0.23 ± 0.30		
Subj #3	Metric	No ACC	2X	4X	6X	8X	10X
Ours	PSNR	54.0232	39.3264	36.4982	35.1042	34.6145	33.6208
	SSIM	0.9899	0.9399	0.8999	0.8738	0.8626	0.8440
	MSE	0.0040	0.1168	0.2240	0.3087	0.3456	0.4344
[14]	PSNR	Inf	38.6929	35.6087	33.9137	33.2628	32.2565
	SSIM	1	0.9200	0.8450	0.7804	0.7440	0.7035
	MSE	0	0.1351	0.2749	0.4061	0.4718	0.5948
Subj #3	Metric	No ACC	4X	8X	10X		
Ours	R2*	65.38 ± 24.39	90.28 ± 37.10	122.03 ± 56.39	162.09 ± 67.02		
	PDFF	8.69 ± 2.67	9.37 ± 3.36	10.89 ± 4.06	12.16 ± 4.88		
	QSM	-0.30 ± 0.07	-0.32 ± 0.12	-0.11 ± 0.20	-0.27 ± 0.85		
[14]	R2*	65.27 ± 25.28	89.98 ± 43.13	116.61 ± 65.33	150.41 ± 85.18		
	PDFF	9.10 ± 3.42	11.47 ± 5.05	15.50 ± 6.91	17.54 ± 8.11		
	QSM	-0.32 ± 0.08	-0.35 ± 0.16	-0.49 ± 0.32	0.02 ± 1.06		

due to undersampling artifacts across the echoes. Compared to Kang et al. [14] (6X ACC and 6 motion states), the proposed method (6X ACC and 6 motion states) produces a less noisy representation after the $\nabla \circ \Delta_e^+$ operation, as a result of reduced undersampling artifacts.

Reconstructed images of a healthy volunteer instructed to perform deep breathing are shown in Fig. 3 with different undersampling factors. Compared to Kang et al. [14] where image quality rapidly deteriorates due to increased undersampling artifacts, the proposed method demonstrates the ability to maintain visual image quality. The relative differences between the reference (Kang et al. with no ACC) and the proposed method (vs Kang et al.) were 28% (vs 31%), 39% (vs 45%), 46% (vs 55%), 49% (vs 60%), and 54% (vs 67%) for 2X to 10X, confirming the difference in visual image quality. For the other two subjects, the relative differences were: 23% (vs 24%), 32% (vs 35%), 35% (vs 40%), 41% (vs 47%), and 44% (vs 50%) for 2X to 10X (Subj #1); and 25% (vs 27%), 32% (vs 36%), 37% (vs 44%), 42% (vs 51%), and 44% (vs 54%) for 2X to 10X (Subj #2). Furthermore, consistently improved PSNR, SSIM, and MSE in Table 1 agree with the above observations and relative differences. R2*, PDFF, and QSM of a healthy subject with acceleration factors ranging from no ACC to 10X are shown in Fig. 4. Unlike Kang et al. [14] where image quality notably deteriorates, which may be characterized by "salt and pepper" noise, the proposed method shows minimally affected image quality. ROI-based R2*, PDFF, and QSM measurements of healthy volunteers are presented in Table 1. The proposed method exhibits smaller standard deviations across all undersampling factors than Kang et al. [14], suggesting a higher precision level. The visual assessment of Fig. 4 can be confirmed by standard deviations of the measurement.

5 Discussion and Conclusion

Although the feasibility of the proposed method was only demonstrated on a small number of subjects, the proposed novel composite TV and its solution algorithm has enabled 5D CS reconstruction of non-Cartesian mGRE MRI data across a wide range of *effective* undersampling factors, from 6X to 60X. Future work will include a larger patient cohort, particularly those with known iron overload or steatosis to further strengthen the study. In addition, the proposed method will be applied to prospectively undersampled k-space data to validate its robustness in providing reliable liver tissue parameters, comparing with our recent study that demonstrated the feasibility of 5D reconstruction by exploiting sparsity across spatial, motion, and echo dimensions [15].

Acknowledgment. This work was supported in part by National Institutes of Health under Award Numbers: R21EB033985 and R21DK137146.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Armstrong, T., et al.: Free-breathing liver fat quantification using a multiecho 3D stack-of-radial technique. Magn. Reson. Med. **79**(1), 370–382 (2018)
2. Bilgic, B., Goyal, V.K., Adalsteinsson, E.: Multi-contrast reconstruction with Bayesian compressed sensing. Magn. Reson. Med. **66**(6), 1601–1615 (2011)
3. Bilgic, B., et al.: Improving parallel imaging by jointly reconstructing multi-contrast data. Magn. Reson. Med. **80**(2), 619–632 (2018)
4. Blomgren, P., Chan, T.F.: Color TV: total variation methods for restoration of vector-valued images. IEEE Trans. Image Process. **7**(3), 304–309 (1998)
5. Bresson, X., Chan, T.F., et al.: Fast dual minimization of the vectorial total variation norm and applications to color image processing. Inverse Probl. Imaging **2**(4), 455–484 (2008)
6. Bustin, A., Lima da Cruz, G., Jaubert, O., Lopez, K., Botnar, R.M., Prieto, C.: High-dimensionality undersampled patch-based reconstruction (HD-PROST) for accelerated multi-contrast MRI. Magn. Reson. Med. **81**(6), 3705–3719 (2019)
7. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
8. Feng, L., Axel, L., Chandarana, H., Block, K.T., Sodickson, D.K., Otazo, R.: XD-GRASP: golden-angle radial MRI with reconstruction of extra motion-state dimensions using compressed sensing. Magn. Reson. Med. **75**(2), 775–788 (2016)
9. Gurney, P.T., Hargreaves, B.A., Nishimura, D.G.: Design and analysis of a practical 3D cones trajectory. Magn. Reson. Med. Off. J. Int. Soc. Magn. Reson. Med. **55**(3), 575–582 (2006)
10. Haldar, J.P., Liang, Z.P.: Joint reconstruction of noisy high-resolution MR image sequences. In: 2008 5th IEEE International Symposium on Biomedical Imaging: from Nano to Macro, pp. 752–755. IEEE (2008)
11. Haldar, J.P., et al.: Improved diffusion imaging through SNR-enhancing joint reconstruction. Magn. Reson. Med. **69**(1), 277–289 (2013)
12. Jiang, W., et al.: Motion robust high resolution 3D free-breathing pulmonary MRI using dynamic 3D image self-navigator. Magn. Reson. Med. **79**(6), 2954–2967 (2018)
13. Kang, M., Wen, Y., Carl, M., Behr, G., Otazo, R., Kee, Y.: Free-breathing R2* mapping for hepatic iron quantification using respiratory motion-resolved 3D multi-echo UTE cones MRI. In: Proceedings of the 2022 ISMRM & ISMRT Annual Meeting, London, England, UK (Abstract# 1338)
14. Kang, M., Behr, G.G., Jafari, R., Gambarin, M., Otazo, R., Kee, Y.: Free-breathing high isotropic resolution quantitative susceptibility mapping (QSM) of liver using 3D multi-echo UTE cones acquisition and respiratory motion-resolved image reconstruction. Magn. Reson. Med. **90**(5), 1844–1858 (2023)
15. Kang, M., Otazo, R., Behr, G., Kee, Y.: 5D image reconstruction exploiting space-motion-echo sparsity for accelerated free-breathing quantitative liver MRI. Med. Image Anal. **102**, 103532 (2025)
16. Karsa, A., Punwani, S., Shmueli, K.: The effect of low resolution and coverage on the accuracy of susceptibility mapping. Magn. Reson. Med. **81**(3), 1833–1848 (2019)
17. Kee, Y., et al.: Free-breathing mapping of hepatic iron overload in children using 3D multi-echo UTE cones MRI. Magn. Reson. Med. **85**(5), 2608–2621 (2021)
18. Molina, R., Mateos, J., Katsaggelos, A.K., Vega, M.: Bayesian multichannel image restoration using compound gauss-Markov random fields. IEEE Trans. Image Process. **12**(12), 1642–1654 (2003)

19. Ong, F., Lustig, M.: Sigpy: a python package for high performance iterative reconstruction. In: Proceedings of the ISMRM 27th Annual Meeting, Montreal, Quebec, Canada, vol. 4819 (2019)
20. Parikh, N., Boyd, S., et al.: Proximal algorithms. Found. Trends® Optim. **1**(3), 127–239 (2014)
21. Schneider, M., et al.: Free-breathing fat and R2* quantification in the liver using a stack-of-stars multi-echo acquisition with respiratory-resolved model-based reconstruction. Magn. Reson. Med. **84**(5), 2592–2605 (2020)
22. Tan, Z., et al.: Free-breathing liver fat, R2* and B0 field mapping using multi-echo radial flash and regularized model-based reconstruction. IEEE Trans. Med. Imaging **42**(5), 1374–1387 (2022)
23. Trzasko, J., Manduca, A.: Group sparse reconstruction of vector-valued images. In: Proceedings of the International Society for Magnetic Resonance in Medicine, vol. 19, p. 2839 (2011)
24. Trzasko, J.D.: Exploiting local low-rank structure in higher-dimensional MRI applications. In: Wavelets and Sparsity XV, vol. 8858, pp. 551–558. SPIE (2013)
25. Yang, J., Yin, W., Zhang, Y., Wang, Y.: A fast algorithm for edge-preserving variational multichannel image restoration. SIAM J. Imag. Sci. **2**(2), 569–592 (2009)
26. Zhou, D., Cho, J., Zhang, J., Spincemaille, P., Wang, Y.: Susceptibility underestimation in a high-susceptibility phantom: dependence on imaging resolution, magnitude contrast, and other parameters. Magn. Reson. Med. **78**(3), 1080–1086 (2017)

AffinityUMamba: Uncertainty-Aware Medical Image Segmentation via Probabilistic Weak Supervision Beyond Gold-Standard Annotations

Yukun Zhang[1], Guisheng Wang[2], William Henry Nailon[3], and Kun Cheng[1(✉)]

[1] Beijing University of Posts and Telecommunications, Beijing, China
{zyk,kcheng}@bupt.edu.cn
[2] Department of Radiology, Third Medical Center of Chinese PLA General Hospital, Beijing, China
wangguisheng@301hospital.com.cn
[3] Department of Oncology Physics and The School of Engineering, NHS Lothian and The University of Edinburgh, Edinburgh, UK
w.nailon@ed.ac.uk

Abstract. Owing to its superior soft tissue contrast, Magnetic Resonance Imaging (MRI) has become a cornerstone modality in clinical practice. This prominence has driven extensive research on MRI-based segmentation, supported by the proliferation of publicly available benchmark datasets. Despite employing multi-expert consensus protocols to ensure annotation quality in public datasets, the inherent label noise, particularly prevalent at lesion boundary regions, remains unavoidable. To address this fundamental challenge, we introduce a novel machine learning paradigm that reframes dataset annotations as probabilistic weak supervision rather than deterministic gold standards. We proposed AffinityUMamba, a novel dual-branch Unet-like framework that synergistically integrates convolutional operations with state space models, leveraging local feature coherence and global contextual agreement. And a Local Affinity-guided Label Refinement (LALR) module to identify potential noisy labels in the training data and produce refined pseudo labels. A unified uncertainty constraint paradigm combining margin-based logit smoothing with local affinity refinement, enabling simultaneous optimization of segmentation accuracy and confidence calibration. Training is stabilized through a composite objective combining topological preservation constraints with margin-aware uncertainty penalization, enabling joint optimization of structural coherence and detail fidelity. We comprehensively evaluated the proposed method on 12 public datasets spanning multiple modalities: 10 MRI, 1 Ultrasound, and 1 CT. The results of our experiments demonstrate improved segmentation performance and reduced prediction uncertainty.

Keywords: Medical Image Segmentation · Uncertainty · Weak Supervision

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 35–45, 2026.
https://doi.org/10.1007/978-3-032-04947-6_4

1 Introduction

Medical image analysis has evolved from rule-based workflows to deep learning paradigms, driven by architectures like nnU-Net [9]. While early research emphasized texture classification [4] and radiomics [13], contemporary efforts focus on lesion segmentation where precision impacts clinical decisions [23]. Magnetic Resonance Imaging (MRI) dominates soft tissue analysis due to unparalleled contrast resolution for neurological [5], cardiac [3], and abdominal structures [24]. However, its sensitivity to tissue heterogeneity introduces critical challenges: ambiguous lesion boundaries and intensity non-uniformities propagate annotation inconsistencies, even in multi-expert consensus protocols [11].

From our perspective, the public datasets utilize voter/averaging manner to obtain a sub-optimal gold standard, while conventional supervised learning frameworks treat annotations as deterministic gold standards, ignoring the inherent uncertainty introduced in medical image manual labeling. This approach propagates toxic training errors, particularly at boundary regions where the highest inter-observer variability is commonly discovered. Gal et al. [12] defined the aleatoric and epistemic uncertainty that were introduced in the common pipeline of deep learning segmentation models. Many researchers focus on the model architecture and training schemes to reduce epistemic (model) uncertainty, but very limited work pays attention to aleatoric (data) uncertainty in medical image segmentation.

Our philosophy here is to acknowledge the gaze annotation effort from experienced oncologists/doctors at the homogeneous interior of the region of interest (ROI), while delivering discriminative knowledge through end-to-end weakly supervised training at inhomogeneous boundary regions. Affinity has been proven beneficial for improving segmentation in weakly supervised semantic segmentation [1], which refers to the spatial and semantic relationships between neighboring pixels that share similar characteristics in medical images. A multi-task framework that uses auxiliary tasks and cross-task affinity learning to enhance weakly-supervised semantic segmentation is proposed using only image-level ground-truth labels [27]. Another end-to-end weakly-supervised semantic segmentation method based on Transformers, leverages attention mechanisms to learn affinity and refine pseudo labels for improved segmentation accuracy [22]. State Space Sequence Models (SSMs) Mamba [6], provide an effective alternative by modeling long-range dependencies and complex relationships in sequential data. When combined with CNNs, SSMs are capable of retaining local details and boundary features while simultaneously capturing global information, which proves advantageous in addressing complex boundaries and potential label noise [14].

To address these challenges, we propose a paradigm shift from deterministic annotation to probabilistic weak supervision, weakening pixel-level supervision from ground truth labels at boundary regions. The main contributions of this work are threefold: (1) We proposed AffinityUMamba, a novel dual-branch Unet-like framework that synergistically integrates convolutional operations with state space models, leveraging local feature coherence and global contextual agree-

ment; (2) We proposed the Local Affinity-guided Label Refinement (LALR) module to identify potential noisy label in the training data and produce refined pseudo labels; (3) A unified uncertainty constraint paradigm combining margin-based logit smoothing with local affinity refinement, enabling simultaneous optimization of segmentation accuracy and confidence calibration.

2 Method

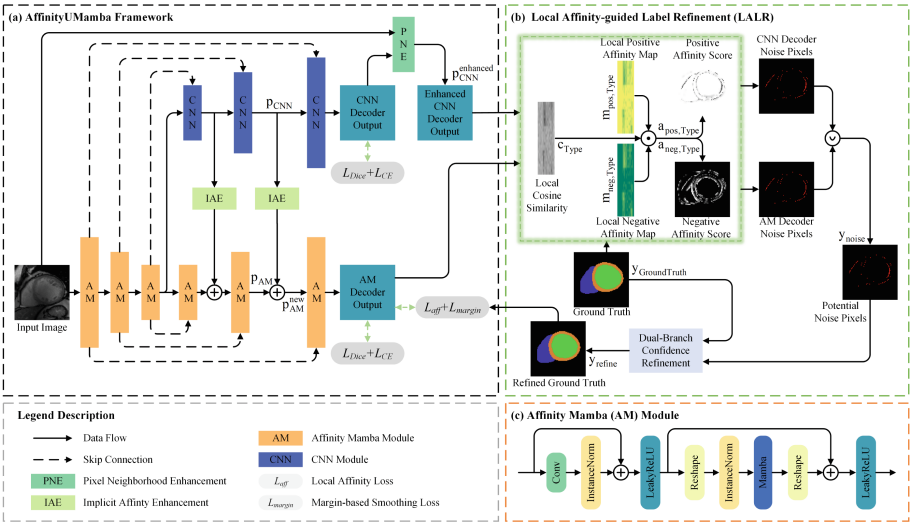


Fig. 1. Illustration of our AffinityUMamba framework.

Our framework integrates collaborative modules for probabilistic weak supervision, as shown in Fig. 1. Let $x \in \mathbb{R}^{H\times W\times D}$ be the input image, and $y \in \mathbb{R}^{H\times W\times D}$ the corresponding ground truth or predicted category map. The feature map is $p \in \mathbb{R}^{C\times H\times W\times D}$, where C is the number of classes, and H, W, and D are the height, width, and depth of the image, respectively. Other variables include the local affinity map $m \in \mathbb{R}^{27\times H\times W\times D}$, where $k = 3$ denotes the size of the local $3 \times 3 \times 3$ neighborhood, and the neighborhood $\mathcal{N}^i(k)$ refers to the region at position i. Additionally, the index j refers to a voxel within $\mathcal{N}^i(k)$, affinity scores $a \in \mathbb{R}^{H\times W\times D}$, and the local cosine similarity $c \in \mathbb{R}^{27\times H\times W\times D}$. To ensure numerical stability in the computations, a small value ϵ is added during division operations.

2.1 Implicit Affinity Enhancement (IAE)

Affinity Mamba (AM) module as shown in Fig. 1(c). The IAE module enhances the feature discriminability of the AM branch by leveraging pixel affinity relationships from the CNN branch, aligning CNN's local semantics with AM's global semantics through adaptive neighborhood interactions.

$$p_{\mathrm{AM}}^{i,\mathrm{new}} = \sum_{j \in \mathcal{N}^i(k)} \mathbb{I}\left(p_{\mathrm{CNN}}^{j,i} \geq \frac{1}{|\mathcal{N}^i(k)|} \sum_{j \in \mathcal{N}^i(k)} p_{\mathrm{CNN}}^{j,i} \right) p_{\mathrm{AM}}^{j} + p_{\mathrm{AM}}^{i} \tag{1}$$

where $\mathbb{I}$ is the indicator function, returning 1 if the condition is true and 0 otherwise. The mechanism encodes semantic relationships among neighboring pixels by comparing each pixel's value to the neighborhood's average, enhancing the corresponding features in AM based on these implicit affinities. This dual-branch Unet co-regularization framework improves boundary delineation by ensuring only statistically significant correlations influence the update, ensuring reliable tissue differentiation despite intensity inhomogeneity.

2.2 Pixel Neighborhood Enhancement (PNE)

We assume that image signals remain relatively reliable despite noisy annotations, containing essential tissue boundary information. This module enhances local continuity in CNN decoder output features by directly leveraging multi-scale intensity relationships from the input image, addressing ground truth inaccuracies caused by inter-observer variability. By constructing adaptive Gaussian kernels at different scales, it explicitly captures pixel correlations based on raw intensity patterns, complementing CNN's inherent local modeling strengths. The computation involves:

$$\begin{aligned} \hat{x}^{j,i} &= \exp\left(-\frac{|x^{j,i} - x^i|^2}{\sigma + \epsilon} \right), \quad \hat{x}^{j,i} \leftarrow \frac{\hat{x}^{j,i}}{\sum_{j \in \mathcal{N}^i(k)} \hat{x}^{j,i}} \\ p_{\mathrm{CNN}}^{i,enhanced} &= \sum_{k \in \{3,5,7\}} \frac{1/\sqrt{k}}{\sum_{k'} 1/\sqrt{k'}} \sum_{j \in \mathcal{N}^i(k)} \hat{x}^{j,i} p_{\mathrm{CNN}}^{j,i} + p_{\mathrm{CNN}}^{i} \end{aligned} \tag{2}$$

It is expected to maintain structural coherence while improving incomplete boundaries which may be caused by partial volume effects.

2.3 Local Affinity-Guided Label Refinement (LALR)

The LALR module derives the noise label suggestion from both CNN and AM decoders, using affinity-guided annotation inconsistencies where local feature coherence and global contextual agreement are combined.

We quantify the intrinsic affinity credibility between neighboring features through local cosine similarity, which evaluates the semantic consistency of the feature. This is formulated as:

$$c_{\mathrm{Type}}^{j,i} = \frac{p_{\mathrm{Type}}^{j,i} \cdot p_{\mathrm{Type}}^{i}}{|p_{\mathrm{Type}}^{j,i}| \cdot |p_{\mathrm{Type}}^{i}| + \epsilon}, \quad \mathrm{Type} \in \{\mathrm{CNN}, \mathrm{AM}\} \tag{3}$$

The dual-decoder verification mechanism operates through dynamically generated local affinity maps that encode the semantic relationships among neighboring pixels:

$$m_{\mathrm{pos,Type}}^{j,i} = \mathbb{I}(y^j = y_{\mathrm{Type}}^{i}), \quad m_{\mathrm{neg,Type}}^{j,i} = 1 - m_{\mathrm{pos,Type}}^{j,i} \tag{4}$$

where y^j denotes the category at position i from the ground truth (for LALR in Sect. 2.3) or refined ground truth (for following Local Affinity Loss in Sect. 2.4), and y^i_{Type} denotes the category at position i from the CNN decoder prediction, AM decoder prediction or refined ground truth.

Potential noise pixels are identified through the affinity score assessment, which detects pixels where both decoders exhibit conflicting affinity patterns-low confidence in same category feature alignment but high confidence in different category dissimilarity. The calculations are performed as follows:

$$\begin{aligned} a^i_{\text{pos,Type}} &= \frac{\sum_{j\in\mathcal{N}^i(k)} c^{j,i}_{\text{Type}} m^{j,i}_{\text{pos,Type}}}{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{pos,Type}} + \epsilon}, \quad a^i_{\text{neg,Type}} = \frac{\sum_{j\in\mathcal{N}^i(k)} c^{j,i}_{\text{Type}} m^{j,i}_{\text{neg,Type}}}{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{neg,Type}} + \epsilon} \\ y^i_{\text{noise}} &= \bigcup_{\text{Type}} \left(a^i_{\text{pos,Type}} < \frac{1}{2} \right) \cap \left(a^i_{\text{neg,Type}} > \frac{1}{2} \right) \end{aligned} \tag{5}$$

The refinement process employs entropy-weighted decision fusion to integrate dual-branch confidence:

$$y^i_{\text{refine}} = y^i_{\text{noise}} \cdot \mathbb{I}(E^i_{\text{CNN}} > E^i_{\text{AM}}) \cdot y^i_{\text{AM}} + (1 - y^i_{\text{noise}}) \cdot y^i_{\text{GroundTruth}} \tag{6}$$

where $E^i_{\text{Type}} = \text{Entropy}(p^i_{\text{Type}})$ measures prediction uncertainty. When the CNN decoder branch exhibits higher entropy (greater uncertainty in local predictions), the AM decoder branch's prediction y^i_{AM} is prioritized for correction.

2.4 Loss Function

Local Affinity Loss enhances feature discriminability by enforcing semantic consistency between positive/negative samples in the AM decoder predictions using the refined ground truths. When computing the local positive and negative affinity maps, $m^{j,i}_{\text{pos,Type}}$ and $m^{j,i}_{\text{neg,Type}}$, y^j represents the label at position i from the refined ground truth y^i_{refine} (as detailed in Sect. 2.3), ensuring semantic consistency:

$$\mathcal{L}_{\text{aff}} = \sum_i \frac{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{pos,refine}} m^{j,i}_{\text{neg,AM}}}{\sqrt{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{pos,refine}}}} + \frac{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{neg,refine}} m^{j,i}_{\text{pos,AM}}}{\sqrt{\sum_{j\in\mathcal{N}^i(k)} m^{j,i}_{\text{neg,refine}}}} \tag{7}$$

Margin-Based Smoothing Loss inspired by previous work [19], we introduce dynamically adjusted, variance-aware boundary constraint, reducing extreme prediction values and alleviating overconfidence in noisy boundaries:

$$\begin{aligned} v^i &= \sqrt{\text{ReLU}\left(\sum_{j\in\mathcal{N}^i(k)} |p^{j,i}_{\text{AM}}|^2 - |\sum_{j\in\mathcal{N}^i(k)} p^{j,i}_{\text{AM}}|^2 \right)} \\ \mathcal{L}_{\text{margin}} &= \sum_i \text{ReLU}\left(\max_c(p^i) - p^i - 8 \cdot \frac{v^i - \min(v^i)}{\max(v^i) - \min(v^i)} \right) \end{aligned} \tag{8}$$

where constant 8 is set following [19].

Overall Objective: The two losses complement the traditional Dice-CE loss:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}}_{\text{Supervision}} + \underbrace{0.1(\mathcal{L}_{\text{Dice}}^{\text{refine}} + \mathcal{L}_{\text{CE}}^{\text{refine}})}_{\text{Refined Supervision}} + 0.1\mathcal{L}_{\text{margin}} + 0.01\mathcal{L}_{\text{aff}} \tag{9}$$

'Supervision' uses ground truth to guide CNN and AM decoder predictions, while 'Refined Supervision' uses refined ground truth for both branches. This joint optimization balances model calibration and discrimination performance.

3 Experiments and Results

3.1 Datasets and Implementations

We validated our model on 12 public medical image segmentation datasets, including ACDC [3] (MRI: Left Ventricle, Right Ventricle, Myocardium), iSeg2017 [25] (MRI: Cerebrospinal Fluid, Gray Matter, White Matter), Brats2020 [18] (MRI: Whole Tumor, Tumor Core, Enhancing Tumor), ISLES2022 [8] (MRI: Ischemic Stroke), PROMISE2012 [16] (MRI: Prostate), MyoPS2020 [29] (MRI: Left Ventricle, Right Ventricle, etc.), MSD [2] Heart (MRI: Left Atrium), MSD [2] Hippocampus (MRI: Anterior Hippocampus, Posterior Hippocampus), AMOS2022 [10] (MRI: Liver, Right Kidney, etc.), ATLAS2022 [15] (MRI: Liver, Hepatic Tumor), CuRIOUS2022 [26] (Ultrasound: Tumor), and FLARE2021 [17] (CT: Liver, Kidney, Spleen, Pancreas). Datasets were split into 4:1:1 training, validation, and test sets. The training was performed on an NVIDIA GeForce RTX 4090 GPU using the SGD optimizer with a 0.01 learning rate for 500 epochs. During inference, the segmentation result is the AM branch output. Performance was evaluated using Dice Coefficient (DSC), 95% Hausdorff Distance (HD95) for discrimination, and Expected Calibration Error (ECE) and Classwise Expected Calibration Error (CECE) with $M = 15$ bins focusing on foreground regions to highlight method differences.

3.2 Results

As shown in Table 1, AffinityUMamba outperforms other methods. For instance, the largest DSC improvement is on the ATLAS2022 dataset (+0.030), compared to nnUNet's 0.746. In terms of HD95, the greatest reduction occurs on the FLARE dataset (−14.058), compared to nnUNet's 21.785. This is primarily due to our method's multi-module collaborative design: the AM module models global anatomical dependencies using Mamba with linear time complexity, capturing global topological constraints; the PNE aggregates neighborhood grayscale features adaptively with a multi-scale Gaussian kernel to improve boundary smoothness and connectivity; the LALR corrects noisy labels via a dual-branch confidence selection mechanism. Additionally, the IAE module dynamically weights neighborhood information based on feature similarity thresholds, alleviating boundary blurring in low-contrast regions.

Table 1. Segmentation performance comparison (Dice Coefficient [DSC] and 95% Hausdorff Distance [HD95]) across 12 datasets. **Bold**: best results; Underlined: second-best.

Dataset	nnUNet [9]		AttentionUNet [21]		UNet++ [28]		SegResNet [20]		SwinUNETR [7]		Ours	
	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95
ACDC	0.913	1.318	0.844	2.162	0.897	1.704	0.874	2.168	0.876	1.961	**0.925**	**1.122**
iSeg2017	0.919	1.138	0.912	1.138	0.914	1.192	0.908	1.197	0.909	1.192	**0.922**	**1.137**
Brats2020	0.847	7.437	0.824	10.258	0.847	8.219	0.831	8.954	0.846	9.834	**0.852**	**5.197**
ISLES2022	0.779	4.331	0.787	**3.281**	0.765	4.109	0.760	4.126	0.768	5.516	**0.816**	3.999
PROMISE2012	0.872	3.645	0.854	4.531	0.862	4.462	0.840	6.201	0.844	5.876	**0.882**	**3.490**
MSD Hippocampus	0.886	1.413	0.883	1.474	0.881	1.439	0.872	1.499	0.883	1.468	**0.887**	**1.367**
ATLAS2022	0.746	20.156	0.701	30.817	0.725	27.663	0.694	30.147	0.700	28.665	**0.776**	**14.609**
CuRIOUS2022	0.782	36.928	0.742	56.612	0.767	46.608	0.764	35.916	0.763	39.452	**0.802**	**35.221**
MyoPS2020	0.702	11.276	0.629	16.448	0.630	15.636	0.641	15.591	0.654	14.291	**0.716**	**10.682**
AMOS2022	0.865	5.618	0.821	14.250	0.813	10.408	0.796	22.308	0.826	8.358	**0.879**	**5.087**
MSD Heart	0.930	3.106	0.931	3.782	0.933	3.606	0.923	4.405	0.927	3.859	**0.934**	**3.031**
FLARE	0.913	21.785	0.871	47.009	0.845	55.506	0.835	62.160	0.851	37.776	**0.927**	**7.727**

Table 2. Calibration performance comparison (Expected Calibration Error [ECE] and Classwise ECE [CECE]) across 12 datasets.

Dataset	nnUNet [9]		AttentionUNet [21]		UNet++ [28]		SegResNet [20]		SwinUNETR [7]		Ours	
	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE	ECE	CECE
ACDC	0.0613	0.0314	0.0777	0.0837	0.1178	0.0816	0.1726	0.1222	0.0531	0.0753	**0.0493**	**0.0267**
iSeg2017	0.0412	0.0310	0.0542	0.0384	0.1487	0.0996	0.0404	0.0327	0.0367	0.0298	**0.0345**	**0.0284**
Brats2020	0.1682	0.3114	0.1604	0.3154	0.1491	0.3130	0.1610	0.3144	0.1669	0.3161	**0.1232**	**0.3105**
ISLES2022	0.2150	0.2270	0.1844	0.1984	0.1314	0.1972	0.1999	0.2838	0.1722	0.2695	**0.1240**	**0.1746**
PROMISE2012	0.1164	**0.1569**	0.1031	0.2802	0.1083	0.2794	0.1593	0.3406	0.1231	0.2880	**0.0753**	0.1653
MSD Hippocampus	0.0904	0.0663	0.0938	0.1202	0.0946	0.1299	0.1382	0.1432	**0.0552**	0.0943	0.0829	**0.0623**
ATLAS2022	0.1794	0.1693	0.1971	0.1866	0.1694	0.1728	0.1492	0.1633	0.1731	0.1796	**0.0810**	**0.1002**
CuRIOUS2022	0.1346	0.3875	0.1458	0.3139	0.1471	0.4051	0.1539	0.4470	0.1394	0.4056	**0.0788**	**0.2708**
MyoPS2020	0.1530	0.0545	0.0571	0.0543	0.1098	0.0539	**0.0499**	0.0542	0.1239	0.0549	0.1535	**0.0538**
AMOS2022	0.3509	0.0509	0.4016	0.0615	0.3643	0.0542	0.4190	0.0642	0.3077	0.0457	**0.0966**	**0.0126**
MSD Heart	0.0258	0.0398	0.0533	0.1223	0.0686	0.1276	0.0725	0.1537	0.0574	0.1155	**0.0251**	**0.0381**
FLARE2021	0.2117	0.1820	**0.0980**	**0.0486**	0.3450	0.2710	0.3564	0.2667	0.3177	0.1640	0.1844	0.3782

AffinityUMamba also demonstrates significant calibration improvements across datasets, as shown in Table 2. The largest ECE reduction occurs on the AMOS2022 dataset (−0.2111), compared to SwinUNETR's 0.3077. For CECE, the largest decrease also occurs on AMOS2022 (−0.0331), compared to SwinUNETR's 0.0457. This advantage stems from key design elements: Margin-based Smoothing Loss dynamically perceives boundary constraints, suppressing over-confidence in noisy boundaries, while the Local Affinity Loss enforces semantic consistency between pseudo-labels and refined ground truths. The LALR successfully avoids over-fitting to the toxic noisy labels, consequently reducing the prediction uncertainty originating from data uncertainty and improving the model segmentation performance.

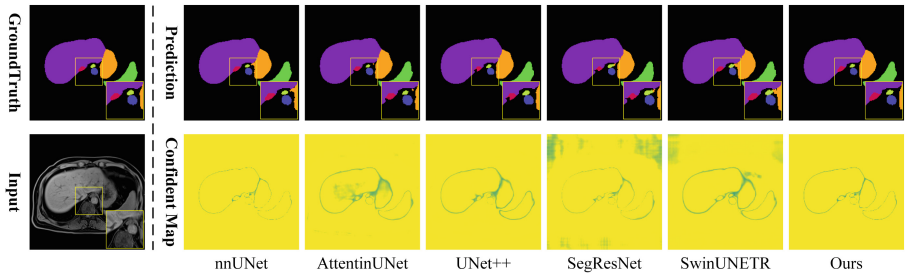


Fig. 2. Segmentation outputs and confidence maps on the AMOS2022 dataset. Confidence maps represent predictive entropy on a 0-1 scale, with yellow indicating higher confidence. Anatomical labels: liver (purple), gallbladder (orange), spleen (green), pancreas (red), kidney (blue). (Color figure online)

Table 3. Ablation study evaluating module contributions on the ACDC dataset.

Target	Discrimination								Calibration	
	Right Ventricle		Myocardium		Left Ventricle		Avg.			
	DSC	HD95	DSC	HD95	DSC	HD95	DSC	HD95	ECE	CECE
AffinityUMamba (w/o LALR, IAE, PNE)	0.9015	1.4570	0.9043	1.1439	**0.9565**	**1.0207**	0.9214	1.2005	0.0496	0.0291
AffinityUMamba (w/o LALR)	0.9089	1.3045	0.9023	1.1398	0.9543	1.0407	0.9219	1.1617	0.0541	0.0316
AffinityUMamba$_{ours}$	**0.9143**	**1.1898**	**0.9047**	**1.1353**	0.9548	1.0414	**0.9246**	**1.1222**	**0.0493**	**0.0267**

In Fig. 2, visual predictions on the AMOS2022 dataset show that our method produced the best boundary delineation, especially for hard regions such as the gallbladder (orange) and pancreas (red), with a significant reduction in misclassified pixels at organ boundary. The significantly calibrated prediction uncertainty is demonstrated on the pixel-wise entropy confidence map, where high uncertainty is precisely concentrated on complete organ boundaries with a clean background at homogeneous regions.

Table 3 presents ablation study results. Compared to the baseline method (AffinityUMamba w/o LALR, IAE, PNE), adding IAE and PNE improves discriminative performance slightly (DSC: 0.9214 → 0.9219, HD95: 1.2005 → 1.1617), but calibration performance decreases (ECE: 0.0496 → 0.0541, CECE: 0.0291 → 0.0316), due to the absence of the LALR mechanism. Despite more ambiguous boundaries in the right ventricle and myocardium, segmentation performance still improves (DSC: Right Ventricle 0.9089 → 0.9143, Myocardium 0.9023 → 0.9047, HD95: Right Ventricle 1.3045 → 1.1898, Myocardium 1.4398 → 1.1353). In contrast, the left ventricle, with clearer boundaries, shows limited improvement (DSC: 0.9543 → 0.9548, HD95: 1.1617 → 1.1222).

4 Conclusion

This study presents AffinityUMamba, a weakly supervised framework that addresses annotation uncertainty in medical image segmentation through weak

supervision. By integrating convolutional networks with state space models, our approach effectively captures anatomical dependencies while mitigating boundary ambiguities. The synergistic architecture combines global context modeling with adaptive local refinement, demonstrating exceptional performance across 12 multi-modal datasets. Experimental results show significant improvements in both discriminative and calibration performance compared to existing state-of-the-art methods. The framework resolves annotation-caused data uncertainty, proving particularly valuable for MRI soft tissue analysis while maintaining promising generalizability across CT and ultrasound modalities. This work establishes a new insight and a reliability-aware paradigm for medical image segmentation using public datasets, bridging the gap between computational models and practical clinical decision-making needs.

Acknowledgments. This study was supported by the Special Scientific Research Program for Capital Health Development (grant number 2024-1-5041).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4981–4990 (2018)
2. Antonelli, M., et al.: The medical segmentation decathlon. Nat. Commun. **13**(1), 4128 (2022)
3. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
4. Castellano, G., Bonilha, L., Li, L., Cendes, F.: Texture analysis of medical images. Clin. Radiol. **59**(12), 1061–1069 (2004)
5. Despotović, I., Goossens, B., Philips, W.: MRI segmentation of the human brain: challenges, methods, and applications. Comput. Math. Methods Med. **2015**(1), 450341 (2015)
6. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) BrainLes 2021. LNCS, vol. 12962, pp. 272–284. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-08999-2_22
8. Hernandez Petzsche, M.R., et al.: Isles 2022: a multi-center magnetic resonance imaging stroke lesion segmentation dataset. Sci. Data **9**(1), 762 (2022)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
10. Ji, Y., et al.: AMOS: a large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Adv. Neural. Inf. Process. Syst. **35**, 36722–36732 (2022)

11. Jungo, A., et al.: On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 682–690. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_77
12. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? Adv. Neural. Inf. Process. Syst. **30** (2017)
13. Lambin, P., et al.: Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer **48**(4), 441–446 (2012)
14. Li, H., Qin, P., Yuan, X., Chen, Z., et al.: HCMA-UNet: a hybrid CNN-mamba UNet with inter-slice self-attention for efficient breast cancer segmentation. arXiv preprint arXiv:2501.00751 (2025)
15. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Sci. Data **9**(1), 320 (2022)
16. Litjens, G., et al.: Evaluation of prostate segmentation algorithms for MRI: the promise12 challenge. Med. Image Anal. **18**(2), 359–373 (2014)
17. Ma, J., et al.: Fast and low-GPU-memory abdomen CT organ segmentation: the flare challenge. Med. Image Anal. **82**, 102616 (2022)
18. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATs). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
19. Murugesan, B., Liu, B., Galdran, A., et al.: Calibrating segmentation networks with margin-based label smoothing. Med. Image Anal. **87**, 102826 (2023)
20. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_28
21. Oktay, O., Schlemper, J., Folgoc, L.L., et al.: Attention U-net: learning where to look for the pancreas. ArXiv abs/1804.03999 (2018)
22. Ru, L., Zhan, Y., Yu, B., Du, B.: Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16846–16855 (2022)
23. Sherer, M.V., et al.: Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: a critical review. Radiother. Oncol. **160**, 185–191 (2021)
24. Tian, Z., Liu, L., Zhang, Z., Fei, B.: PSNet: prostate segmentation on MRI based on a convolutional neural network. J. Med. Imaging **5**(2), 021208 (2018)
25. Wang, L., et al.: Benchmark on automatic six-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. IEEE Trans. Med. Imaging **38**(9), 2219–2230 (2019)
26. Xiao, Y., Fortin, M., Unsgård, G., Rivaz, H., Reinertsen, I.: Resect: a clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries. Med. Phys. **44**(7), 3875–3882 (2017)
27. Xu, L., Ouyang, W., Bennamoun, M., Boussaid, F., Sohel, F., Xu, D.: Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6984–6993 (2021)

28. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
29. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans. Pattern Anal. Mach. Intell. **41**(12), 2933–2946 (2018)



Anomaly Detection by Clustering DINO Embeddings Using a Dirichlet Process Mixture

Nico Schulthess(✉) and Ender Konukoglu

Computer Vision Lab, ETH Zurich, Zurich, Switzerland
nico.schulthess@vision.ee.ethz.ch

Abstract. In this work, we leverage informative embeddings from foundational models for unsupervised anomaly detection in medical imaging. For small datasets, a memory-bank of normative features can directly be used for anomaly detection which has been demonstrated recently. However, this is unsuitable for large medical datasets as the computational burden increases substantially. Therefore, we propose to model the distribution of normative DINOv2 embeddings with a Dirichlet Process Mixture model (DPMM), a non-parametric mixture model that automatically adjusts the number of mixture components to the data at hand. Rather than using a memory bank, we use the similarity between the component centers and the embeddings as anomaly score function to create a coarse anomaly segmentation mask. Our experiments show that through DPMM, embeddings of DINOv2, despite being trained on natural images, achieve very competitive anomaly detection performance on medical imaging benchmarks and can do this while at least halving the computation time at inference. Our analysis further indicates that normalized DINOv2 embeddings are generally more aligned with anatomical structures than unnormalized features, even in the presence of anomalies, making them great representations for anomaly detection. The code is available at https://github.com/NicoSchulthess/anomalydino-dpmm.

Keywords: Unsupervised Anomaly Detection · Dirichlet Process Mixture Model · Foundation Model

1 Introduction

Anomaly detection focuses on identifying samples that deviate from the norm, such as the detection of lesions in medical imaging or defects in industrial inspection. Here, the norm is defined by normal example samples, for instance by images of healthy volunteers in medical contexts or intact objects in industrial settings. Of particular interest is the unsupervised setting, where no anomalous samples are used during training. This avoids introducing unwanted biases towards certain types of anomalies, does not introduce any potential class imbalances, and alleviates the need for annotating anomalous samples for each modality and anatomy.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 46–56, 2026.
https://doi.org/10.1007/978-3-032-04947-6_5

One line of work to tackle unsupervised anomaly detection (UAD) are reconstruction-based approaches, where a generative model is trained to reconstruct normal images. The underlying hypothesis is that since the model is trained to reconstruct normal images well, it will do a poor job in anomalous regions and the reconstruction error will highlight such regions. Popular choices for the reconstruction model are adaptations of autoencoders [6,24], variational autoencoders [7,8,34], masked autoencoders [27], generative adversarial networks [29], and denoising diffusion probabilistic models [31]. Instead of reconstructing images, a feature reconstruction objective can be used in a knowledge distillation setup where a student network is trained on normal data to predict similar features to a pretrained teacher network [23]. Then, the embeddings are assumed to differ between the student and teacher networks for anomalous data while they are similar for normal data.

Another line of work aims to model the distribution of the features extracted from normal samples. One choice is to use normalizing flow models to map the normal patch feature distribution to a tractable distribution [14]. However, normalizing flow models can assign high probabilities to anomalous samples and should thus be used with caution for anomaly detection [20,25]. SPADE [9] constructs a memory bank of normal embeddings and uses the average distance to k nearest neighbors as an anomaly score for a given sample. PaDiM [11] creates a probabilistic representation of the features by modeling feature distribution for each pixel by a multivariate Gaussian. PatchCore [28] and ProtoAD [17] employ a memory bank as in SPADE, however, both methods compress the memory bank to reduce its size using a greedy coreset selection and a nonparametric hierarchical clustering, respectively. AnomalyDINO [10] stores patch features from DINOv2 [26] using only a few normal images to construct a memory bank.

The comparison with large memory banks usually leads to accurate anomaly detections, however at the cost of substantially long runtimes and high memory utilization. Having small yet expressive memory banks is crucial for anomaly detection on a large scale as it would be required for medical imaging.

In this work, we aim to effectively compress the memory bank of AnomalyDINO. We propose to model the patch features using a Dirichlet process mixture model (DPMM) [1,13] with Gaussian component distributions, which is a Gaussian mixture model without a fixed number of components that instead determines the necessary number of components based on the data, allowing for a more flexible, data-driven approach. We fit the DPMM using a batched expectation maximization algorithm [19]. We use the component means as prototypes for the normal distribution and define the anomaly score as the similarity to these prototypes. For an overview of our approach refer to Fig. 1.

Our framework achieves competitive performance on the BMAD benchmark [3] while being highly efficient to compute thanks to the small number of prototypes in use to model the normative distribution. Further, our analyses indicate that features from DINOv2 trained on natural images are useful for anomaly detection in medical imaging, especially after normalization.

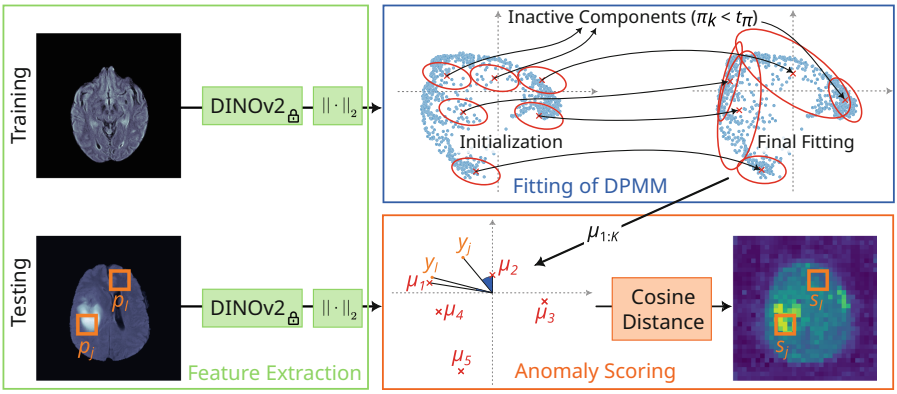


Fig. 1. In the training phase, we fit a Dirichlet process mixture model (DPMM) with Gaussian components (indicated in red) to the DINOv2 patch embeddings of the normal samples. After fitting the DPMM to the normal embeddings, we use the cosine distance between each patch feature and the closest component mean to obtain a patch-level anomaly score, which we interpolate for a pixel-wise anomaly map. (Color figure online)

2 Methods

2.1 Fitting the Dirichlet Process Mixture Model (DPMM)

Following [19], we fit the DPMM with Gaussian components using a batched expectation maximization (EM) approach. For this, we consider a DPMM truncated to K components. For sufficiently large K, the truncated DPMM is more flexible than a standard Gaussian mixture model as components can vanish in the DPMM if they are not required. We define the probability density of a sample y under the DPMM with parameters Φ as

$$p\left(y|\Phi\right) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y|\theta_k), \tag{1}$$

where $\theta_k = (\mu_k, \Sigma_k)$ and the component weights $\{\pi_k\}_{k=1}^K$ are defined as

$$\pi_k = \begin{cases} v_1, & \text{if } k = 1 \\ v_k \prod_{j=1}^{k-1} (1 - v_j), & \text{otherwise,} \end{cases} \tag{2}$$

with $v_k \sim \text{Beta}(1, \alpha)$ for $k \in \{1, \ldots, K-1\}$, $v_K = 1$, and α being the concentration parameter of the Dirichlet process. Given a batch of embeddings $y_{1:N}$, we want to estimate the parameters $\Phi = (v_{1:K-1}, \theta_{1:K})$ and α by maximizing

$$Q_\Phi(\Phi, \Phi^{(t-1)}) = \mathbb{E}_{x_{1:N}|y_{1:N}, \theta_{1:K}^{(t-1)}, \alpha^{(t-1)}} \left[\log p(x_{1:N}, y_{1:N}, v_{1:K}|\theta_{1:K}, \alpha)\right] \tag{3}$$

and

$$Q_\alpha(\alpha, \alpha^{(t-1)}) = \mathbb{E}_{x_{1:N}|y_{1:N},\theta_{1:K}^{(t-1)},\alpha^{(t-1)}} \left[\log \int p(x_{1:N}, y_{1:N}, v_{1:K}|\theta_{1:K}, \alpha)\, \mathrm{d}v_{1:K} \right] \tag{4}$$

with respect to Φ and α, respectively, where $x_n \in \{1, \ldots, K\}$ is the random variable for the component assignment of the embedding y_n.

For Gaussian components, we can obtain the update formulas

$$\theta_k^{(t)} = \left(\mu_k^{(t)}, \Sigma_k^{(t)}\right) = \left(\frac{\bar{m}_k^{(t)}}{\sum_{j=1}^K \bar{p}_j^{(t)}}, \quad \frac{\bar{c}_k^{(t)}}{\sum_{j=1}^K \bar{p}_j^{(t)}} - \mu_k^{(t)} \mu_k^{(t)^\top} \right) \tag{5}$$

$$v_k^{(t)} = \frac{\bar{p}_k^{(t)}}{\bar{p}_k^{(t)} + \alpha^{(t-1)} - 1 + \sum_{j=k+1}^K \bar{p}_j^{(t)}}, \tag{6}$$

where $\bar{p}_k$, $\bar{m}_k$, and $\bar{c}_k$ are moving averages of the sufficient statistics defined by

$$\bar{p}_k^{(t)} = (1 - \gamma_t)\bar{p}_k^{(t-1)} + \gamma_t\, \mathbb{E}_y \left[\Pr(x = k|y, \Phi^{(t-1)}) \right] \tag{7}$$

$$\bar{m}_k^{(t)} = (1 - \gamma_t)\bar{m}_k^{(t-1)} + \gamma_t\, \mathbb{E}_y \left[y \Pr(x = k|y, \Phi^{(t-1)}) \right] \tag{8}$$

$$\bar{c}_k^{(t)} = (1 - \gamma_t)\bar{c}_k^{(t-1)} + \gamma_t\, \mathbb{E}_y \left[yy^\top \Pr(x = k|y, \Phi^{(t-1)}) \right]. \tag{9}$$

Finally, the update formula for the concentration parameter α is

$$\alpha^{(t)} = \frac{N - 1}{\sum_{k=1}^{N-1} \Psi\left(\alpha^{(t-1)} + 1 + \bar{C}_k^{(t)} + \bar{C}_{>k}^{(t)}\right) - \Psi\left(\alpha^{(t-1)} + \bar{C}_{>k}^{(t)}\right)}, \tag{10}$$

with $\bar{C}_k^{(t)} = B \cdot \bar{p}_k^{(t)}$, $\bar{C}_{>k}^{(t)} = B \sum_{j=k+1}^K \bar{p}_j^{(t)}$, and B being the batch size. Further details on the derivation of the formulas above can be found in [19].

2.2 Anomaly Scoring

Given the DPMM fitted to the normal distribution of patch embeddings, we can evaluate the alignment of a test sample with the modeled distribution using several metrics. Specifically, we consider the likelihood of the most probable component as well as the cosine similarity and Euclidean distance to the nearest component center. As shown in Fig. 2, certain components tend to model specific anatomical structures. Among the three metrics, cosine similarity provides the most consistent assignment of anatomical structures across different samples, particularly for anomalous samples. Using the full likelihood of the model in Eq. (1) is not an option, as in our experiments the full likelihood is usually dominated by the likelihood of only one component. Based on this observation, we 1) choose to normalize the embeddings y for training and testing and 2) use cosine similarity as our anomaly score function.

During fitting the DPMM, some components can vanish as their weight π_k shrinks towards zero. To suppress all components with vanishing weights, we

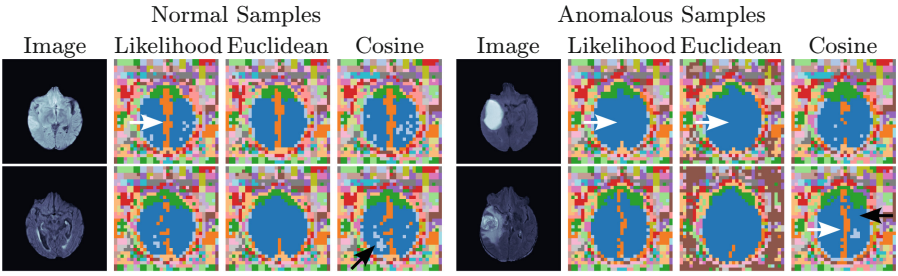


Fig. 2. For two normal and two anomalous images, we visualized the index of the closest component for each patch in terms of component likelihood ($\text{argmax}_k \mathcal{N}(y_n|\mu_k, \Sigma_k)$), Euclidean distance ($\text{argmin}_k ||y_n - \mu_k||_2$), and cosine similarity ($\text{argmax}_k y_n^\top \mu_k / (||y_n||_2 \cdot ||\mu_k||_2)$). To highlight the consistency over multiple samples, we selected samples showing similar anatomical structures for this visualization, with and without lesions. Different brain structures of the normal images are seemingly represented by different components: tissues at the centerline by the orange component (white arrow), ventricles by the bright blue component (black arrow). For the anomalous images, we can observe a similar assignment when using cosine similarity. However, the component likelihood and the Euclidean distance fail to assign these structures to the correct component in several anomalous cases. (Color figure online)

apply a threshold t_π to the component weights first, and the anomaly score function becomes

$$s(y_n) = \max_{k:\pi_k > t_\pi} \frac{y_n^\top \mu_k}{||y_n||_2 ||\mu_k||_2}. \tag{11}$$

Since DINOv2 only creates embeddings of image patches, this yields a coarse, patch-wise anomaly score map. To obtain pixel-wise anomaly scores, we interpolate the patch-wise anomaly scores to full resolution.

To create an anomaly segmentation mask, we threshold the pixel-level anomaly score map. To stay in line with the unsupervised setting, we select the threshold such that a predefined false positive rate on the healthy validation set is achieved as in [8].

3 Experiments and Results

3.1 Experimental Setup

Model Architecture. As backbone, we use the small architecture of DINOv2 with the official pretrained weights[1], leading to a 384-dimensional embedding space. We truncate the DPMM to 500 components and use diagonal covariance matrices for the Gaussian components. We set the discount factor in the moving average γ_t to 0.2 to accumulate the statistics over many batches and the component weight threshold t_π to $1e{-}6$. We train the DPMM with a batch size of 12 images, which are resized to a resolution of 448×448, resulting in batches

[1] https://github.com/facebookresearch/dinov2 under the Apache 2.0 License.

of 12288 embedding vectors. We trained the DPMM for 40 epochs and selected the model achieving the highest log-likelihood on the normal validation samples.

Datasets. We conducted our experiments using the anomaly segmentation datasets from the BMAD benchmark [3], which include BraTS2021 [2] (brain MRI), BTCV [21] and LiTs [4] (liver CT), and RESC [16] (retinal OCT). BraTS2021 comprises 8179 normal and 3119 anomalous images, the liver CT datasets have 2468 normal and 733 anomalous images, and RESC contains 5408 normal and 809 anomalous images.

Baselines. We compare our work with multiple unsupervised anomaly detectors, namely DRAEM [33], UTRAD [5], RD4AD [12], STFPM [30], PaDiM [11], PatchCore [28], CFA [22], CFLOW [14] and AnomalyDINO [10]. For PatchCore, we applied the proposed reduction to 10% of the memory bank as well as a much stronger reduction to 1024 and to 150 prototypes, which we refer to as few-shot setting. For AnomalyDINO, we tested the full-shot and one-shot setting, with the one-shot setting corresponding to 1024 prototypes.

Evaluation Metrics. We compare our method with state-of-the-art anomaly detectors using the area under the receiver operator characteristic curve (AUROC) and the area under the precision recall curve (AUPR) at pixel-level. We also report the per-sample runtime and maximum memory utilization for the anomaly detection stage, measured with a batch size of 1 and one separate thread for data loading using an NVIDIA A6000 GPU with 48 GB of VRAM. Further, we compare the anomaly segmentation performance with AnomalyDINO using the Dice score at 1%, 5%, and 10% false positive rate on the normal validation images.

To test the statistical significance of our results over PatchCore with 150 prototypes, we run a paired permutation test on the image-wise AUROC and AUPR scores with 10000 permutations for the BraTS2021 and the RESC datasets. Using image-wise scores is necessary as the anomaly score maps for PatchCore and our method are constructed using Euclidean and cosine distance to the nearest prototypes, respectively, and thus cannot be compared directly.

3.2 Results

We provide a quantitative comparison of our method with state-of-the-art approaches in Table 1. Over all datasets, only AnomalyDINO in the full-shot setting manages to outperform our approach consistently. However, AnomalyDINO comes with enormous memory requirements and with long runtimes. For BraTS and BTCV+LiTs, the AUROC for our approach is only 2.13% and 1.83% lower than for the best-performing method, respectively. With the exception of the one-shot AnomalyDINO on the RESC dataset, our method runs more than 1.8 times faster than all better-performing methods. This trade-off is also

Table 1. AUROC and AUPR scores for pixel-level anomaly detection as well as per-sample runtime (without data loading or metric computation) and memory utilization of the anomaly detection on the test set. The methods with the suffix "R50" use WideResNet50 [32] as the backbone instead of ResNet18 [15]. PatchCore-R50 in the standard setting is not available for the BraTS2021 dataset as the coreset subsampling could not fit into the GPU memory. We averaged AUROC and AUPR scores for AnomalyDINO in the one-shot setting over 25 runs.

	BraTS2021				BTCV+LiTs				RESC			
Method	AUROC [%]	AUPR [%]	Runtime [ms]	Memory [GB]	AUROC [%]	AUPR [%]	Runtime [ms]	Memory [GB]	AUROC [%]	AUPR [%]	Runtime [ms]	Memory [GB]
DRAEM	92.70	38.66	149	0.5	93.82	16.01	125	0.5	86.78	31.86	130	0.5
UTRAD	92.81	24.63	43	0.9	84.28	1.25	45	0.9	55.52	6.00	42	0.9
RD4AD	97.81	52.11	139	0.4	94.35	3.48	118	0.4	96.92	67.56	117	0.4
STFPM	96.64	45.45	122	0.0	94.70	3.20	106	0.0	93.85	45.07	108	0.0
STFPM-R50	97.32	52.87	133	0.2	96.14	5.10	118	0.2	94.78	52.66	111	0.2
PaDiM	85.77	8.21	145	0.2	89.77	1.66	127	0.2	90.00	31.07	124	0.2
PaDiM-R50	83.36	6.89	158	3.7	90.48	1.79	131	3.7	92.12	35.31	129	3.7
PatchCore	98.34	67.37	202	3.4	96.67	5.66	132	0.7	97.08	70.47	150	2.0
PatchCore-R50	N/A				96.63	5.99	160	1.9	97.55	72.62	218	5.0
CFA	97.27	54.15	134	0.3	97.29	7.65	125	0.3	91.65	26.52	116	0.3
CFLOW	95.18	32.63	255	1.0	97.13	11.63	223	0.9	94.79	53.21	219	0.9
AnomalyDINO	97.71	58.69	422	33.1	97.24	17.91	107	6.9	94.04	50.45	984	38.1
PatchCore-R50[a]	95.38	55.07	149	0.2	45.70	0.32	130	0.1	93.28	61.88	132	0.1
PatchCore-R50[b]	95.02	42.68	158	0.2	16.31	0.23	139	0.1	78.70	25.98	136	0.1
▲	96.57	44.74			98.17	14.30			92.24	42.63		
AnomalyDINO[a*]	94.61	31.21	24	0.2	94.60	5.98	26	0.2	89.52	36.98	49	0.4
▼	84.97	8.72			89.45	2.93			85.10	26.51		
Ours	96.21[†]	43.43[†]	39	1.6	95.46	9.02	35	1.5	90.20 [†]	41.66[†]	58	2.0

[a] few-shot with 1024 prototypes
[b] few-shot with 150 prototypes
[*] max (▲), average, and min (▼) value over all seeds provided
[†] statistically significant improvement over PatchCore-R50[†] (permutation test with $p < 1\%$)

visualized in Fig. 3. Our method is one of the furthest towards low runtimes and high AUROC and AUPR scores, indicating a better trade-off than state-of-the-art methods.

Our approach ends up with roughly 120 to 150 prototypes, while the remaining components vanish. Comparing with similarly sized coresets for PatchCore

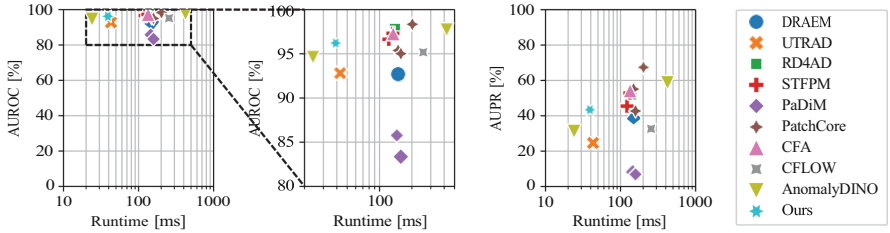


Fig. 3. Visualization of the trade-off between runtime and anomaly segmentation performance on the BraTS2021 dataset. Methods with high anomaly segmentation performance and low runtime are at the top left in each plot. (Color figure online)

Table 2. Dice scores for anomaly segmentation for three different thresholds selected to achieve a fixed false positive rate (FPR) on the normal validation set. Results for the one-shot setting are averaged over 25 samples.

	BraTS2021			BTCV+LiTs			RESC		
FPR	10%	5%	1%	10%	5%	1%	10%	5%	1%
AnomalyDINO (full-shot)	29.26	38.16	54.24	7.19	10.48	18.69	39.28	45.82	52.30
AnomalyDINO (one-shot)	31.89	37.45	32.98	6.17	8.20	10.63	36.48	40.34	33.88
Ours	31.64	41.45	45.87	6.19	8.52	14.76	38.00	43.74	43.49

or the one-shot setting for AnomalyDINO, our approach outperforms both in almost all metrics, while still having a substantially lower runtime compared to PatchCore. The implementation of AnomalyDINO is more optimized for efficiency than ours, for instance, they use Faiss [18] for the nearest neighbor search and thus reach faster runtimes. Using a similar optimization, we expect our method to reach even lower runtimes.

We report Dice scores for our approach and both settings of AnomalyDINO in Table 2. Not surprisingly, AnomalyDINO in the full-shot setting yields the highest scores for most of the thresholds. However, our approach outperforms the one-shot AnomalyDINO in all but one case, usually by quite some margin. Despite using only a few prototypes, our method demonstrates its effectiveness in reducing the gap to methods with many prototypes.

3.3 Ablation Studies

We analyze the choice of cosine distance over Euclidean distance and the log-likelihood for our anomaly metric as well as the choice of normalizing the DINOv2 features. The resulting AUROC and AUPR scores are in Table 3. Generally, the results using normalized features are higher than for unnormalized features. The detection with cosine distance leads to the highest scores for most of the experiments, further supporting the findings of our preliminary analysis in Sect. 2.2.

Table 3. Ablation results for the anomaly score function and for the normalization of the DINOv2 features.

Anomaly Score Function	Feature Normalization	BraTS2021		BTCV+LiTs		RESC	
		AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Likelihood	False	92.85	23.56	94.62	5.90	85.92	29.04
Euclidean	False	93.52	29.04	93.99	7.54	86.26	35.47
Cosine	False	95.06	38.41	95.30	7.96	89.78	42.52
Likelihood	True	95.49	33.44	95.66	9.92	89.27	33.80
Euclidean	True	96.17	42.68	95.44	9.22	90.15	41.00
Cosine	True	96.21	43.43	95.46	9.02	90.20	41.66

4 Conclusion

In this work, we proposed an unsupervised anomaly detection framework that leverages the expressive power of DINOv2 embeddings and models their distribution using a Dirichlet Process Mixture model (DPMM), which automatically adjusts the number of components to the data at hand. Experimental results on the BMAD benchmark [3] demonstrate that our method achieves competitive performance while significantly reducing the computational cost. Currently, the feature extraction with DINOv2 is the main bottleneck, so future research could explore this setup with distilled models to further improve efficiency.

Acknowledgments. This project was funded by the Swiss National Science Foundation (SNF) (205320_200877).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Stat. **2**(6), 1152–1174 (1974)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv: 2107.02314 (2021)
3. Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., Li, X.: BMAD: benchmarks for medical anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 4042–4053 (2024)
4. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., et al.: The liver tumor segmentation benchmark (LiTS). Med. Image Anal. **84**, 102680 (2023)
5. Chen, L., You, Z., Zhang, N., Xi, J., Le, X.: UTRAD: anomaly detection and localization with u-transformer. Neural Netw. **147**, 53–62 (2022)
6. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)

7. Chen, X., Pawlowski, N., Glocker, B., Konukoglu, E.: Normative ascent with local gaussians for unsupervised lesion detection. Med. Image Anal. **74**, 102208 (2021)
8. Chen, X., You, S., Tezcan, K.C., Konukoglu, E.: Unsupervised lesion detection via image restoration with a normative prior. Med. Image Anal. **64**, 101713 (2020)
9. Cohen, N., Hoshen, Y.: Sub-image anomaly detection with deep pyramid correspondences. arXiv preprint arXiv: 2005.02357 (2021)
10. Damm, S., Laszkiewicz, M., Lederer, J., Fischer, A.: AnomalyDINO: boosting patch-based Few-shot Anomaly Detection with DINOv2. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025)
11. Defard, T., Setkov, A., Loesch, A., Audigier, R.: PaDiM: a patch distribution modeling framework for anomaly detection and localization. In: Pattern Recognition. ICPR International Workshops and Challenges, pp. 475–489 (2021)
12. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9737–9746 (2022)
13. Ferguson, T.S.: Bayesian density estimation by mixtures of normal distributions. In: Recent Advances in Statistics, pp. 287–302 (1983)
14. Gudovskiy, D., Ishizaka, S., Kozuka, K.: CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 98–107 (2022)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
16. Hu, J., Chen, Y., Yi, Z.: Automated segmentation of macular edema in OCT using deep neural networks. Med. Image Anal. **55**, 216–227 (2019)
17. Huang, C., Kang, Z., Wu, H.: A prototype-based neural network for image anomaly detection and localization. Neural Process. Lett. **56**(3), 169 (2024)
18. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2019)
19. Kimura, T., Tokuda, T., Nakada, Y., Nokajima, T., Matsumoto, T., Doucet, A.: Expectation-maximization algorithms for inference in Dirichlet processes mixture. Pattern Anal. Appl. **16**(1), 55–67 (2013)
20. Kirichenko, P., Izmailov, P., Wilson, A.G.: Why normalizing flows fail to detect out-of-distribution data. In: Advances in Neural Information Processing Systems, vol. 33, pp. 20578–20589 (2020)
21. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proceedings of the MICCAI Multi-Atlas Labeling beyond Cranial Vault—Workshop Challenge, vol. 5, p. 12 (2015)
22. Lee, S., Lee, S., Song, B.C.: CFA: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. IEEE Access **10**, 78446–78454 (2022)
23. Liu, M., Jiao, Y., Lu, J., Chen, H.: Anomaly detection for medical images using teacher-student model with skip connections and multiscale anomaly consistency. IEEE Trans. Instrum. Meas. **73**, 1–15 (2024)
24. Lu, S., Zhang, W., Zhao, H., Liu, H., Wang, N., Li, H.: Anomaly detection for medical images using heterogeneous auto-encoder. IEEE Trans. Image Process. **33**, 2770–2782 (2024)

25. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? In: International Conference on Learning Representations (2019)
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., et al.: DINOv2: learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024)
27. Rashmi, K., Das, A., Matcha, N., Ram, K., Sivaprakasam, M.: Ano-swinMAE: unsupervised anomaly detection in brain MRI using swin transformer based masked auto encoder. In: Medical Imaging with Deep Learning (2024)
28. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14318–14328 (2022)
29. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: F-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. **54**, 30–44 (2019)
30. Wang, G., Han, S., Ding, E., Huang, D.: Student-teacher feature pyramid matching for anomaly detection. In: The British Machine Vision Conference (BMVC) (2021)
31. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: AnoDDPM: anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 650–656 (2022)
32. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 87.1–87.12 (2016)
33. Zavrtanik, V., Kristan, M., Skočaj, D.: DRAEM - a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8330–8339 (2021)
34. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised anomaly localization using variational auto-encoders. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 289–297. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_32

Aorta Multi-class Segmentation via Anatomically Constrained Plane Detection

Jonghoon An, Dong Hyun Lee, So Hyun Kim, Taejin Moon, and Minyoung Chung(✉)

School of Software, Soongsil University, Seoul, Republic of Korea
chungmy@ssu.ac.kr

Abstract. Accurate multi-class segmentation of the aorta in medical CT images is essential for the effective diagnosis and treatment of blood flow abnormalities. However, achieving precise segmentation in multi-zone remains challenging due to the lack of visible boundaries and the similarity in intensity between zones. Although existing methods incorporate anatomical features such as global geometric constraints and landmark-based alignment, they often struggle when these features are difficult to extract, such as in regions with asymmetric deformation or extreme curvature due to dissection. This limitation of relying solely on simple anatomical cues underscores the need to learn and model complex anatomical interrelationships for robust segmentation. To overcome these challenges, we propose a plane detection-based segmentation framework that is constrained by anatomical features and their relationships to accurately detect planes between zones. Specifically, our method detects planes by localizing centerpoints and regressing the corresponding normal vectors, while anatomical landmarks further refine the position and orientation of these planes. Additionally, anatomical regularization losses enforce geometric consistency among these components, thereby enhancing both accuracy and stability of the detected planes. The entire framework is implemented as an end-to-end architecture, enabling efficient learning. The experimental results on the AortaSeg24 dataset demonstrate that our approach achieves state-of-the-art performance. Our code is publicly available at https://github.com/jjong0225/ACP.

Keywords: Anatomical constraints · Anatomical regularization · Aorta multi-class segmentation · Plane detection

1 Introduction

Aortic segmentation in medical CT images is a critical step in various clinical procedures, particularly for addressing blood flow abnormalities [1–3]. For successful interventions, detailed structural and functional analysis is essential, which is achieved by dividing the aorta into multiple zones based on clinically relevant anatomical definitions, such as the Society for Vascular Surgery/Society

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 57–66, 2026.
https://doi.org/10.1007/978-3-032-04947-6_6

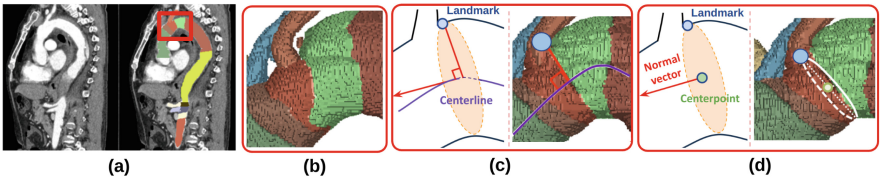


Fig. 1. (a) Multi-zone image and label. (b) Previous works (SwinUNETR [6]). (c) Anatomical definition of multi-zone class boundaries with ground-truth. (d) Proposed plane detection-based segmentation (Ours).

of Thoracic Surgeons (SVS/STS) zones. [4]. Multi-zone segmentation enables the automated measurements of the diameters and volumes, which are essential for accurate device selection, and precise stent placement [5]. Given that these measurements directly impact treatment efficacy [7], achieving high segmentation accuracy is paramount [5,8]. While considerable progress has been made in aortic segmentation, multi-zone segmentation remains challenging due to the inherent similarity of class intensities and the lack of clear boundaries between zones (Fig. 1a, b).

To address these challenges, significant research has leveraged anatomical knowledge to constrain segmentation outputs in line with expected anatomical properties [9–15]. Since inter-zone planes are defined by the aortic structure and its landmarks (Fig. 1c), two primary strategies have emerged: imposing geometric constraints across zones [5,9–11] and aligning planes using landmarks [12–14]. Geometric constraints approaches focus on constraining structural consistency across zones. In the AortaSeg24 challenge [5], A5 employed Skeleton Recall Loss [16] to ensure connectivity, while A2 applied centerline boundary Dice [17] to enhance centerline integrity. Conversely, landmark-driven plane approaches directly define planes by leveraging landmarks and centerlines. The planes are defined to pass through the detected landmarks while being orthogonal to the centerline (Fig. 1c). For example, [12] derived landmarks through curvature analysis, whereas [13] incorporated an auxiliary network for landmark prediction. Although these studies resolve visually ambiguous multi-zone boundaries, they often struggle to produce accurate segmentation when these features are poorly extracted such as in regions with asymmetric deformation or extreme curvature due to dissection [15,18]. This limitation of relying solely on simple anatomical cues underscores the need to learn and model complex anatomical interrelationships for robust segmentation.

In this paper, we propose an anatomically constrained plane detectionbased segmentation framework that ensures accurate, anatomically aligned planes. Specifically, the framework decomposes the plane into three components: (1) the centerpoint, which defines its approximate location along the centerline, (2) landmarks, which serve as anatomical reference points to guide both its location and alignment, and (3) the normal vector, which determines its precise

alignment by accounting for the aortic structure. Furthermore, our anatomical regularization enforces the geometric relationships among these components. This guarantees that the predicted planes adhere to the anatomical priors, thus improving segmentation accuracy. We implemented the entire framework including segmentation and plane detection into an end-to-end framework, enabling efficient learning.

To summarize, the contributions of this study are as follows:

- We propose a novel plane detectionbased segmentation framework.
- We introduce an anatomy-constrained plane detection that learns the anatomical components of planes and their geometric relationships.
- We implemented an efficient end-to-end architecture that jointly learns plane detection and aortic segmentation.
- Our framework outperforms state-of-the-art methods in segmentation accuracy on the AortaSeg24 dataset.

2 Anatomically Constrained Plane Detection

We propose an anatomically constrained plane detection-based segmentation framework. An overview of our framework is illustrated in Fig. 2a. Our framework employs plane detection to address visually ambiguous boundaries by integrating anatomical constraints. Specifically, we model each plane with three key components (centerpoints, landmarks, and normal vectors) and learn their anatomical relationships. The plane position is first estimated using centerpoints and landmarks extracted from a keypoint detection head (Fig. 2b, Sect. 2.1). Then, the normal vector for each plane class, conditioned on the centerpoints and landmarks of the corresponding class, is estimated via a normal vector regression head (Fig. 2b, Sect. 2.2). Finally, an anatomical regularization term is applied to capture the geometric relationship among these components (Fig. 2c, Sect. 2.3).

2.1 Centerpoint and Landmark Detection

Our framework detects planes using seven arterial branch landmarks and nine centerpoints derived from their centerline (Fig. 2a). Specifically, landmarks are identified as the arterial branches closest to the boundary points, while the centerpoints represent the midpoints of the boundaries between adjacent zones.

To obtain keypoints $\mathbf{K}$, composed of landmarks $\mathbf{L}$ and centerpoints $\mathbf{C}$, we adopt a heatmap-based keypoint detection framework following [13]. This approach estimates keypoints by learning their spatial probability distribution. Specifically, shared features $\mathbf{f}_{share}$ are passed through a keypoint detection head (KD Head), composed of convolutional layers, to generate heatmaps for landmarks $\mathbf{h_L}$ and centerpoints $\mathbf{h_C}$ (Fig. 2b). These heatmaps are converted into trainable keypoint coordinates using keypoint localization, which computes the

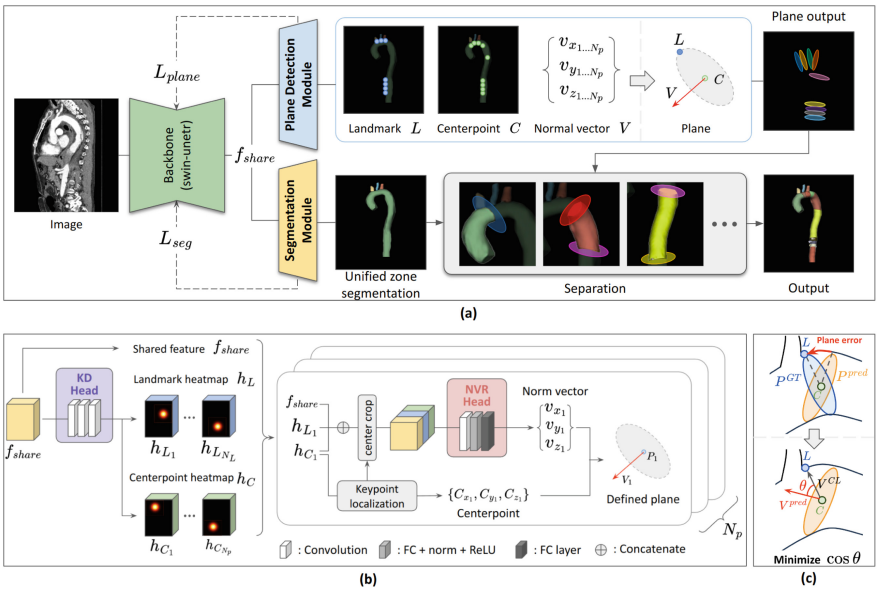


Fig. 2. (a) Overview of our framework for plane-based multi-class segmentation, where multi-zone segmentation is defined by partitioning the predicted unified zone segmentation with detected planes. (b) Plane detection module estimates class-wise planes using centerpoint and normal vector. The keypoint detection head (KD Head) generates heatmaps for landmarks and centerpoints, which are then fused with features and passed to the normal vector regression head (NVR Head) to predict class-specific normal vectors. (c) Anatomical regularization refines planes by enforcing geometric constraints among anatomical components. Plane error, visualized as the blue-orange discrepancy, is quantified by the angle θ between V^{pred} and V^{CL}. Enforcing $\theta = 90°$ minimizes the error and ensures anatomical consistency.

weighted mean of the heatmap and the coordinate map $\mathbf{r}$ through element-wise multiplication, as follows:

$$\mathbf{K} = \frac{\sum_{x,y,z} h(x,y,z) \cdot \mathbf{r}(x,y,z)}{\sum_{x,y,z} h(x,y,z)}, \tag{1}$$

where $\mathbf{r}(x,y,z)$ is the spatial coordinates and $h(x,y,z)$ is the heatmap value. To train the model with heatmaps, we use Gaussian heatmaps centered on keypoints as ground truth (GT) and optimize the model using focal loss $\mathcal{L}_{\text{key}}$ [14], following [13].

2.2 Normal Vector Regression

We directly predict the parametric representation of a plane by regressing its normal vector $\mathbf{V} = (v_x, v_y, v_z)$. To ensure that the predicted normal vector aligns with the centerpoints and anatomical landmarks (Fig. 2b), class-specific normal

vectors are regressed while being conditioned on the corresponding keypoint heatmap predictions. Specifically, the regression pipeline begins with the shared features $\mathbf{f}_{share}$ and keypoint heatmaps ($\mathbf{h}_L$ and $\mathbf{h}_C$) as inputs. These inputs are cropped around the detected centerpoints, focusing on the regions of interest. The cropped features are then fed into the normal vector regression head (NVR Head), which predicts the normal vector $\mathbf{V}^{pred}$ through a series of convolution layers and fully connected layers (Fig. 2b). For training, we minimize the vector regression loss $\mathbf{L}_{vec}$, which penalizes the L1 difference between predicted normal vectors $\mathbf{V}^{pred}$ and GT normal vectors $\mathbf{V}^{GT}$, as follows:

$$\mathcal{L}_{vec} = \frac{1}{N_p} \sum_{i=1}^{N_p} \left\| \mathbf{V}_i^{pred} - \mathbf{V}_i^{gt} \right\|_1 . \tag{2}$$

2.3 Anatomical Regularization

To ensure geometric consistency across predicted plane components, we introduce an anatomical regularization. This regularization leverages the anatomical definition of multi-zone regions, where the centerpoint and landmark vector of each plane are constrained to lie on that plane defined by the normal vector (Fig. 2c).

Specifically, we first convert the heatmap into landmark and centerpoint coordinates using (1). We then compute the centerpointlandmark vector $\mathbf{V}^{CL}$ by subtracting the centerpoint from the landmark and normalizing it. Finally, we measure the orthogonality between the $\mathbf{V}^{CL}$ and the $\mathbf{V}^{pred}$ using cosine similarity, which is incorporated into the anatomical regularization loss $\mathcal{L}_{reg}$, as follows:

$$\mathcal{L}_{reg} = \frac{1}{N_p} \sum_{i=1}^{N_p} \left(\mathbf{V}_i^{CL} \cdot \mathbf{V}_i^{pred} \right)^2 . \tag{3}$$

Minimizing $\mathcal{L}_{reg}$ enforces geometrical consistency among the plane's components, preserving anatomical structure and enabling accurate segmentation.

2.4 End-to-End Architecture

We implemented an end-to-end architecture that jointly trains plane detection and aortic segmentation. The process begins with a shared backbone that extracts shared features, $\mathbf{f}_{share}$, which are then fed to task-specific modules for plane detection and segmentation. By jointly learning from a shared representation, each module remains independent in its predictions while leveraging shared information, enhancing efficiency and consistency across tasks. We employ SwinUNETR [6] as the shared backbone and segmentation framework for unified zone class and 13 artery classes. Following its original implementation, the model optimizes a segmentation loss $\mathcal{L}_{seg}$ combining Dice loss [14] and Focal loss [19]. All losses are integrated into a joint loss formulation, as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{key}\mathcal{L}_{key} + \lambda_{vec}\mathcal{L}_{vec} + \lambda_{reg}\mathcal{L}_{reg}, \tag{4}$$

where λ_{key}, λ_{vec}, and λ_{reg} are weighting coefficients for $\mathcal{L}_{key}$, $\mathcal{L}_{vec}$, and $\mathcal{L}_{reg}$, respectively.

3 Experiments

Experiments were conducted on the AortaSeg24 Challenge dataset [5], the only publicly available benchmark for multi-class segmentation of aortic structures in CTA scans. The dataset is divided into 40 training and 10 test cases. All images were resampled to an isotropic resolution of 1 mm^3, intensity-normalized to [0,1], and cropped using labels. Additional labels for plane parameters are derived using the centroid and principal component analysis (PCA) of class boundary points. The derived ground truth (GT) plane parameters effectively characterize the segmentation structure, achieving a Dice Similarity Coefficient [20] of 0.9732 for the segmentation obtained through plane-based separation. During the test phase, this separation process involves refining the predicted segmentation by extracting the largest connected component, followed by partitioning into subregions based on the predicted plane parameters. Each subregion is then assigned a class label corresponding to its associated plane.

All models, including our proposed method and competing baselines, were trained from scratch without any pre-training, under identical conditions and rigorously tuned to ensure fair comparison.

3.1 Evaluation Metrics

We evaluate segmentation performance using the Dice Similarity Coefficient (DSC), the 95th percentile hausdorff distance (HD95), and the average symmetric surface distance (ASSD). DSC quantifies the overlap ratio between the predicted and GT segmentations, providing a measure of overall segmentation accuracy. In contrast, HD95 and ASSD assess surface-based errors, capturing the geometric discrepancy between predicted and GT planes.

Additionally, to assess clinical adaptability, we include diameter measurements, which are crucial for applications such as surgical planning, device selection, and minimally invasive interventions [21]. Specifically, we define the diameter error (DE) as the absolute difference in distance between the predicted and GT diameters, measured in millimeters (mm).

3.2 Experimental Results

For evaluation, we compare two widely used segmentation models (3D U-Net [22] and SwinUNETR [6]) along with three aortic segmentation models (CIS-UNet [23], A5 [5], an nnU-Net [24] variant model with geometric constraints, and M-SL [13], a landmark-driven plane approach). Table 1 presents the quantitative segmentation performance, reporting the mean values of DSC, HD95, ASSD, and DE. Our method outperforms the baseline SwinUNETR and multiple state-of-the-art (SOTA) approaches, demonstrating superior segmentation accuracy, surface delineation, and clinical applicability.

Table 1. Segmentation performance comparison (mean ± std). Higher is better for DSC (↑); lower is better for HD95, ASSD, and DE (↓).

Method	DSC ↑	HD95 ↓	ASSD ↓	DE ↓
3D U-Net [22]	0.729 ± 0.064	8.875 ± 2.98	2.444 ± 0.79	6.328 ± 1.83
SwinUNETR [6]	0.718 ± 0.057	9.044 ± 3.00	2.484 ± 0.76	5.471 ± 2.03
CIS-UNet [23]	0.748 ± 0.063	8.360 ± 2.77	2.249 ± 0.75	4.810 ± 1.86
A5 [5]	0.750 ± 0.045	7.587 ± 2.45	2.156 ± 0.57	3.992 ± 1.18
M-SL [13]	0.732 ± 0.060	8.483 ± 2.18	2.266 ± 0.64	3.453 ± 1.40
Ours	**0.793 ± 0.039**	**5.990 ± 2.06**	**1.745 ± 0.43**	**3.016 ± 1.11**

Table 2. Zone-wise DSC performance for each zone, denoted as Z_i, where i represents the zone index.

Method	Z_0	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9
3D U-Net [22]	0.919	0.619	0.712	0.706	0.715	0.858	0.668	0.599	0.635	0.855
SwinUNETR [6]	0.905	0.597	0.729	0.720	0.749	0.869	0.520	0.587	0.625	0.879
CIS-UNet [23]	**0.926**	0.656	0.719	0.667	0.748	0.882	0.648	0.661	0.689	0.884
A5 [5]	0.895	0.592	0.709	0.693	0.796	0.884	0.690	0.654	0.680	0.903
M-SL [13]	0.908	0.613	0.717	0.705	0.783	0.863	0.577	0.608	0.678	0.866
Ours	0.924	**0.666**	**0.752**	**0.726**	**0.827**	**0.900**	**0.759**	**0.731**	**0.740**	**0.908**

Our model improves DSC by 10.5% over the baseline and by 6% over CIS-UNet, the best-performing method for DSC, demonstrating consistently higher segmentation accuracy. Beyond DSC, our method reduces HD95 by 33% and ASSD by 31% compared to the baseline and by 28% and 22%, respectively, compared to A5, the best-performing method for surface accuracy. These results highlight our method's precise plane detection, reducing noise and improving segmentation accuracy. Furthermore, DE is reduced by 40% compare to the baseline and by 24% compared to M-SL, the best-performing method for DE. This improvement underscores our method's adaptability to clinical applications by precisely estimating diameters through accurate and stable plane detection. The consistently lower standard deviations across all metrics further confirm that our improvements are not driven by outlier handling, but reflect robust and consistent performance gains across cases, including those with asymmetric deformation or extreme curvature. Table 2 further validates these findings, demonstrating consistent performance gains across most anatomical zones, except for Zone 0. Figure 3 qualitatively illustrates that our method produces accurate plane, whereas SOTA methods exhibit increased noise and structural inconsistencies. These results reinforce the effectiveness of our approach in both segmentation accuracy and clinical reliability.

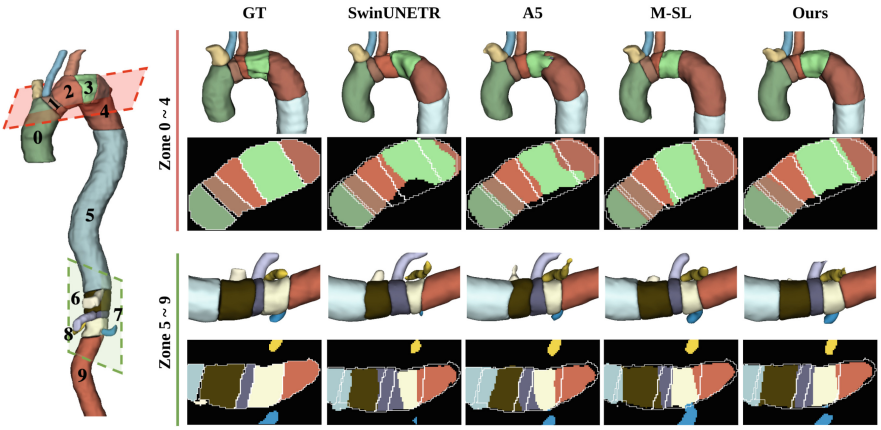


Fig. 3. Qualitative comparison of registration methods. The left panel visualizes overall zone locations and slice positions for comparison. The right panels present qualitative comparison, with the GT class boundary shown as a white contour. While other methods exhibit segmentation noise or inaccuracies in plane delineation, our approach achieves precise segmentation and plane estimation that closely aligns with the GT.

Table 3. Ablation study on our anatomically constrained plane detection framework.

Method	DSC ↑	HD95 ↓	ASSD ↓	DE ↓
Baseline (SwinUNETR [6])	0.7179	9.0437	2.4843	5.4712
+ landmark	0.7205	8.4696	2.3384	3.1487
+ plane detection	0.7682	7.2032	2.0194	3.1319
+ anatomical regularization	**0.7932**	**5.9906**	**1.7452**	**3.0159**

3.3 Ablation Study of Plane Detection

To assess the impact of individual components in our anatomically constrained plane detection framework, we performed an ablation study under consistent experimental settings. The study began with the baseline, followed by progressive additions of key components: (1) landmarks, (2) plane detection, and (3) anatomical regularization.

Table 3 presents the results, demonstrating that each component incrementally improves segmentation accuracy. The introduction of anatomical landmarks improves segmentation accuracy by defining a landmark-driven plane, reducing HD95 by 6.35%, ASSD by 5.87%, and DE by 42.45%. This suggest that leveraging anatomical priors effectively delineates ambiguous boundaries, mitigating noise and inaccuracies. Incorporating plane detection further enhances segmentation by providing spatial priors for both position and orientation, leading to a 6.62% improvement in DSC. This underscores the effectiveness of integrating multiple anatomical priors to refine segmentation accuracy. Finally, anatomical regularization enforces geometric consistency, further reducing HD95 by 16.83%,

ASSD by 13.58% and improving DSC by 3.25%. This suggest that beyond a simple combination of anatomical cues, modeling the interrelations among plane components enables more precise and robust segmentation. These results suggest that leveraging anatomical priors guides segmentation toward anatomically consistent and robust outcomes.

4 Conclusion

We presented an anatomically constrained plane detection-based segmentation framework for aortic multi-class segmentation. Our framework models the complex anatomical interrelationships of planes, ensuring not only high segmentation accuracy but also precise estimation of clinically relevant metrics. While our current work focuses on multi-zone segmentation of the aorta, future efforts will extend this methodology to other vascular structures, such as arteries, to further validate its generalizability and clinical utility.

Acknowledgments. This work was supported by the Material&parts technology development project (20024280, Development of a Focused Ultrasound Therapeutic Device for BBB Opening and Neuromodulation) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Mavridis, C., Economopoulos, T.L., Benetos, G., Matsopoulos, G.K.: Aorta segmentation in 3D CT images by combining image processing and machine learning techniques. Cardiovasc. Eng. Technol. **15**(3), 359–373 (2024)
2. Mayer, C., et al.: Aortic segmentations and their possible clinical benefits. In: MICCAI Challenge on Segmentation of the Aorta, pp. 135–140. Springer (2023)
3. Lareyre, F., et al.: Artificial intelligence in vascular surgical decision making. In: Seminars in Vascular Surgery, vol. 36, pp. 448–453. Elsevier (2023)
4. Lombardi, J.V., et al.: Society for vascular surgery (SVS) and society of thoracic surgeons (STS) reporting standards for type b aortic dissections. Ann. Thorac. Surg. **109**(3), 959–981 (2020)
5. Imran, M., et al.: Multi-class segmentation of aortic branches and zones in computed tomography angiography: the aortaseg24 challenge. arXiv preprint arXiv:2502.05330 (2025)
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop, pp. 272–284. Springer (2021)
7. Mendoza, D.D., et al.: Impact of image analysis methodology on diagnostic and surgical classification of patients with thoracic aortic aneurysms. Ann. Thorac. Surg. **92**(3), 904–912 (2011)
8. Krebs, J.R., et al.: Volumetric analysis of acute uncomplicated type b aortic dissection using an automated deep learning aortic zone segmentation model. J. Vasc. Surg. **80**(4), 1025–1034 (2024)

9. Xie, L., et al.: Semi-supervised region-connectivity-based cerebrovascular segmentation for time-of-flight magnetic resonance angiography image. Comput. Biol. Med. **149**, 105972 (2022)
10. Shit, S., et al.: clDice-a novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16560–16569 (2021)
11. Chen, D., et al.: Multi-stage learning for segmentation of aortic dissections using a prior aortic anatomy simplification. Med. Image Anal. **69**, 101931 (2021)
12. Pirentis, A., Kalogerakos, P.D., Mojibian, H., Elefteriades, J.A., Lazopoulos, G., Papaharilaou, Y.: Automated ascending aorta delineation from ECG-gated computed tomography images. Med. Biol. Eng. Comput. **60**(7), 2095–2108 (2022)
13. Yang, J., et al.: Segment aorta and localize landmarks simultaneously on noncontrast CT using a multitask learning framework for patients without severe vascular disease. Comput. Biol. Med. **160**, 107002 (2023)
14. Saitta, S., et al.: A deep learning-based and fully automated pipeline for thoracic aorta geometric analysis and planning for endovascular repair from computed tomography. J. Digit. Imaging **35**(2), 226–239 (2022)
15. Pepe, A., et al.: Automated cross-sectional view selection in CT angiography of aortic dissections with uncertainty awareness and retrospective clinical annotations. Comput. Biol. Med. **165**, 107365 (2023)
16. Kirchhoff, Y., et al.: Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. In: European Conference on Computer Vision, pp. 218–234. Springer (2024)
17. Shi, P., Hu, J., Yang, Y., Gao, Z., Liu, W., Ma, T.: Centerline boundary dice loss for vascular segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 46–56. Springer (2024)
18. Kaladji, A., Spear, R., Hertault, A., Sobocinski, J., Maurel, B., Haulon, S.: Centerline is not as accurate as outer curvature length to estimate thoracic endograft length. Eur. J. Vasc. Endovasc. Surg. **46**(1), 82–86 (2013)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
20. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
21. Artzner, C., et al.: Assessing the accuracy of an artificial intelligence-based segmentation algorithm for the thoracic aorta in computed tomography applications. Diagnostics **12**(8), 1790 (2022)
22. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
23. Imran, M., et al.: CIS-UNet: multi-class segmentation of the aorta in computed tomography angiography via context-aware shifted window self-attention. Comput. Med. Imaging Graph. **118**, 102470 (2024)
24. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)



Automated Characterization of Myocardial Scar Topological Patterns for Ventricular Tachycardia Screening

Xicheng Sheng[1], Yang Zhang[1], Lei Li[2], Bailiang Chen[3,4], Freddy Odille[3,4], and Xiahai Zhuang[1(✉)]

[1] School of Data Science, Fudan University, Shanghai, China
zxh@fudan.edu.cn
[2] Department of Biomedical Engineering, National University of Singapore, Singapore, Singapore
[3] CIC-IT 1433, INSERM, Université de Lorraine and CHRU Nancy, Nancy, France
[4] IADI U1254, INSERM and Université de Lorraine, Nancy, France
https://zmiclab.github.io/

Abstract. Ventricular tachycardia screening is crucial for early intervention and prevention of life-threatening cardiac events. Myocardial scar topology on late gadolinium enhancement (LGE) MRI offers detailed structural insights that may be closely associated with the mechanisms underlying ventricular tachycardia. However, accurate characterization presents challenges due to the substantial shape variability of myocardium, indistinct boundaries, small scar volumes, and potential issues with image quality. In this study, we present PolarNet, a novel framework for automatic scar segmentation and topological pattern characterization in polar coordinates. The framework incorporates a boundary-aware segmentation branch that explicitly models boundaries essential for scar characterization (endocardium, scar-start, scar-end, and epicardium), ensuring geometric consistency and anatomical coherence. Our method outperforms nnU-Net in both scar segmentation and topological pattern characterization. Code will be available at https://github.com/Sheng-xc/VTS_PolarNet.

Keywords: Ventricular Tachycardia · LGE MRI · Scar Topology · Polar Transformation

1 Introduction

Ventricular tachycardia (VT) is a potentially life-threatening arrhythmia characterized by rapid and irregular heartbeats originating in the ventricles [13], often associated with post-infarction myocardial scarring. Prophylactic implantable cardioverter-defibrillators (ICD) are recommended for post-infarct patients with low left ventricular (LV) ejection fractions [14]. However, ICD implantation offers

X. Sheng and Y. Zhang—These two authors contributed equally.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 67–76, 2026.
https://doi.org/10.1007/978-3-032-04947-6_7

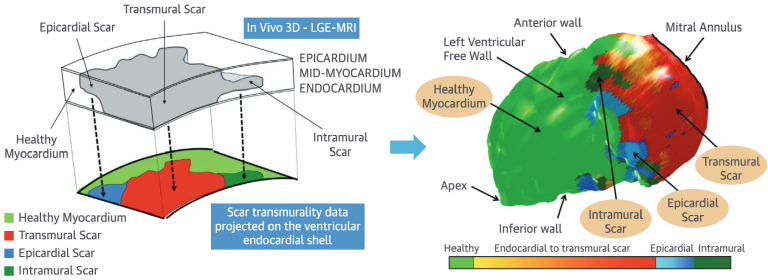


Fig. 1. Illustration of a myocardial scar with heterogeneous topological characteristics. Adapted from Fig. 1A and 1B of *JACC: Cardiovascular Imaging*, Vol. 14, de Chillou et al., "Magnetic Resonance Imaging Screening for Postinfarct Life-Threatening Ventricular Arrhythmia", pp. 2479–2481, Copyright (2021), with permission from Elsevier and the authors [1].

no benefit to patients without VT episodes while exposing them to potential complications, whereas LVEF-based criteria omit many at-risk individuals [17]. Therefore, more refined evaluations are needed to assess the risk of VT.

Myocardial scar tissue plays a critical role in the development of VT. Scars create heterogeneous substrates that can lead to re-entrant circuits, which are the primary mechanism of VT [3]. The topological transmurality of infarct scar, reflecting its relative position within the myocardial wall, may serve as an important marker for VT risk in postinfarct patients eligible for prophylactic ICD implantation [1]. As shown in Fig. 1, endocardial mapping of myocardial scar involves projecting scar tissue onto the endocardial shell. Each point on this projected scar surface is further classified into subtypes based on its topological relationship to the endocardial and epicardial boundaries [1].

To perform the projection, scar segmentation is essential. Numerous methodologies have been adapted for this task, ranging from traditional image processing techniques such as thresholding and region growing, to classical machine learning algorithms like SVMs and CRFs [9], and more recent deep learning architectures [10], including UNet-based models [15] and transformer-based networks [11]. Boundary quality has received limited attention, with some studies applying morphology-based [11] or signed-distance based [18] methods. Nonetheless, existing methods still struggle to accurately localize scars within the very thickness of myocardium on LGE MRI, largely due to the irregular morphology and blurred boundaries of pathological scars [12]. These inherent complexities often lead to fragmented segmentation outputs that lack anatomical consistency, making them unsuitable for applications in myocardial scar topology analysis.

Following scar segmentation, topology-based scar characterization emerged as a critical but underexplored issue. Current research in myocardial scar characterization for VT risk assessment predominantly relies on intensity-based measures of scar heterogeneity. For instance, methods based on maximum pixel signal have been adopted to identify core and border zone for arrhythmogenic sub-

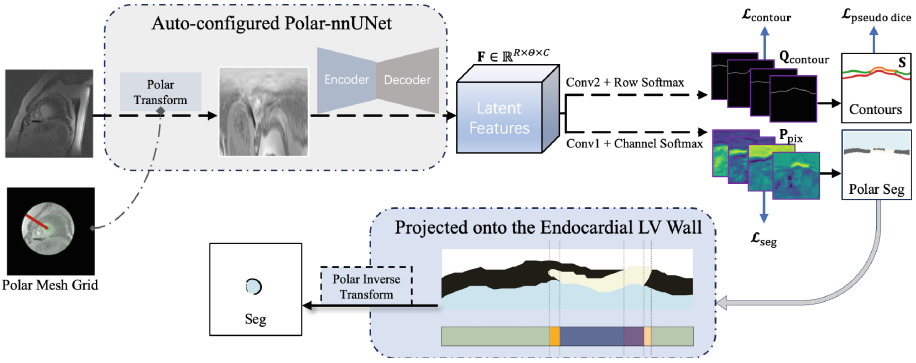


Fig. 2. The proposed framework for endocardial mapping of myocardial scar, consisting of 3 steps: 1) Anatomy-informed polar transformation, 2) Boundary-aware segmentation, 3) Projection-based subtype classification.

strates on semi-automatically [4] or automatically [16] delineated myocardium and scar. Similarly, standard deviation-based thresholding methods for dense and nondense scar quantification have been performed on VGG16-initialized UNet scar segmentations [5] for prediction of major adverse cardiac events. However, these approaches do not capture the topological heterogeneity of scars, which may offer a complementary perspective for VT risk stratification.

In this work, we propose PolarNet, a novel topology-aware framework designed for boundary-sensitive myocardial scar characterization. By leveraging polar transformation, the framework enables more intuitive endocardial mapping of myocardial scar. Its dual-branch architecture integrates pixel-wise scar segmentation with explicit boundary regression, thereby enhancing anatomical coherence and segmentation accuracy. To the best of our knowledge, this is the first deep learning-based framework designed to automate the characterization of scar topological patterns, offering clinically valuable insights for VT screening.

2 Methodology

Figure 2 presents the proposed PolarNet framework, which is designed to address the clinical demands of myocardial scar topological pattern characterization by unifying multi-class segmentation and boundary-aware regularization. The methodology is structured around three pillars: anatomy-informed polar transformation (Sect. 2.1), dual-branch segmentation and boundary regression (Sect. 2.2), and projection-based subtype classification (Sect. 2.3).

2.1 Anatomy-Informed Polar Transformation

The polar projection nature of endocardial mapping [2] motivates us to directly address the problem in the polar coordinate system for efficiency. To align with

the radial geometry of LV, we transform an input MRI slice $\mathbf{I} \in \mathbb{R}^{H \times W}$ into a polar representation $\mathbf{I}_{\text{polar}} \in \mathbb{R}^{R \times \Theta}$ centered at the LV centroid predicted by an anatomical segmentation model trained on public LGE MRI datasets. This transformation is implemented via grid sampling [8], using a normalized mesh grid $\mathbf{G} \in [-1, 1]^{R \times \Theta \times 2}$. Each target coordinate (r_i, θ_j) in the polar output $\mathbf{I}_{\text{polar}}$ corresponds to a sampling location in the original MRI slice $\mathbf{I}$, computed as

$$\mathbf{G}[i, j] = \left(\rho \cdot \frac{i}{R} \cos(\frac{2j\pi}{\Theta}), -\rho \cdot \frac{i}{R} \sin(\frac{2j\pi}{\Theta}) \right),$$

with $\rho \in [0, 1]$ controlling the field-of-view based on the anatomical segmentation.

2.2 Boundary-Aware Scar Segmentation

PolarNet employs a dual branch architecture inspired by structured surface segmentation for retina OCT [6]. It is built on a shared auto-configured residual encoder nnU-Net [7] backbone to extract hierarchical latent features $\mathbf{F} \in \mathbb{R}^{R \times \Theta \times C}$. These features are passed through task-specific convolutional layers to produce logits for segmentation and boundary prediction.

Segmentation Branch. The standard segmentation branch predicts pixel-wise probabilities $\mathbf{P}_{\text{pix}} \in \mathbb{R}^{R \times \Theta \times 4}$ for scar, healthy myocardium, left ventricle blood pool and background. A dice and cross-entropy loss is used for optimization:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Dice}}.$$

Boundary Branch. This branch predicts four boundaries critical for topological pattern characterization of scar: endocardium ($k = 1$), scar-start ($k = 2$), scar-end ($k = 3$), and epicardium ($k = 4$). While contour regression alone cannot segment scar due to its irregular shape, it guides attention to boundaries. This branch outputs boundary probabilities $\mathbf{Q}_{\text{contour}} \in \mathbb{R}^{R \times \Theta \times 4}$ via a row-wise softmax, where $\mathbf{Q}_k[r, \theta]$ represents the probability of boundary k at angle θ occurring at radial position r. Cross-entropy loss is adopted to penalize deviations from the ground-truth boundary positions:

$$\mathcal{L}_{\text{contour}} = -\frac{1}{4\Theta} \sum_{k=1}^{4} \sum_{\theta=0}^{\Theta-1} \log \mathbf{Q}_k[\mathbf{Y}_k[\theta], \theta],$$

where $\mathbf{Y}_k[\theta]$ is the ground-truth radial position of boundary k at angle θ. As in [6], the predicted boundary position $\hat{\mathbf{S}}_k[\theta]$ is obtained using a soft-argmax operation: $\hat{\mathbf{S}}_k[\theta] = \sum_{r=0}^{R-1} r \mathbf{Q}_k[r, \theta]$. To ensure topological constraints, we apply a recursive operation: $\mathbf{S}_{k+1}[\theta] = \mathbf{S}_k[\theta] + \text{ReLU}(\hat{\mathbf{S}}_{k+1}[\theta] - \mathbf{S}_k[\theta])$, where $\mathbf{S}_k$ is the final boundary position for boundary k at angle θ. This ensures that boundaries are ordered correctly (i.e., endocardium $\leq$ scar-start $\leq$ scar-end $\leq$ epicardium).

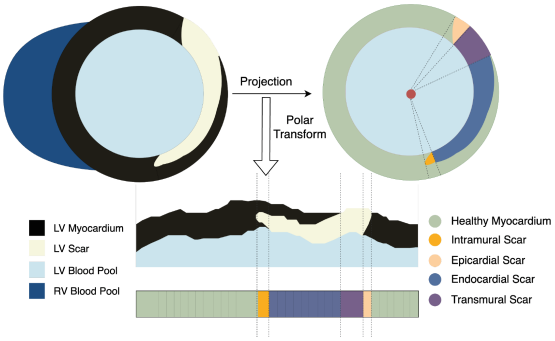


Fig. 3. Scar subtype classification via projection under polar transformation

For the endocardial boundary, epicardial boundary and scar envelope, pseudo Cartesian dice scores are computed to formulate the pseudo dice (p-dice) loss:

$$\mathcal{L}_{\text{pseudo dice}} = 2 - 2\frac{\sum_{\theta=0}^{\Theta-1} \min(\mathbf{Y}_3[\theta], \mathbf{S}_3[\theta])^2 - \max(\mathbf{Y}_2[\theta], \mathbf{S}_2[\theta])^2}{\sum_{\theta=0}^{\Theta-1} (\mathbf{Y}_3[\theta]^2 - \mathbf{Y}_2[\theta]^2) + (\mathbf{S}_3[\theta]^2 - \mathbf{S}_2[\theta]^2)} - \sum_{k=1,4} \frac{\sum_{\theta=0}^{\Theta-1} \min(\mathbf{Y}_k[\theta], \mathbf{S}_k[\theta])^2}{\sum_{\theta=0}^{\Theta-1} \mathbf{Y}_k[\theta]^2 + \mathbf{S}_k[\theta]^2}$$

further enforcing alignment between the predicted and ground truth boundaries across coordinate systems.

2.3 Projection-Based Scar Subtype Classification

The polar segmentation output is projected along the radial axis to characterize the topological distribution of myocardial scar. As illustrated in Fig. 3, this projection enables classification of each angular position θ into healthy myocardium or one of four scar subtypes—transmural, endocardial, epicardial, and intramural—based on whether the radial line intersects the scar region and whether the scar contacts the endocardium and/or epicardium. This subtype classification facilitates quantification of scar topological heterogeneity.

3 Experiments

3.1 Materials

Data Acquisition and Pre-processing. We collected LGE MRI data within one month prior to ICD implantation from 181 post-infarct patients across five centers. All images were resampled to an unified resolution of 1.36 mm $\times$ 1.36 mm, with an intensity normalization via Z-score. The dataset was randomly devided into 126 training, 18 validation and 37 test cases.

Implementation. All experiments were conducted on a workstation equipped with an AMD EPYC 7742 64-Core Processor and an NVIDIA GeForce RTX 3090 GPU. The framework was implemented using the PyTorch library and integrated into the nnU-Net framework for efficient data processing and model training. The training process utilized stochastic gradient descent with a linear decay rate for optimization. All models were trained for 200 epochs with a batch size of 8.

Gold Standard and Evaluation. All LGE MRIs were manually labeled by an experienced cardiologist to delineate the myocardium and scar regions, which served as the ground truth for scar segmentation. The projected scar label was generated based on the manual annotations along 192 evenly spaced radii from the LV centroid, with five classes: endocardial scar, transmural scar, intramural scar, epicardial scar and healthy myocardium.

For the evaluation of scar pattern characterization, we measured performance metrics on the projected cylindrical map. Specifically, Dice, sensitivity (SEN) and specificity (SPE) were calculated for the binary classification map, where all scar subtypes were aggregated to represent the myocardial area at risk. To address the class imbalance inherent in the distribution of topological scar subtypes, we computed micro-averaged SEN and SPE, as well as the generalized Dice score (GDice)—the micro-averaged counterpart of Dice—for the multi-class classification of distinct scar subtypes.

For the evaluation of the conventional scar segmentation task, we computed standard segmentation metrics, including Dice, Hausdorff distance (HD), average symmetric surface distance (ASSD), SEN, and SPE.

3.2 Results

Results on the Task of Scar Subtype Classification and Quantification. Table 1 presents the results of the ablation study (all metrics evaluated in polar coordinates). Regarding the quantification of scar subtypes, the proposed PolarNet achieves the best GDice (0.548 ± 0.134) and SEN (0.561 ± 0.131), the latter being particularly important for screening applications in clinical practice. Compared to nnU-Net, the drop in performance observed in Polar-nnUNet may stem from the loss of fine-grained details introduced during the coordinate transformation. However, this limitation is effectively mitigated by PolarNet's dual-branch architecture, which explicitly incorporates boundary supervision. Notably, the addition of the p-Dice loss further enhances performance, contributing 1.1% and 2.2% gains in mean GDice and SEN, respectively. SPE remains consistently high ($> 94\%$) across all methods, with no statistically significant differences observed. For the overall myocardial area at risk, performance remains comparable or slightly improves after polar transformation. Addition of the boundary branch further enhances Dice and SEN. While p-Dice has limited effect on Dice/SEN, it helps recover SPE ($> 86\%$), which otherwise declines without it.

Figure 4 visualizes the results of scar subtype classification and quantification in a representative subject. Each ring represents a short-axis slice, ordered from

Table 1. Evaluation on the task of scar subtype classification and quantification. Results shown in (mean ± standard deviation). **Bold** indicates best performance, underline indicates second best performance, asterisk (*) indicates the statistically significant difference ($p < 0.05$) given by a Wilcoxon signed-rank test between our proposed model and each comparison model.

Model	Model variants			Scar subtypes			Myocardial area at risk		
	polar	contour	p-dice	GDice	SEN	SPE	Dice	SEN	SPE
nnU-Net [7]	×	×	×	0.542* ±0.125	0.532 ±0.121*	**0.950 ±0.020**	0.793 ±0.165*	0.783 ±0.175	**0.891 ±0.084**
Polar-nnUNet	✓	×	×	0.521 ±0.144*	0.517 ±0.137*	0.944 ±0.022	0.796 ±0.169*	0.795 ±0.169	0.879 ±0.093
PolarNet (w/o $\mathcal{L}_{\text{pseudo dice}}$)	✓	✓	×	0.537 ±0.125*	0.539 ±0.117*	0.943 ±0.020	0.806 ±0.163	**0.832 ±0.152**	0.806 ±0.163
PolarNet (Proposed)	✓	✓	✓	**0.548 ±0.134**	**0.561 ±0.131**	0.942 ±0.019	**0.807 ±0.173**	0.827 ±0.164	0.860 ±0.098

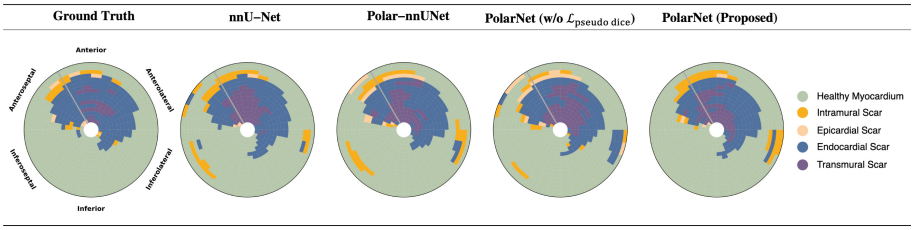


Fig. 4. Qualitative results on a typical subject for the task of scar subtype classification and quantification. Presented in a bullseye-like layout, each ring represents the classification output from a short-axis slice, with color-coded subtypes. Our proposed method (last column) yields more consistent predictions of scar presence (non-green) and improved detection of intramural scars (orange). (Color figure online)

apical (innermost) to basal (outermost). Despite the challenge, our proposed model produces more consistent predictions of scar presence and more accurate identification of non-transmural subtypes—particularly intramural scars, which are key substrates for re-entrant VT. Notably, all methods struggle in the basal inferolateral region, likely due to image blur inherent in LGE MRI.

Results on the Task of Scar Segmentation. Our projection-based subtype classification relies on accurate scar segmentation, making segmentation quality a critical prerequisite. Table 2 summarizes the ablation study results (all metrics evaluated in Cartesian coordinates). Except for HD, the full model outperforms all ablated models, with Dice and SEN showing statistically significant gains, and ASSD improving consistently though not always significantly. The higher HD is not significant and mainly caused by outliers. Each component contributes incrementally to the overall performance. Figure 5 visualizes segmentation results on typical slices. Polar transformation improves the overall ability to distinguish scar regions, while the addition of the boundary branch reduces discontinuities

Table 2. Quantitative evaluation on the task of scar segmentation. Results shown in (mean ± standard deviation). **Bold** indicates best performance, underline indicates second best performance, asterisk (*) indicates the statistically significant difference ($p < 0.05$) given by a Wilcoxon signed-rank test between our proposed model and each comparison model.

Model	Metrics				
	Dice	HD(mm)	ASSD(mm)	SEN	SPE
nnU-Net [7]	0.558 ±0.148*	**13.446 ±13.034**	3.606 ±4.082*	0.552 ±0.165*	0.999 ±0.001
Polar-nnUnet	0.561 ±0.137*	13.516 ±14.826	3.387 ±4.313	0.558 ±0.150*	0.999 ±0.001
PolarNet (w/o $\mathcal{L}_{\text{pseudo dice}}$)	0.572 ±0.137*	13.469 ±15.005	3.379 ±4.990	0.582 ±0.151*	0.999 ±0.001
PolarNet (Proposed)	**0.581 ±0.139**	13.614 ±15.076	**3.340 ±3.934**	**0.599 ±0.150**	0.999 ±0.001

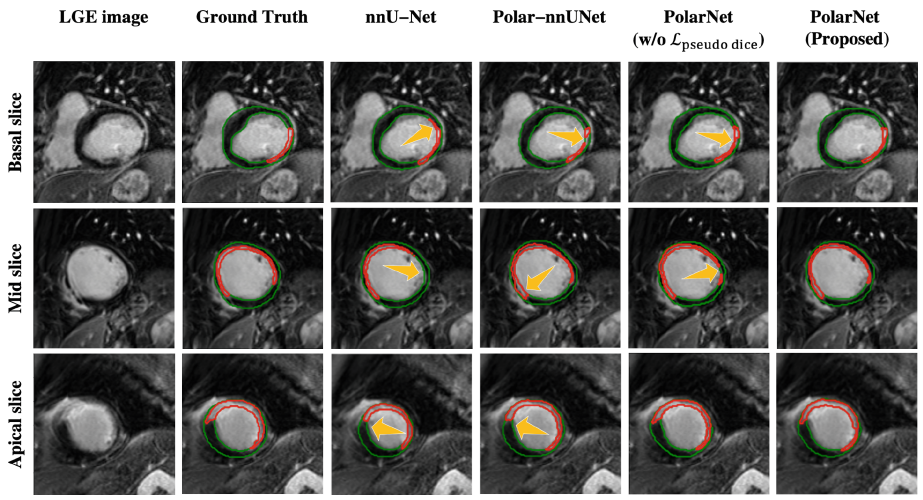


Fig. 5. Qualitative results on typical slices for the task of scar segmentation. Yellow arrows indicate outputs likely to result in implausible topological pattern characterization. The proposed method (last column; scar in red, myocardium in green) yields segmentations more suitable for scar topology analysis, showing better continuity, improved localization (rows 1–2), and sharper edge (row 3). (Color figure online)

and sharpens scar edges. These improvements help explain the superior performance observed in scar subtype classification and quantification.

4 Conclusion

Our study presents an automated framework to address the scar topological pattern characterization in LGE MRI, which is valuable in VT screening. Unlike

conventional segmentation tasks, this work emphasizes the critical examination of intricate topological characteristics of myocardial scar tissue—a challenging endeavor that remains largely unexplored in existing literature. The framework's performance was quantitatively assessed using established metrics including GDice and sensitivity etc., while qualitative validation through visual analysis demonstrated alignment with clinical expert assessments. However, we recognize that this specialized clinical application may require the development of more sophisticated evaluation methodologies beyond the current conventional metrics. It should be noted that the present validation was conducted exclusively on a dataset consisting of patients with LVEF $\leq 35\%$. Future work may extend to multi-center post-infarction cohorts to improve the generalizability of our method and enable deeper analysis for enhanced clinical relevance.

Acknowledgments. This work was funded by the National Natural Science Foundation of China (grant No. 62372115) and Shanghai Municipal Education Commission-Artificial Intelligence Initiative to Promote Research Paradigm Reform and Empower Disciplinary Advancement Plan (grant No. 24KXZNA13). The authors would also like to acknowledge the support from Fédération Français de Cardiologie.

Disclosure of Interests. The authors declare no competing interests.

References

1. Chillou, C., et al.: Magnetic resonance imaging screening for postinfarct life-threatening ventricular arrhythmia. Cardiovasc. Imaging **14**(12), 2479–2481 (2021)
2. Codreanu, A., et al.: Electroanatomic characterization of post-infarct scars. J. Am. Coll. Cardiol. **52**(10), 839–842 (2008)
3. Donahue, J.K., Chrispin, J., Ajijola, O.A.: Mechanism of ventricular tachycardia occurring in chronic myocardial infarction scar. Circ. Res. **134**(3), 328–342 (2024)
4. Fernández-Armenta, J., et al.: Use of myocardial scar characterization to predict ventricular arrhythmia in cardiac resynchronization therapy. EP Europace **14**(11), 1578–1586 (2012)
5. Ghanbari, F., et al.: AI cardiac MRI scar analysis aids prediction of major arrhythmic events in the multicenter DERIVATE registry. Radiology **307**(3), e222239 (2023)
6. He, Y., et al.: Structured layer surface segmentation for retina OCT using fully convolutional regression networks. Med. Image Anal. **68**, 101856 (2021)
7. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
8. Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems, vol. 28, pp. 2017–2025 (2015)
9. Karim, R., et al.: Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement MR images. Med. Image Anal. **30**, 95–107 (2016)
10. Lalande, A., et al.: Deep learning methods for automatic evaluation of delayed enhancement-MRI. The results of the EMIDEC challenge. Med. Image Anal. **79**, 102428 (2022)

11. Lecesne, E., Simon, A., Garreau, M., Barone-Rochette, G., Fouard, C.: Segmentation of cardiac infarction in delayed-enhancement MRI using probability map and Transformers-based neural networks. Comput. Methods Programs Biomed. **242**, 107841 (2023)
12. Li, L., et al.: MyoPS: a benchmark of myocardial pathology segmentation combining three-sequence cardiac magnetic resonance images. Med. Image Anal. **87**, 102808 (2023)
13. Mahida, S., et al.: Cardiac imaging in patients with ventricular tachycardia. Circulation **136**(25), 2491–2507 (2017)
14. Priori, S.G., et al.: 2015 ESC guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: The task force for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death of the European society of cardiology (ESC)endorsed by: Association for European paediatric and congenital cardiology (AEPC). Eur. Heart J. **36**(41), 2793–2867 (2015)
15. Qiu, J., et al.: MyoPS-Net: myocardial pathology segmentation with flexible combination of multi-sequence CMR images. Med. Image Anal. **84**, 102694 (2023)
16. Thomsen, A.F., et al.: Scar border zone mass and presence of border zone channels assessed with cardiac magnetic resonance imaging are associated with ventricular arrhythmia in patients with ST-segment elevation myocardial infarction. Europace European Pacing, Arrhythmias, and Cardiac Electrophysiology: Journal of the Working Groups on Cardiac Pacing, Arrhythmias, and Cardiac Cellular Electrophysiology of the European Society of Cardiology **25**(3), 978–988 (2023)
17. Waks, J.W., Buxton, A.E.: Risk stratification for sudden cardiac death after myocardial infarction. Annu. Rev. Med. **69**(1), 147–164 (2018)
18. Zhang, J., Xie, Y., Liao, Z., Verjans, J., Xia, Y.: EfficientSeg: a simple but efficient solution to myocardial pathology segmentation challenge. In: Zhuang, X., Li, L. (eds.) MyoPS 2020. LNCS, vol. 12554, pp. 17–25. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-65651-5_2

Blind Restoration of High-Resolution Ultrasound Video

Chu Chen[1,2(✉)], Kangning Cui[1,2], Pasquale Cascarano[3], Wei Tang[1,2], Elena Loli Piccolomini[4], and Raymond H. Chan[2,5]

[1] Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong
[2] Hong Kong Centre for Cerebro-Cardiovascular Health Engineering, Hong Kong Science Park, Hong Kong
chuchen4-c@my.cityu.edu.hk
[3] Department of the Arts, University of Bologna, Bologna, Italy
[4] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
[5] Department of Operations and Risk Management and School of Data Science, Lingnan University, Tuen Mun, Hong Kong

Abstract. Ultrasound imaging is widely applied in clinical practice, yet ultrasound videos often suffer from low signal-to-noise ratios (SNR) and limited resolutions, posing challenges for diagnosis and analysis. Variations in equipment and acquisition settings can further exacerbate differences in data distribution and noise levels, reducing the generalizability of pre-trained models. This work presents a self-supervised ultrasound video super-resolution algorithm called Deep Ultrasound Prior (DUP). DUP employs a video-adaptive optimization process of a neural network that enhances the resolution of given ultrasound videos without requiring paired training data while simultaneously removing noise. Quantitative and visual evaluations demonstrate that DUP outperforms existing super-resolution algorithms, leading to substantial improvements for downstream applications.

Keywords: Ultrasound · Video Super-resolution · Deep Image Prior · Self-supervised Learning · Ejection Fraction

1 Introduction

Ultrasound (US) imaging is an essential diagnostic tool in modern medicine, widely utilized in various clinical applications, including cardiology [3,22], obstetrics [2,14], and musculoskeletal [21] imaging. Its ability to provide visualization of internal structures, coupled with its non-invasive nature, makes US a preferred imaging technique. Furthermore, the dynamic recording of US videos captures much more information about tissues than single images, making diagnosis heavily reliant on video analysis. Applications include breast lesion detection [28,31]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 77–87, 2026.
https://doi.org/10.1007/978-3-032-04947-6_8

and cardiovascular monitoring [9,11,24], where metrics like Ejection Fraction (EF) play a critical role. However, the quality of US data often suffers from both inherent and technical limitations. Acoustic noise, caused by scattering and absorption of sound waves in complex tissue structures, degrades image clarity, while limited resolution, stemming from constraints in transducer technology and frequency ranges, affects the ability to capture fine anatomical details. These challenges not only affect diagnostic accuracy but also impact downstream tasks like EF prediction, where precise visualization of cardiac structures and motion is essential. Addressing these limitations requires continuous advancements in US technology to enhance visual quality and diagnostic reliability through super-resolution (SR) algorithms.

Related Works. The development of large training datasets has significantly advanced the research on SR algorithms for natural images [1] and videos [16,19]. These datasets enable supervised training of neural networks, resulting in models that achieve exceptional performance. Among all, deep convolutional neural networks (CNN) [15,30] have shown remarkable capabilities in single image SR (SISR) tasks. To ensure temporal correlation in video enhancement, techniques such as the flow-based [27] and attention-based method [32] are integrated with CNNs. However, significant shifts in data distributions make these pre-trained models struggle with generalization. Consequently, artifacts are often introduced in the upsampled videos.

In the field of US image SR, acquiring paired low-resolution (LR) and high-resolution (HR) datasets is particularly challenging, leading researchers to either create simulated datasets for supervised training or explore unsupervised/self-supervised approaches. Choi et al. [8] modified the SRGAN [12] to enhance B-mode US images with low lateral resolution, improving their similarity to HR images. A U-net-style network [26] is trained on simulated pairs for super-resolved US images. Liu et al. [17] proposed ZSSR-Cycle, a zero-shot and self-supervised generative adversarial network framework for US image SR.

One particularly intriguing development in image processing is the Deep Image Prior (DIP) method [25], which demonstrates that CNN structures can inherently capture and enhance image features without relying on any predefined training dataset. In fact, it has been shown that CNNs can more effectively replicate and enhance the structural characteristics present in images compared to arbitrary noise patterns [6]. In the field of SISR, DIP has achieved competitive performance compared to other supervised training models. Variations of DIP have also been proposed for video processing. The Deep Video Prior (DVP) [13] adopted DIP for individual video frame with iteratively reweighted training strategy to address the multi-modal inconsistencies. The Recursive Deep Prior Video (RDPV) [5] proposed a recursive updating rule for CNN optimization when dealing with each frame for the light Time-Lapse Microscopy video SR. While DIP-based methods effectively capture underlying image structures, their ability to fit fine details can also lead to overfitting, where noise and artifacts are misinterpreted as part of the high-frequency details that SR methods seek

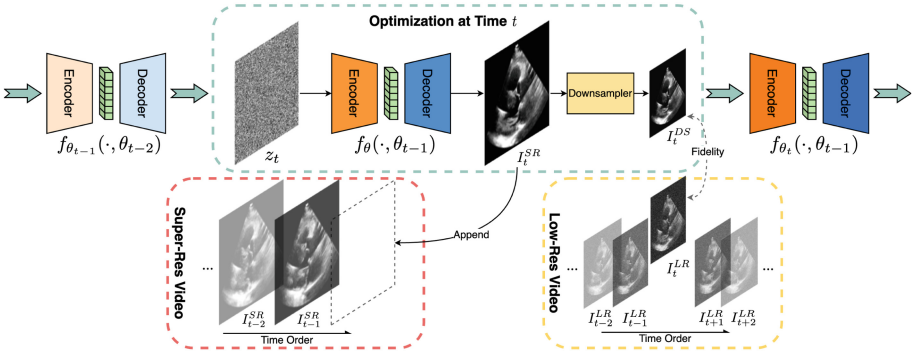


Fig. 1. Schematic of the DUP algorithm.

to enhance. These unwanted patterns adversely affect diagnostic and analytical applications based on the SR-enhanced US videos.

Contributions. US video restoration faces fundamental challenges from lacking paired data, heterogeneous distributions, and multi-factorial corruptions. We address these through a blind restoration method followed by extensive evaluations. The key contributions of this work are as follows:

- We present a self-supervised framework Deep Ultrasound Prior (DUP) for US video SR, eliminating the need for paired training data through video-specific optimization.
- DUP adopts a Weight Inheritance (WIn) strategy and dual regularization for CNN optimization, accelerating convergence and achieving information sharing among successive frames while removing noise and artifacts that appeared in US videos.
- We systematically compare DUP with existing SR techniques under various levels of noise, which demonstrates the superior performance and robustness of our approach in US video restoration.
- By evaluating EF prediction from restored cardiac videos, DUP achieves the lowest error, highlighting its potential to improve clinical diagnostic accuracy.

2 Method

2.1 Problem Formulation

The inverse problem of frame-wise video SR can be formulated as

$$I_t^{LR} = \mathcal{A} I_t^{HR} + e_t, \quad t = 1, 2, \ldots, T, \tag{1}$$

where $\mathcal{A}$ represents the measurement operator, $I_t^{LR} \in \mathbb{R}^{MN}$ is the vectorized tth LR video frame with spatial resolution of $M \times N$, T is the number of frames,

$I_t^{HR} \in \mathbb{R}^{s^2MN}$ is the underlying HR frame with upscaling factor $s \in \mathbb{Z}^+$, and $e_t \in \mathbb{R}^{MN}$ is the measurement error. In this work, we focus on the linear degradation operator $\mathcal{A} \in \mathbb{R}^{MN \times s^2MN}$. Since the SR problem is known to be strongly ill-posed [29], the reconstructed SR frame I_t^{SR} that solves this system is not unique while sensitive to the measurement error e_t. The common approach to address this problem is to add prior constraints on the desired I_t^{SR}, narrowing the space of the possible solution. We formulate the underdetermined systems in Eq. (1) into a set of variational regularized optimization problems:

$$I_t^{SR} = \underset{I_t}{\operatorname{argmin}} \frac{1}{2}||I_t^{LR} - \mathcal{A}I_t||_2^2 + \lambda R(I_t), \quad t = 1, 2, \ldots, T, \tag{2}$$

where $\{I_t\}_{t=1}^T$ are variables. In Eq. (2), the SR video $\{I_t^{SR}\}_{t=1}^T$ is estimated by minimizing the sum of the ℓ_2 fidelity term and the regularization term R for each frame, where λ is the regularization parameter that balances the two terms.

2.2 Deep Ultrasound Prior

The ability to reconstruct and augment detail highlights CNNs' potential for self-supervised learning techniques, improving the spatial resolution and clarity of US video. We now explain our proposed DUP method in detail.

This algorithm produces super-resolved US video in a frame-by-frame manner following the time order. Let integer $t \in [1, T]$ be an index of the video frames and $z_t \in \mathbb{R}^{s^2MN}$ be a random image with targeted spatial resolution. DUP utilizes an encoder-decoder CNN treated as a mapping from the random image z_t to SR frame $I_t^{SR} = f_\theta(z_t, \omega) : \mathbb{R}^{s^2MN} \times \mathbb{R}^\Theta \to \mathbb{R}^{s^2MN}$, where the sub-index θ is the trainable parameters of the CNN and Θ is the total number of parameters, the second entry $\omega \in \mathbb{R}^\Theta$ defined initialization of CNN.

During the reconstruction process of SR images, we aim to enhance only the contrast details while preventing the amplification of random noise and artifacts and even removing them. A widely recognized approach for achieving this is the total variation (TV) model [23], which replaces the regularization term R in Eq. (2) with a TV regularization term R_1:

$$R_1(I_t) = \sum_{i=1}^{s^2MN} (|(D_h I_t)_i| + |(D_v I_t)_i|) \tag{3}$$

where D_h and D_v are the first-order finite difference discrete operators along the horizontal and vertical directions, respectively.

However, optimizing with only total variation regularization can result in staircase effects appearing in the image. To address the piece-wise constant artifacts, we additionally introduce a higher-order (HO) [4,7] term R_2:

$$R_2(I_t) = \frac{1}{2} \sum_{i=1}^{s^2MN} (|(D_h I_t)_i|^2 + |(D_v I_t)_i|^2). \tag{4}$$

Hence, the optimization problem in Eq. (2) is solved by optimizing the parameters of the CNN as

$$\theta_t = \underset{\theta}{\operatorname{argmin}} \frac{1}{2}||I_t^{LR} - \mathcal{A}f_\theta(z_t, \omega)||_2^2 + \lambda_1 R_1(f_\theta(z_t, \omega)) + \lambda_2 R_2(f_\theta(z_t, \omega)), \quad (5)$$

thereby the SR frames can be reconstructed by $I_t^{SR} = f_{\theta_t}(z_t, \omega)$.

The overview of the proposed DUP framework is shown in Fig. 1. We now explain in detail the technical advancement of DUP.

WIn Strategy. For any frame-by-frame video enhancement algorithm, leveraging temporal correlation is essential for maintaining high-quality results. DUP employs a Weight Inheritance (WIn) strategy to utilize the continuity of adjacent frames for improved efficiency, allowing the CNN to converge faster while accumulating optimization iterations. The WIn strategy in DUP operates as follows: 1. The CNN is randomly initialized with parameter θ_0 as processing the first video frame. After optimization based on the first low resolution frame I_1^{LR}, the optimal parameter of the network is θ_1 and corresponding super-resolved image $I_1^{SR} = f_{\theta_1}(z_1, \theta_0)$; 2. For subsequent frames I_t^{LR} $(t \geq 2)$, CNN is initialized with the optimized parameters θ_{t-1} from the preceding frame I_{t-1}^{LR}, and the super-resolved image is generalized by $I_t^{SR} = f_{\theta_t}(z_t, \theta_{t-1})$. This WIn strategy enhances the computational efficiency in frame-by-frame video processing. By sharing network parameters, CNNs utilize information from neighboring frames, exploiting the temporal correlations and achieving faster convergence within a limited number of iterations. Furthermore, continuous weight inheritance means the CNN effectively trains on more samples as it processes additional frames, leading to better SR performance through extended iterations.

Early Stopping and Input Update. Since DUP follows a self-supervised learning strategy, there is no need to consider generalization beyond the given video. However, excessive optimization can introduce unwanted artifacts into the image. Specifically, over-optimizing the fidelity term may lead to overfitting measurement errors, whereas excessive optimization of the regularization term can cause over-smoothing. To mitigate these issues, the iterative process is early-stopped based on a criterion that monitors the loss over a fixed window. Training stops when the loss value stabilizes. With the WIn strategy, the CNN typically requires fewer iterations after processing the first frame, allowing DUP to begin monitoring the loss after fewer training steps.

Furthermore, according to DIP, the input to the CNN is a randomly generated noise image z_1. Due to the strong similarities and periodic patterns exhibited among frames in a US video, using the same noise image as input for every frame (i.e., $z_t = z_1, t \geq 2$) may cause the CNN to undergo lazy training, leading to outputs that remain overly similar to the first frame. In response, DUP updates the noise image for each frame as follows: 1. $z_1 = n_1$; 2. $z_t = z_{t-1} + \sigma n_t, (t \geq 2)$, where each pixel of the random image $\{n_t\}_{t=1}^T \in \mathbb{R}^{s^2MN}$ is $i.i.d$ sampling from $\mathcal{N}(0, 1)$, and σ is the standard deviation of the added noise

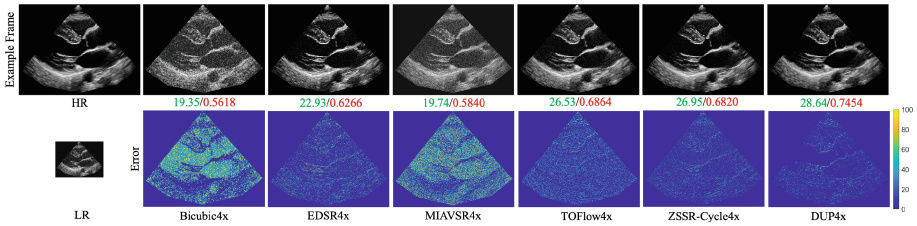


Fig. 2. Visual comparisons of 4× SR results. The paired HR and LR example frames are shown on the far left. The upper row displays the SR frames from various methods, with quantitative results (PSNR (dB)/SSIM) shown underneath, while the lower row is the corresponding error maps.

in each update. This formulation allows the CNN to learn a mapping from the random noise input to the residual between successive frames.

Further implementation details are provided in the Sect. 3.2.

3 Experiments

3.1 Datasets

Our validation combines video quality assessment and downstream tasks using two datasets: 1) EchoNet-LVH [10] (1024×768 resolution, 90–200 frames/clip) for restoration evaluation. We simulate LR scenarios through normalization and downsampling with Gaussian noise of various standard deviations (std 0.01–0.1), comparing outputs against original HR videos. 2) EchoNet-Dynamic [20] (112×112 resolution, 28-1002 frames/clip) for application testing. We randomly selected 100 videos from each dataset for the respective evaluation scenarios.

3.2 Implement Details

In this part, we explain the implementation of our method and experiments. A Lanczos kernel is used for performing the degradation process $\mathcal{A}$, which aligns with the real-world US image degradation characteristics [18]. The CNN is an Encoder-Decoder-style architecture featuring skip concatenation. The architecture consists of four encoder and decoder base units each, which perform convolutions using 128 feature maps, along with batch normalization layers and Leaky ReLU activations. Downsampling is achieved via convolutional layers with a stride of two, while up-sampling is by a Lanczos operator. For the first video frame, we allow up to 3000 iterations, implementing early stopping after 2000 iterations with a patience of 100. For subsequent frames, we reduce the maximum to 2000 iterations, with early stopping commencing at 1000 iterations and patience of 50. σ is set to 0.03 and $\lambda_1 = \lambda_2 = 0.01$ in Eq. (5). For the downstream task, we employ EchoCLIP [9], a CLIP-like vision-language model trained on medical reports paired with US videos. The restored videos (each frame resized

Table 1. Ablation studies.

Method	*HO*	*TV*	PSNR (dB)↑	SSIM↑
DIP (w/o WIn)	✗	✗	25.53	0.7133
	✓	✗	27.16	0.7432
	✗	✓	28.49	0.7432
	✓	✓	28.63	0.7440
DUP	✗	✗	25.08	0.7069
	✓	✗	26.85	0.7409
	✗	✓	28.85	0.7722
	✓	✓	**28.99**	**0.7740**

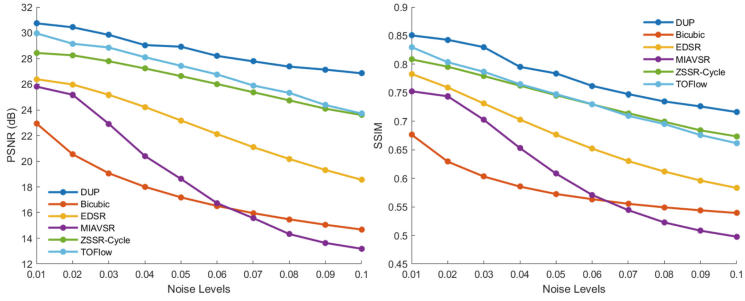


Fig. 3. Quantitative evaluations on 4× SR videos corrupted by multi-level noise.

to 224 × 224) generated by different SR methods are fed into EchoCLIP to perform zero-shot EF prediction. All experiments were conducted in Python 3.8.17 on a PC equipped with Intel® Xeon® Silver 4210 Processor CPU 2.20 GHz and Nvidia GeForce RTX 3090 GPU with 24G of memory.

3.3 Image-Based Evaluations

We compared DUP against four categories of methods: (1) Bicubic interpolation, (2) SISR learning-based architectures (EDSR [15] and DIP [25]), (3) video SR-based methods (TOFlow [27] and MIAVSR [32]), and (4) SR framework for US (ZSSR-Cycle [17]). The visual comparison results in Fig. 2, based on an LR video with 0.05 std of additive noise, reveal that DUP notably outperforms other methods, specifically in reconstructing clear textures of the ventricles with significantly lower error. In contrast, the results from other methods exhibit noticeable errors after restoration, compromising image clarity. This visual superiority aligns with the quantitative metrics, where DUP attains the highest peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) values. To evaluate the robustness against noise as demonstrated in Fig. 3, DUP consistently achieves the best reconstruction performance under different levels of noise corruption, while other baselines degrade rapidly. From both quantitative and qualitative

Table 2. EF Prediction Error

SR Method	Bicubic	EDSR	RDN	ZSSR-Cycle	DUP
Mean Error (%) ↓	10.26	8.69	8.46	7.60	**7.23**

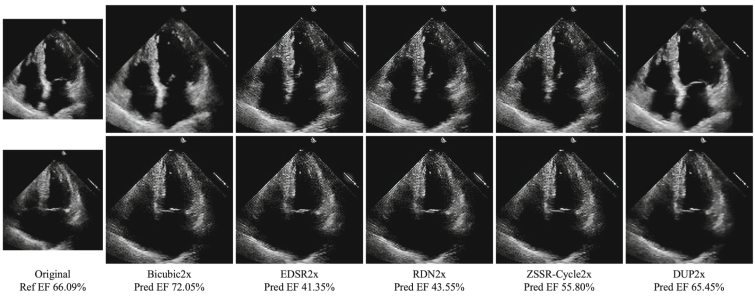


Fig. 4. Visualization of maximum diastole (upper) and minimum systole (lower) detected within cardiac cycles.

aspects, DUP yields a superior result with rich details and less noise. In order to analyze the role of regularizations and WIn, we conduct ablation studies on these setups. As listed in Table 1, the best SNR and structural detail are achieved when both HO and TV priors are incorporated into the target loss, regardless of the presence of WIn. Without WIn, DUP defaults to the same training strategy as DIP, resulting in more iterations(3000 iters/frame) and longer optimization time. Despite this, DUP with the WIn strategy outperforms and achieves greater efficiency.

3.4 Downstream Evaluations

To evaluate the practical utility of our proposed method, we assess its performance on the downstream task of EF prediction. Since the EF prediction model only accepts 2× SR results and both MIAVSR and TOFlow are limited to 4×, we substituted these two baselines with RDN [30] for this part of the study. Table 2 compares the mean prediction error of EF across different SR preprocessing methods, where DUP achieves the lowest mean error of 7.23%. This demonstrates the potential of DUP in enhancing the quality of US videos for diagnostic tasks. Figure 4 visualizes a sample EF prediction result with detected minimum systole and maximum diastole frames within cardiac cycles. EchoCLIP identified frames corresponding to the two heart phases that align with the manually annotated ground truth from the DUP SR results, while those extracted from other SR videos displayed discrepancies. Consequently, the improvement in visual quality directly correlates with the lower error in EF prediction observed in Fig. 4. The sample further validates the practical applicability of DUP in clinical settings.

4 Conclusion

In this work, we propose a self-supervised learning method, Deep Ultrasound Prior (DUP), for US video restoration, eliminating the need for HR video supervision. DUP increases spatial resolution while using CNNs' implicit regularization to capture detailed features accurately. By incorporating the WIn strategy with an early stopping and input update mechanism, the network can perceive and share local frame information while improving convergence speed and performance. With the addition of TV regularization and HO terms, DUP effectively eliminates the measurement error (e.g., noise and artifacts) commonly found in US videos. We demonstrate the superiority of DUP by comparing the quality of super-resolved US videos and their robustness against noise with state-of-the-art methods. Ablation studies confirm the necessity of the WIn strategy and explicit regularization terms for effective video restoration. Furthermore, we validate DUP's capabilities as a pre-processing method in downstream tasks, showing that the restored videos enhance the accuracy of EF predictions for cardiac. This work establishes DUP as a comprehensive solution for US video restoration and overcoming the bottlenecks of its clinical applications.

Acknowledgments. This work is partially supported by HKRGC (grant number CityU11301120, C1013-21GF, CityU11309922, CityU9380162), ITF (grant number LU BGR 105824, MHP/054/22), and the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126–135 (2017)
2. Alfirevic, Z., Stampalija, T., Medley, N.: Fetal and umbilical doppler ultrasound in normal pregnancy. Cochrane Database Syst. Rev. (2015)
3. Aly, I., et al.: Cardiac ultrasound: an anatomical and clinical review. Transl. Res. Anat. **22**, 100083 (2021)
4. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imag. Sci. **3**(3), 492–526 (2010)
5. Cascarano, P., Comes, M.C., Mencattini, A., Parrini, M.C., Piccolomini, E.L., Martinelli, E.: Recursive deep prior video: a super resolution algorithm for time-lapse microscopy of organ-on-chip experiments. Med. Image Anal. **72**, 102124 (2021)
6. Cascarano, P., Franchini, G., Porta, F., Sebastiani, A.: On the first-order optimization methods in deep image prior. J. Verif. Valid. Uncertain. Quantif. **7**(4), 041002 (2022)
7. Chan, T., Marquina, A., Mulet, P.: High-order total variation-based image restoration. SIAM J. Sci. Comput. **22**(2), 503–516 (2000)

8. Choi, W., Kim, M., HakLee, J., Kim, J., BeomRa, J.: Deep CNN-based ultrasound super-resolution for high-speed high-resolution b-mode imaging. In: 2018 IEEE International Ultrasonics Symposium (IUS), pp. 1–4 (2018)
9. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nat. Med., 1–8 (2024)
10. Duffy, G., et al.: High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. JAMA Cardiol. **7**(4), 386–395 (2022)
11. Duffy, G., et al.: High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. JAMA Cardiol. **7**(4), 386–395 (2022)
12. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
13. Lei, C., Xing, Y., Ouyang, H., Chen, Q.: Deep video prior for video consistency and propagation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 356–371 (2022)
14. Leung, K.Y.: Applications of advanced ultrasound technology in obstetrics. Diagnostics **11**(7), 1217 (2021)
15. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144 (2017)
16. Liu, C., Sun, D.: On Bayesian adaptive video super resolution. IEEE Trans. Pattern Anal. Mach. Intell. **36**(2), 346–360 (2013)
17. Liu, H., Liu, J., Hou, S., Tao, T., Han, J.: Perception consistency ultrasound image super-resolution via self-supervised cycleGAN. Neural Comput. Appl., 1–11 (2021)
18. Meijering, E.H.W., Niessen, W.J., Pluim, J.P.W., Viergever, M.A.: Quantitative comparison of sinc-approximating kernels for medical image interpolation. In: Taylor, C., Colchester, A. (eds.) MICCAI 1999. LNCS, vol. 1679, pp. 210–217. Springer, Heidelberg (1999). https://doi.org/10.1007/10704282_23
19. Nah, S., et al.: NTIRE 2019 challenge on video deblurring and super-resolution: dataset and study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
20. Ouyang, D., et al.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580** (2020)
21. Patil, P., Dasgupta, B.: Role of diagnostic ultrasound in the assessment of musculoskeletal diseases. Ther. Adv. Musculoskelet. Dis. **4**(5), 341–355 (2012)
22. Pellicori, P., et al.: Ultrasound imaging of congestion in heart failure: examinations beyond the heart. Eur. J. Heart Fail. **23**(5), 703–712 (2021)
23. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1–4), 259–268 (1992)
24. Tang, W., Cui, K., Chan, R.H., Morel, J.M.: Bilateral signal warping for left ventricular hypertrophy diagnosis. arXiv preprint arXiv:2411.08819 (2024)
25. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
26. Van Sloun, R.J., Solomon, O., Bruce, M., Khaing, Z.Z., Eldar, Y.C., Mischi, M.: Deep learning for super-resolution vascular ultrasound imaging. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1055–1059. IEEE (2019)
27. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. Int. J. Comput. Vision **127**, 1106–1125 (2019)
28. Yoon, J.H., et al.: Validation of the fifth edition BI-RADS ultrasound lexicon with comparison of fourth and fifth edition diagnostic performance using video clips. Ultrasonography **35**(4), 318 (2016)

29. Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., Zhang, L.: Image super-resolution: the techniques, applications, and future. Signal Process. **128**, 389–408 (2016)
30. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018)
31. Zhao, G., Kong, D., Xu, X., Hu, S., Li, Z., Tian, J.: Deep learning-based classification of breast lesions using dynamic ultrasound video. Eur. J. Radiol. **165**, 110885 (2023)
32. Zhou, X., Zhang, L., Zhao, X., Wang, K., Li, L., Gu, S.: Video super-resolution transformer with masked inter&intra-frame attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 25399–25408 (2024)



Bridging Classical and Learning-Based Iterative Registration Through Deep Equilibrium Models

Yi Zhang, Yidong Zhao, and Qian Tao(✉)

Department of Imaging Physics, Delft University of Technology, Delft, The Netherlands
{y.zhang-43,y.zhao-8,q.tao}@tudelft.nl

Abstract. Deformable medical image registration is traditionally formulated as an optimization problem. While classical methods solve this problem iteratively, recent learning-based approaches use weight-tied neural networks to mimic this process by unrolling the prediction of deformation fields in a fixed number of steps. However, classical methods typically converge after sufficient iterations, but learning-based unrolling methods lack a theoretical convergence guarantee and show instability empirically. In addition, unrolling methods have a practical bottleneck at training time: GPU memory usage grows linearly with the unrolling steps due to backpropagation through time (BPTT). To address both theoretical and practical challenges, we propose DEQReg, a novel registration framework based on Deep Equilibrium Models (DEQ), which formulates registration as an equilibrium-seeking problem, establishing a natural connection between classical optimization and learning-based unrolling methods. DEQReg maintains constant memory usage, enabling theoretically unlimited iteration steps. Through extensive evaluation on the public brain MRI and lung CT datasets, we show that DEQReg can achieve competitive registration performance, while substantially reducing memory consumption compared to state-of-the-art unrolling methods. We also reveal an intriguing phenomenon: the performance of existing unrolling methods first increases slightly then degrades irreversibly when the inference steps go beyond the training configuration. In contrast, DEQReg achieves stable convergence with its inbuilt equilibrium-seeking mechanism, bridging the gap between classical optimization-based and modern learning-based registration methods.

Keywords: Deformable Image Registration · Deep Equilibrium Models

1 Introduction

Deformable image registration (DIR) is a fundamental task in medical image analysis that aims to discover the spatial correspondence between pairs of images

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 88–98, 2026.
https://doi.org/10.1007/978-3-032-04947-6_9

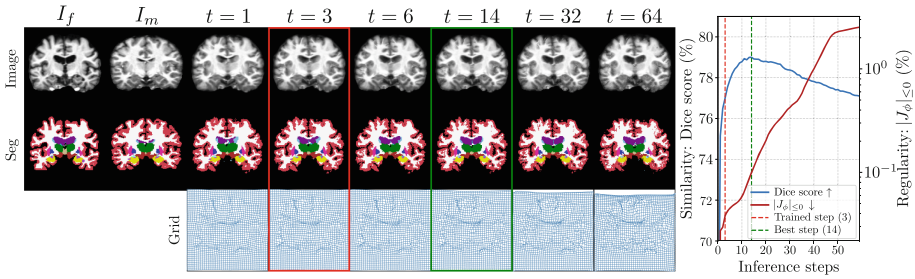


Fig. 1. Performance of a learning-based unrolling method (3 resolution level, 3 steps per level, 9 steps in total). Inference steps t denotes steps on the highest level.

[14,27,28]. DIR can be formulated as an optimization problem to find a parameterized deformation field that aligns the images. The objective typically comprises two terms: image similarity for alignment accuracy, and regularization for deformation smoothness and anatomical plausibility. Historically, this optimization problem has been solved by iterative algorithms such as adaptive gradient descent [2,3,20]. While successful, the iterative nature of these methods leads to long computation time seeking convergence, limiting their real-time clinical applications. With recent advances in deep learning (DL), data-driven approaches have emerged as a promising alternative, by directly inferring the transformation field in a single forward-pass through parametrized neural networks [6,8,23]. More recently, inspired by classical iterative methods, unrolling methods [9,15,23,24,29,30] were proposed to learn the optimization process in multiple steps. Compared to single-step inference, these unrolling methods demonstrated further improved performance [9,24,30].

Despite improved accuracy and generalizability over single-step inference, a fundamental gap exists between learning-based unrolling methods and classical optimization methods. While classical optimization methods pursue convergence for optimal performance and often require >100 iterations, current unrolling methods rely on a limited number of steps (<10) at both training and inference time. At training time, this limitation stems from the reliance on back-propagation through time (BPTT) [26], which causes a linearly growing memory footprint that quickly surpasses the practical GPU capacity. At inference time, extended iterations can be made beyond the training steps, but existing implementations maintain the same iteration steps in both phases. This raises a curious question as to what would happen when inference steps go beyond training ones, which we show in Fig. 1. It can be observed that extended inference steps initially improve the Dice score and then degrade it, while the deformation regularity expressed by percentage of negative Jacobian determinants of deformation field (folded voxels) deteriorates consistently. The empirical results show that learning-based unrolling methods lack a convergence guarantee, which is well-established in classical methods and highly desirable for DIR.

To bridge this conceptual and empirical gap, we propose DEQReg, which seeks a converged estimation of registration fields based on recent Deep Equilibrium Models (DEQ) [5]. DEQReg is the first attempt to formulate the entire learning-based registration process as an end-to-end unsupervised fixed-point problem. It offers both theoretical and practical advantages: first, by explicitly seeking equilibrium points during iterative registration, it establishes a natural connection between classical optimization and learning-based unrolling methods. Second, DEQ's implicit differentiation [5,21] enables potentially unlimited iteration steps with constant memory usage, bypassing the memory limitations of BPTT. Unlike recent works that apply DEQ as a post-processing step for pre-trained models [18] and semi-supervised registration [19], or as a solver for supervised regression and inverse problems [4,13], our work presents two new contributions for DIR:

- We propose DEQReg, the first end-to-end unsupervised registration framework that formulates DIR as a fixed-point equilibrium problem. Extensive experiments on the public brain MRI and lung CT datasets demonstrate that our approach achieves competitive accuracy with constant memory complexity through implicit differentiation.
- We empirically demonstrate that DEQReg achieves stable convergence at inference time while revealing an intriguing phenomenon in current unrolling methods: Their performances are bound to deteriorate after initial slight improvement beyond the predefined training step, highlighting the significance of our principled equilibrium-seeking design.

2 Methods

2.1 Iterative Learning-Based Image Registration via Unrolling

Given a fixed image I_f and a moving image I_m, DIR aims to find a dense displacement field $\mathbf{u}$ that aligns I_m with I_f. The deformation field ϕ is defined as $\phi(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$ for any spatial coordinate $\mathbf{x}$. For notational simplicity, we use ϕ hereafter. DIR can be formulated as an optimization problem:

$$\phi^* = \underset{\phi}{\operatorname{argmin}}\, \mathcal{L}_{\text{sim}}(I_f, I_m \circ \phi) + \lambda \mathcal{L}_{\text{reg}}(\phi), \tag{1}$$

where $\mathcal{L}_{\text{sim}}$ measures image dissimilarity, $\mathcal{L}_{\text{reg}}$ enforces smoothness of the deformation, and λ balances these competing terms. Classical methods typically solve this optimization through gradient descent:

$$\phi_{t+1} = \phi_t - \eta_t \nabla_\phi (\mathcal{L}_{\text{sim}}(I_f, I_m \circ \phi_t) + \lambda \mathcal{L}_{\text{reg}}(\phi_t)), \quad t = 0, 1, \ldots, L, \tag{2}$$

where η_t denotes an adaptive step size. The optimization typically continues until convergence or reaches a maximum iteration count L, which often exceeds 100 steps for sufficient optimization. In iterative learning-based registration, this optimization is often reformulated by unrolling into a fixed number of steps T.

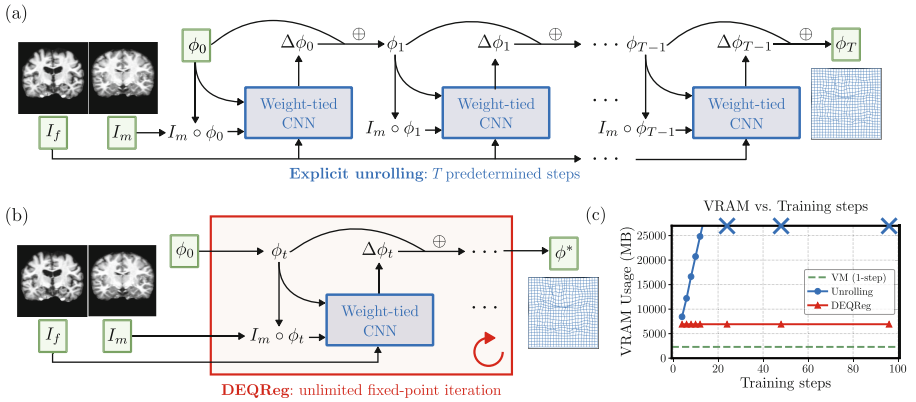


Fig. 2. (a) Illustration of a typical learning-based registration method that uses explicit unrolling of T steps in training. (b) DEQReg uses implicit fixed-point iteration, enabling theoretically unlimited steps in training. (c) Memory usage when training on a 3D brain MRI pair, where blue crosses indicate memory overflow for a 24 GB GPU. (Color figure online)

The update rule is modeled using a weight-tied neural network as shown in Fig. 2, where parameters θ are shared across steps to learn the optimal update function f_θ:

$$\phi_{t+1} = \phi_t + f_\theta(I_f, I_m, \phi_t), \quad t = 0, 1, \ldots, T-1. \tag{3}$$

The update function f_θ is learned by minimizing the expected loss over a training dataset $\mathcal{D} = \{(I_f^i, I_m^i)\}_{i=1}^M$:

$$\min_\theta \mathbb{E}_{(I_f, I_m)\sim\mathcal{D}}[\mathcal{L}(I_f, I_m, \phi_T) + \sum_{t=0}^{T-1} w_t \mathcal{L}(I_f, I_m, \phi_t)], \tag{4}$$

where $w_t > 0$ is the weight for the correction of intermediate states.

Training such unrolling networks through BPTT [26] requires storing *all* intermediate states, leading to $\mathcal{O}(T)$ memory usage and typically restricting steps to $T < 10$ due to hardware constraints.

2.2 Deep Equilibrium Models for Deformable Image Registration

To overcome the limitations of unrolling approaches that can only allow for a fixed number of iterations, we reformulate the iterative process using the Deep Equilibrium Model (DEQ) [5] to find the equilibrium of a fixed-point equation. In its general form, a fixed point $\mathbf{x}^*$ of a function g satisfies $\mathbf{x}^* = g(\mathbf{x}^*)$. Here, we design our network f_θ to learn residual updates, where DEQ seeks an equilibrium solution ϕ^* through residual updates that satisfy this fixed-point condition:

$$\phi^* = g_\theta(I_f, I_m, \phi^*) = \phi^* + f_\theta(I_f, I_m, \phi^*), \tag{5}$$

which is equivalent to finding ϕ^* where $f_\theta(I_f, I_m, \phi^*) = 0$. This formulation provides a principled way to enforce convergence through vanishing updates, analogous to classical optimization but with the learned update by f_θ.

Unlike BPTT that requires storing and backpropagating throughout the forward trajectory, a key advantage of DEQ is its ability to compute derivatives directly from the derived equilibrium state. When the system converges to a fixed point ϕ^* where $f_\theta(I_f, I_m, \phi^*) = 0$, we can leverage the implicit function theorem (IFT) [5,21] to compute gradients without tracking intermediate states:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \phi^*} \left(I - \frac{\partial g_\theta}{\partial \phi^*} \right)^{-1} \frac{\partial g_\theta(I_f, I_m, \phi^*)}{\partial \theta}, \tag{6}$$

where I is an identity matrix of the same size as $\frac{\partial g_\theta}{\partial \phi^*}$. Leveraging IFT, equilibrium points can be derived using any black-box root-finding algorithm, *e.g.* Broyden's method [7]. Figure 2 provides a schematic comparison between DEQ and existing unrolling methods. By focusing on the equilibrium state rather than the iteration trajectory, DEQ achieves constant memory complexity while enabling theoretically unlimited iterations, with convergence design (i.e. fixed point condition).

2.3 Training the DEQ

Despite the memory efficiency of DEQ, directly computing the inverse Jacobian term $(I - \frac{\partial g_\theta}{\partial \phi^*})^{-1}$ in Eq. (6) remains computationally intensive. To address this challenge, we use the phantom gradient approach [11,12], which offers a structured approximation of IFT that preserves the same descent direction while being computationally efficient. The key idea is to approximate the inverse Jacobian through a sequence of damped iterations. Starting from the converged fixed point $\phi^0 = \phi^*$, we generate this sequence with a damping factor $\tau \in (0, 1)$:

$$\phi^{p+1} = \tau g_\theta(I_f, I_m, \phi^p) + (1 - \tau)\phi^p, \quad p = 0, \ldots, K - 1. \tag{7}$$

Using the generated sequence $\{\phi^p\}_{p=0}^{K-1}$, we can approximate the gradient as:

$$\frac{\partial \mathcal{L}}{\partial \theta} \approx \frac{\partial \mathcal{L}}{\partial \phi^*} \mathbf{A}, \quad \text{where} \quad \mathbf{A} = \tau \sum_{k=0}^{K-1} \prod_{s=k+1}^{K-1} \left(\tau \frac{\partial g_\theta}{\partial \phi} \bigg|_{\phi^s} + (1 - \tau) I \right) \frac{\partial g_\theta}{\partial \theta} \bigg|_{\phi^k}. \tag{8}$$

This truncated gradient converges to the exact implicit gradient in Eq. 6 as $K \to \infty$ under mild conditions (see [12] for detailed theoretical analysis). Generally, larger K adds higher-order Neumann terms as in Eq. 8, improving estimation accuracy at a cost of $\mathcal{O}(K)$ total memory consumption. In the case of K=1, the estimation recovers Jacobian-free backpropagation [10], with $\mathcal{O}(1)$ memory complexity but less stability. In DEQReg, we set $K = 2$ to balance accuracy and efficiency.

Additionally, we can leverage intermediate state correction during the forward process to regularize the iteration behavior [4,9,29] in training. Given a

training trajectory $[\phi_0, \phi_1, \ldots, \phi_N, \phi^*]$, we sample S intermediate states with equal intervals, where $\mathcal{S} \subset \{0, 1, \ldots, N\}$ with $|\mathcal{S}| = S$, to compose the total loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}(I_f, I_m, \phi^*) + \gamma \sum_{t \in \mathcal{S}} \mathcal{L}(I_f, I_m, \phi_t), \quad \gamma > 0, \tag{9}$$

The intermediate state gradients are computed via phantom gradients, enabling effective training of the equilibrium-seeking network f_θ.

3 Experiments

Datasets. We evaluated our method on two public datasets: (1) **OASIS** [22], a brain MRI dataset containing 414 T1-weighted scans for inter-subject registration. Following [17], scans were preprocessed and resampled to a size of $128 \times 128 \times 128$, with 30 anatomical segmentation labels (300/30/84 for train/val/test). (2) **NLST** [1], a lung CT dataset from Learn2Reg challenge [16] with 210 intra-subject inhale-exhale pairs. The affinely prealigned scans are of size $224 \times 192 \times 224$, with keypoint annotations and lung masks (170/10/30 for train/val/test).

Evaluation Metrics. We evaluate both anatomical alignment and deformation smoothness. For alignment, we compute the average Dice score and Hausdorff distance (HD) on OASIS using the structures in [17], and keypoint target registration error (TRE) and lung Dice score on NLST. For deformation smoothness, we compute two Jacobian-based metrics ($J_\phi = \nabla\phi$): the percentage of non-diffeomorphic (folded) voxels $|J_\phi|_{\leq 0}$ and the standard deviation of log-Jacobian determinant $\text{std}(\log |J_\phi|)$ quantifying the spatial consistency of volume changes.

Baseline Methods. We compare our method against the following approaches: (1) **elastix** [20], a widely-used classical registration method, configured with 3-level B-spline transformation and grid spacing of 4 voxels; (2) **VoxelMorph** [6], a well-established one-step DL method; and two state-of-the-art unrolling methods: (3) **GraDIRN** [24], which integrates explicit similarity gradients alongside neural network predictions in 3 resolution levels with $T = 3$ per level ($w_t = 0$), and (4) **RIIR** [29] with single resolution and $T = 6$ iterations using exponentially increasing weights ($w_t = 10^{\frac{t-1}{T-1}}$). Our proposed DEQReg employs a fixed-point iteration strategy with maximum $T = 48$ steps with $S = 3$ and $\gamma = 0.5$ in Eq. 9.

Convergence Analysis Setup. In this study, we investigate how DL-based unrolling methods behave when varying inference steps while keeping training steps fixed. Besides GraDIRN and RIIR, we evaluate two additional variants: (1) Unrolling (the vanilla version of GraDIRN without incorporating similarity gradients) and (2) DEQReg-12, DEQReg with $T = 12$ and $S = 0$. For GraDIRN and Unrolling, we vary the steps at the highest resolution.

Implementation Details. For fair comparison, we use the same U-Net [25] configuration as VoxelMorph and RIIR for DEQReg. For each step t, our network takes concatenated $[I_f, I_m \circ \phi_t, \phi_t]$ as input, consistent with other iterative

Table 1. Quantitative comparison of registration methods. Mem. denotes GPU memory usage and Time shows training/inference time per pair. Best performance is in **bold** and second best is underlined, with * denotes statistical significance ($p < 0.05$, Wilcoxon).

OASIS	Dice (%)↑	HD (mm)↓	$\lvert J_\phi\rvert_{\leq 0}$ (%)↓	std($\log \lvert J_\phi\rvert$) ↓	Mem. (MB)	Time (s)
Affine	61.2(7.3)	8.56(1.34)	–	–	–	–
elastix	75.2(3.8)	2.89(0.58)	0.068(0.041)	0.551(0.134)	–	35.45
VoxelMorph	77.8(3.7)	2.74(0.49)	0.063(0.031)	0.601(0.121)	2645	0.20/0.14
GraDIRN	80.3(3.2)	2.47(0.41)	0.071(0.029)	0.568(0.114)	16619	0.59/0.24
RIIR	80.6(3.1)	**2.43**(0.39)	0.058(0.031)	0.556(0.118)	12542	0.60/0.30
DEQReg	**80.7***(3.1)	2.43(0.40)	**0.053***(0.033)	**0.542***(0.123)	6925	1.09/0.98
NLST	TRE (mm)↓	Dice (%)↑	$\lvert J_\phi\rvert_{\leq 0}$ (%)↓	std($\log \lvert J_\phi\rvert$) ↓	Mem. (MB)	Time (s)
Affine	8.43(3.97)	87.3(4.1)	–	–	–	–
elastix	3.24(1.74)	94.8(0.7)	0.110(0.136)	0.636(0.370)	–	32.51
VoxelMorph	4.08(2.31)	95.6(1.2)	0.143(0.290)	0.671(0.506)	1522	0.24/0.12
GraDIRN	2.26(1.87)	96.5(1.2)	0.165(0.278)	0.738(0.512)	9313	0.41/0.18
RIIR	**2.21***(1.78)	96.5(1.1)	0.120(0.250)	0.598(0.481)	7029	0.40/0.16
DEQReg	2.23(2.11)	**96.6**(1.5)	**0.079***(0.078)	**0.586***(0.290)	3983	0.74/0.53

methods. The loss combines local normalized cross correlation (LNCC) with window size 5 as $\mathcal{L}_{\text{sim}}$ and diffusion regularization as $\mathcal{L}_{\text{reg}}$, with $\lambda = 0.1$ (OASIS) and 0.125 (NLST). Models were trained using AdamW ($\text{lr} = 1 \times 10^{-4}$) for 100 epochs. Code and pretrained weights are made publicly available[1].

4 Results and Discussions

Main Results. Table 1 presents a comprehensive comparison of our DEQReg against classical and learning-based approaches, with visual results shown in Fig. 4. DEQReg achieves a statistically significant improvement in Dice scores while maintaining competitive HD on OASIS, and demonstrates comparable performance in TRE and Dice score on NLST. Notably, by decomposing deformation into more steps, DEQReg achieves the lowest percentage of $\lvert J_\phi\rvert_{\leq 0}$ and std($\log \lvert J_\phi\rvert$), indicating superior deformation regularity. Though DEQReg requires relatively higher inference time, this can be mitigated by using solver acceleration, which can largely reduce iterations. Remarkably, DEQReg substantially reduces memory consumption compared to other DL methods.

Convergence Analysis. Figure 3 and Table 2 present the convergence behavior of different methods. The vanilla unrolling method shows a severe performance drop beyond its trained steps. Although GraDIRN and RIIR show modest improvement when inference steps exceed training configuration due to their image similarity-driven updates, they eventually suffer from deteriorating deformation regularity (shown in Fig. 4, row 5) and Dice. In contrast, both DEQReg

[1] https://gitlab.tudelft.nl/ai4medicalimaging/deqreg.

Table 2. Comparison of registration performance at trained steps versus best achievable steps on validation set. Best performance is in **bold** and second best is underlined.

Method	Trained Steps			Best Steps		
	Steps	Dice (%)	$\lvert J_\phi\rvert_{\leq 0}$ (%)	Steps	Dice (%)	$\lvert J_\phi\rvert_{\leq 0}$ (%)
Unrolling	3	77.95(4.67)	0.064(0.031)	3	77.95(4.67)	0.064(0.031)
GraDIRN	3	78.69(4.20)	0.057(0.029)	14	79.44(3.57)	0.078(0.033)
RIIR	6	79.30(3.72)	0.066(0.031)	12	79.39(3.58)	0.062(0.028)
DEQReg-12	12	78.19(4.68)	0.077(0.041)	96	**79.53**(3.62)	0.087(0.039)
DEQReg	48	**79.33**(3.97)	**0.053**(0.030)	96	79.45(3.84)	**0.056**(0.031)

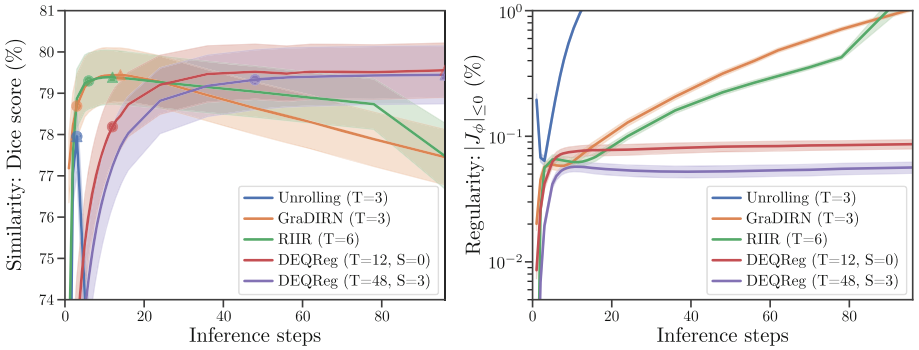


Fig. 3. Convergence analysis for different registration methods. Left: Dice score (%) where • indicates the results at the trained step count and ▲ marks the steps achieving best Dice performance; Right: negative Jacobian determinant ratio (%). Shaded areas show standard deviation. DEQReg exhibits convergence beyond training steps.

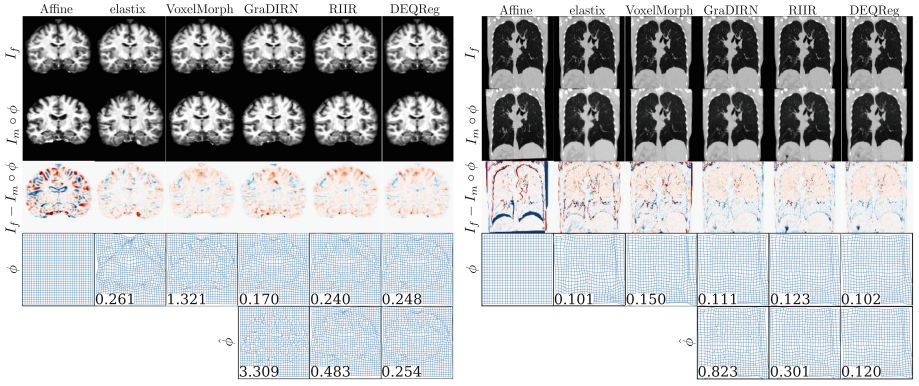


Fig. 4. Qualitative comparison of registration methods on brain MRI and lung CT images. In the last row, all unrolling methods go to the excessive step count of $T = 96$ to evaluate the convergence of ϕ to $\hat{\phi}$. The percentage of $|J_\phi|_{\leq 0}$ is reported at the bottom-left corner to indicate deformation regularity.

variants demonstrate stable convergence. Notably, DEQReg shows potential for improvement with increased iterations, as its optimal Dice occur at maximum steps. Comparing the gap between trained-step and best-achievable performance, methods with intermediate state correction as in Eq. 9 (RIIR and DEQReg) achieve near-optimal results at trained steps, confirming their effectiveness. In particular, the full DEQReg consistently maintains superior deformation regularity while achieving competitive accuracy, as shown in Fig. 3.

5 Conclusion

For deformable medical image registration, classical optimization methods typically converge, but modern learning-based unrolling methods lack the theoretical guarantee of convergence and suffer from memory bottlenecks. This creates a gap between the principled convergence of classical methods and the lack of convergence in learning-based methods. In this work, we propose DEQReg to bridge this gap by reformulating learning-based registration as an equilibrium-seeking problem. Our experiments show that this reformulation enables stable convergence, in contrast to existing unrolling methods that tend to degrade when further unrolled beyond the training steps. DEQReg achieves competitive accuracy and improved regularity with constant memory usage, effectively unifying classical optimization principles and learning-based advantages in speed and flexibility.

Acknowledgments. We gratefully acknowledge the Dutch Research Council (NWO) and Amazon Science for financial and computing support.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aberle, D.R., et al.: Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. **365**(5), 395–409 (2011)
2. Ashburner, J.: A fast diffeomorphic image registration algorithm. Neuroimage **38**(1), 95–113 (2007)
3. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage **54**(3), 2033–2044 (2011)
4. Bai, S., Geng, Z., Savani, Y., Kolter, J.Z.: Deep equilibrium optical flow estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 620–630 (2022)
5. Bai, S., Kolter, J.Z., Koltun, V.: Deep equilibrium models. Adv. Neural Inf. Process. Syst. **32** (2019)
6. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
7. Broyden, C.G.: A class of methods for solving nonlinear simultaneous equations. Math. Comput. **19**(92), 577–593 (1965)

8. De Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Med. Image Anal. **52**, 128–143 (2019)
9. Falta, F., Hansen, L., Heinrich, M.P.: Learning iterative optimisation for deformable image registration of lung CT with recurrent convolutional networks. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 301–309. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_29
10. Fung, S.W., Heaton, H., Li, Q., McKenzie, D., Osher, S., Yin, W.: JFB: Jacobian-free backpropagation for implicit networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36 (2022)
11. Geng, Z., Kolter, J.Z.: TorchDEQ: a library for deep equilibrium models. arXiv preprint arXiv:2310.18605 (2023)
12. Geng, Z., Zhang, X.Y., Bai, S., Wang, Y., Lin, Z.: On training implicit models. Adv. Neural. Inf. Process. Syst. **34**, 24247–24260 (2021)
13. Gilton, D., Ongie, G., Willett, R.: Deep equilibrium architectures for inverse problems in imaging. IEEE Trans. Comput. Imaging **7**, 1123–1133 (2021)
14. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Mach. Vis. Appl. **31**, 1–18 (2020)
15. Hering, A., Ginneken, B., Heldmann, S.: mlVIRNET: multilevel variational image registration network. In: Shen, D., et al. (eds.) MICCAI 2019, Part VI. LNCS, vol. 11769, pp. 257–265. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_29
16. Hering, A., et al.: Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Trans. Med. Imaging **42**(3), 697–712 (2022)
17. Hoopes, A., Hoffmann, M., Fischl, B., Guttag, J., Dalca, A.V.: HyperMorph: amortized hyperparameter learning for image registration. In: Feragen, A., Sommer, S., Schnabel, J., Nielsen, M. (eds.) IPMI 2021. LNCS, vol. 12729, pp. 3–17. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-78191-0_1
18. Hu, J., Gan, W., Sun, Z., An, H., Kamilov, U.S.: A plug-and-play image registration network. arXiv preprint arXiv:2310.04297 (2023)
19. Jena, R., Chaudhari, P., Gee, J.C.: Deep implicit optimization enables robust learnable features for deformable image registration. Med. Image Anal., 103577 (2025)
20. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)
21. Krantz, S.G., Parks, H.R.: The Implicit Function Theorem: History, Theory, and Applications. Springer, New York (2002). https://doi.org/10.1007/978-1-4614-5981-1
22. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. **19**(9), 1498–1507 (2007)
23. Mok, T.C.W., Chung, A.C.S.: Large deformation diffeomorphic image registration with Laplacian pyramid networks. In: Martel, A.L., et al. (eds.) MICCAI 2020, Part III. LNCS, vol. 12263, pp. 211–221. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_21

24. Qiu, H., Hammernik, K., Qin, C., Chen, C., Rueckert, D.: Embedding gradient-based optimization in image registration networks. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. Lecture Notes in Computer Science, vol. 13436, pp. 56–65. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_6
25. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015, Part III. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
26. Sherstinsky, A.: Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D **404**, 132306 (2020)
27. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: a survey. IEEE Trans. Med. Imaging **32**(7), 1153–1190 (2013)
28. Staring, M., vander Heide, U.A., Klein, S., Viergever, M.A., Pluim, J.P.: Registration of cervical MRI using multifeature mutual information. IEEE Trans. Med. Imaging **28**(9), 1412–1421 (2009)
29. Zhang, Y., Zhao, Y., Xue, H., Kellman, P., Klein, S., Tao, Q.: Recurrent inference machine for medical image registration. arXiv preprint arXiv:2406.13413 (2024)
30. Zhao, S., Dong, Y., Chang, E.I., Xu, Y.: Recursive cascaded networks for unsupervised medical image registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10600–10610 (2019)



Cascaded 3D Diffusion Models for Whole-Body 3D 18-F FDG PET/CT Synthesis from Demographics

Siyeop Yoon, Sifan Song, Pengfei Jin, Matthew Tivnan, Yujin Oh, Sekeun Kim, Dufan Wu, Xiang Li, and Quanzheng Li(✉)

Department of Radiology and Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
li.quanzheng@mgh.harvard.edu

Abstract. We propose a cascaded 3D diffusion model framework to synthesize high-fidelity 3D PET/CT volume directly from demographic variables, addressing the growing need for realistic digital twins in oncologic imaging, virtual trials, and AI-driven data augmentation. Unlike deterministic phantoms, which rely on predefined anatomical and metabolic templates, our method employs a two-stage generative process: an initial score-based diffusion model synthesizes low-resolution PET/CT volumes from the demographic variables only, providing global anatomical structures and approximate metabolic activity, followed by a super-resolution residual diffusion model refining spatial resolution. Our framework was trained on 18-F FDG PET/CT scans from the AutoPET dataset and evaluated using organ-wise volume and standardized uptake value (SUV) distributions, comparing synthetic and real data between demographic subgroups. The organ-wise comparison demonstrated strong concordance between synthetic and real images. In particular, most of the deviations in metabolic uptake values remained within 3–5% of the ground truth in sub-group analysis. These findings highlight the potential of cascaded 3D diffusion models to generate anatomically and metabolically accurate PET/CT images, offering a robust alternative to traditional phantoms and enabling scalable, population-informed synthetic imaging for clinical and research applications. Codes can be found at https://github.com/siyeopyoon/TotalGen.

Keywords: Data synthesis · Diffusion Models · PET · CT

1 Introduction

Medical image synthesis has become an essential tool in healthcare, facilitating data augmentation [1], modality translation [16,28], and digital twin simulations [11]. Generative models enable the creation of realistic medical images, addressing challenges related to data scarcity and privacy concerns [14,21]. While early

S. Yoon and S. Song—Equally Contributed.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 99–109, 2026.
https://doi.org/10.1007/978-3-032-04947-6_10

deep generative models such as VAEs [17] and GANs [6] demonstrated promise in medical image synthesis [9,22], diffusion models have recently emerged as the superior alternative, offering enhanced image fidelity and greater control over the generation process [8,12,26]. However, their high computational cost and slow inference speed remain significant challenges. To mitigate these computational demands, Latent diffusion models (LDMs) alleviate memory constraints by performing synthesis in a compressed latent space [24], though the final image quality is heavily dependent on the encoder-decoder network used for latent mapping [15]. In 3D volumetric image synthesis, patch-wise training has gained popularity [28,30]. By breaking a volume into smaller patches, models can operate more efficiently while maintaining high-resolution outputs [31]. Most previous studies on diffusion model-based image synthesis have focused on reconstructing images from undersampled measurements [4,10,30] or performing modality conversion using acquired images [20,27]. However, when models are weakly conditioned or operate unconditionally—without complementary imaging data, the iterative noise removal process often fails to maintain anatomical consistency, resulting in spatial incoherence [23].

The spatial contexts and alignment become even more challenging when models trained on different imaging modalities—such as anatomical and functional images—operate separately. Text-based image generation techniques have been explored as an alternative, using natural language descriptions to synthesize diverse medical images [7]. While such models can produce visually plausible results, they often lack anatomical precision due to the absence of explicit spatial constraints, leading to outputs that may not accurately reflect human anatomy.

To overcome these limitations, we propose an alternative framework that integrates stepwise 3D PET/CT volume synthesis with resolution enhancement, thereby overcoming challenges inherent in weakly conditional synthesis methods. Our approach first generates a coarse anatomical representation using only demographic attributes, establishing spatial relationships and organ layouts in a low-resolution 3D PET/CT volume. This blueprint is then progressively refined using a separated super-resolution residual diffusion model. By decoupling structural generation from visual enhancement, our method reduces dependence on large imaging datasets while streamlining the synthesis pipeline. The proposed framework is evaluated through task-based metrics, including organ volume accuracy and standardized uptake value (SUV) distributions from the AutoPET dataset. Our results confirm that the synthetic images closely align with real demographic-matched data, demonstrating high anatomical fidelity.

2 Methods

We propose a cascaded 3D diffusion model framework that employs a global-to-local synthesis strategy to generate anatomically and metabolically consistent PET/CT images only from demographics (Fig. 1).

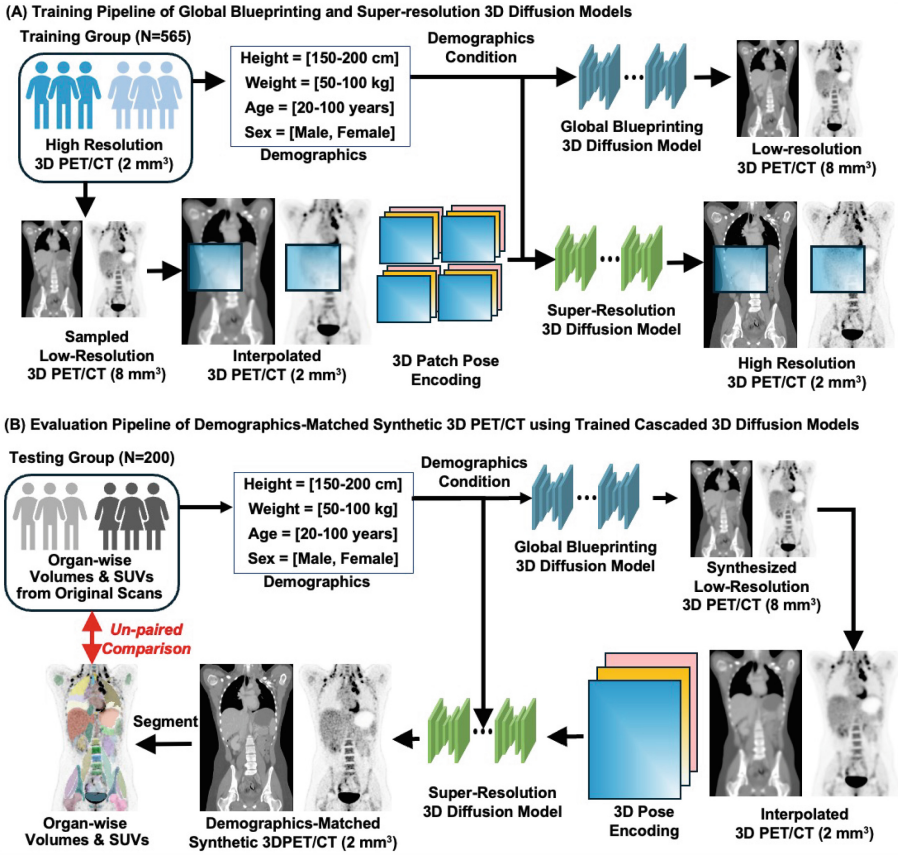


Fig. 1. Overview of the cascaded 3D diffusion framework for demographic-driven PET/CT synthesis. **(A)** During training, demographics guide the global diffusion model to generate a low-resolution 3D PET/CT, which is then interpolated and refined by a super-resolution diffusion model to produce high-resolution outputs. **(B)** For evaluation, the demographics from the testing set are matched and input to the same cascaded process. This yields synthetic PET/CT compared to real data cohorts, focusing on organ-wise volume and SUV metrics.

2.1 Cascaded Synthesis via Conditional Diffusion

Global Anatomical and Functional Synthesis. In the first stage of our cascaded framework, a globally coherent low-resolution PET/CT volume is generated through a score-based diffusion process conditioned on demographic attributes. To formalize this, we define a family of distributions $p(I_{\text{LR}}; \sigma, x_{\text{Demo}})$ obtained by perturbing a low-resolution image with i.i.d. Gaussian noise of standard deviation σ with demographics x_{Demo}. At the maximum noise level $\sigma_{\max}$, the distribution $p(I_{\text{LR}}; \sigma_{\max}, x_{\text{cond}})$ approximates pure Gaussian noise, which

serves as the initialization for the denoising procedure. The evolution from a noisy sample I_{LR} to a clean, structured, low-resolution volume is modeled by a stochastic differential equation (SDE) that combines deterministic drift with stochastic diffusion by [2]:

$$dI_{\mathrm{LR}} = -\frac{1}{2}\beta(t)\,\nabla_{I_{\mathrm{LR}}} \log p(I_{\mathrm{LR}}; \sigma(t), x_{\mathrm{Demo}})\,dt + \sqrt{\beta(t)}\,dW_t, \tag{1}$$

where $\beta(t)$ is a drift coefficient and dW_t represents a standard Wiener process. And $\nabla_{I_{\mathrm{LR}}} \log p(I_{\mathrm{LR}}; \sigma(t), x_{\mathrm{Demo}})$ denotes the score function that guides I_{LR} toward regions of higher probability density as the noise diminishes under the condition x_{Demo}. Numerical integration of the corresponding reverse-time ODE from Eq. (1)—yields a stable and deterministic trajectory from a highly perturbed state to the final low-resolution PET/CT image.

To approximate the score function $\nabla_{I_{\mathrm{LR}}} \log p(I_{\mathrm{LR}}; \sigma(t), x_{\mathrm{Demo}})$, we train a conditional score network $s_\theta(\cdot, \sigma, x_{\mathrm{Demo}})$ using the following loss [26]:

$$\mathcal{L}_{\text{score, LR}} = \mathbb{E}_{\sigma, I_{\mathrm{LR}}}\left[\lambda(\sigma)\left\|s_\theta(I_{\mathrm{LR}} + \sigma\epsilon, \sigma, x_{\mathrm{Demo}}) - I_{\mathrm{LR}}\right\|^2\right], \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\lambda(\sigma)$ is a weighting function that balances the contribution of different noise levels [12]. Multiplication σ to ϵ is to normalize the magnitude of the noise. This scaling allows the loss to remain stable across different noise levels and ensures that the network can generalize its denoising capability across the entire noise schedule. The generation of low resolution volume I_{LR} is then evolved by integrating the reverse-time ODE:

$$\frac{dI_{\mathrm{LR}}}{dt} = -\frac{1}{2}\beta(t)\,\nabla_{I_{\mathrm{LR}}} \log p(I_{\mathrm{LR}}; \sigma(t), x_{\mathrm{Demo}}), \tag{3}$$

from $t = T$ (high noise) down to $t = 0$ (no noise). This integration can be performed using standard numerical solvers (EDM2 Solver [12,13]) that provide a stable and accurate approximation of the trajectory. With the trained score network $s_\theta(\cdot, \sigma, x_{\mathrm{Demo}})$, the iterative denoising process reconstructs the complete anatomical and functional features, establishing a structural blueprint for the subsequent high-resolution synthesis stage.

Super-Resolution via Residual Diffusion and Patch-Wise Training. In the second stage, the low-resolution image I_{LR} obtained from the global synthesis is refined to recover fine anatomical details. First, an upsampled estimate I_{LU} is computed by linear interpolation to the target high-resolution dimensions. Since simple interpolation does not fully restore the textures and edges, we define a residual term, $R = I_{\mathrm{HR}} - I_{\mathrm{LU}}$, where I_{HR} is the true high-resolution image.

Given that the low-resolution volume serves as a structural blueprint, patch-wise training can be employed effectively. The model is conditioned on the noise level as well as on a spatial prior that specifies the 3D location of each patch within the interpolated PET/CT volume, I_{LU}. The 3D domain Ω is partitioned

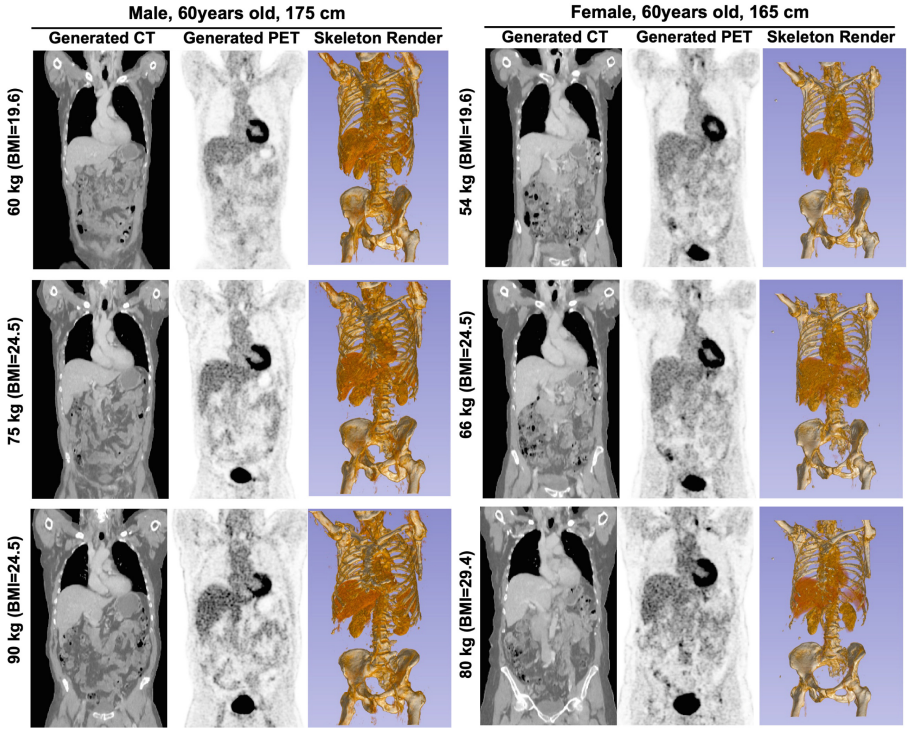


Fig. 2. Representative examples of 18-F FDG PET/CT generated using cascaded 3D diffusion models. The images show CT, 18-F FDG SUV, and 3D renderings for synthetic subjects of the same age (60 years) with different heights (male 175 cm vs. female 165 cm) and BMIs. The results demonstrate plausible anatomical and metabolic differences, including variations in adipose tissue distribution and PET signal heterogeneity.

into N patches, $\Omega = \bigcup_{i=1}^{N} \Omega_i$, and the patch-wise loss is formulated as

$$L_{Patch} = \sum_{i=1}^{N} \mathbb{E}_{R, I_{\mathrm{LU}}, \sigma} \left\| S_\theta \left(R_{\Omega_i} + \sigma\epsilon; I_{\mathrm{LU},\Omega_i}, \sigma, x_{\mathrm{Demo}}, \Omega_i \right) - R_{\Omega_i} \right\|_2^2, \quad (4)$$

where R_{Ω_i} denotes the residual signal corresponding to the positional encoding of patch Ω_i. I_{LU} and x_{Demo} represent the low-resolution PET/CT and demographics, respectively [28,31]. The inclusion of the spatial Ω_i and the low-resolution priors $I_{(\mathrm{LU},\Omega_i)}$ ensures that local reconstructions are consistent with the blueprint. Once the conditional score network S_θ is trained with the above patch-wise loss, it is used to estimate the residuals for high-resolution PET/CT volumes through the reverse-time ODE flow, as defined in Eq. (3). Finally, the super-resolved high-resolution PET/CT volume is obtained by adding the recovered residual R to the upsampled estimate $I_{\mathrm{SR}} = I_{\mathrm{LU}} + R$.

2.2 Dataset and Implementation Details

We utilize the AutoPET dataset [5], which comprises whole-body ^{18}F-FDG PET/CT scans. A total of 765 subjects with complete demographic information (age, sex, height, and weight) were retrospectively selected (565 training and 200 testing subjects). All volumes were resampled to a resolution of 2 mm^3.

To ensure consistent anatomical coverage, including key organs such as the heart, liver, and kidneys, CT volumes were segmented using the TotalSegmentator, and images were cropped to the region extending from the clavicle to the sacrum bones. Each volume was zero-padded to maintain a standardized spatial dimension: $I_{\text{HR}} \in \mathbb{R}^{224\times224\times384}$. The low-resolution volumes were derived from the high-resolution volumes using voxel subsampling. Specifically, each HR volume was sampled by a factor of 4 along each spatial dimension using stratified voxel extraction, defined as $I_{\text{LR}}^{(x,y,z)} = I_{\text{HR}}[x :: 4, y :: 4, z :: 4]$, with Cartesian multiplication of $(x, y, z) \in \{0, 1, 2, 3\}$. This process generates 64 unique low-resolution volumes per subject, significantly augmenting the training set size to low-resolution volumes while preserving anatomical diversity.

To standardize intensity values, CTs were clipped in the range $[-500, 500]$ HU and linearly scaled to $[0, 1]$. PET images were converted to SUV units, clipped to $[0, 25]$, and normalized using $SUV_{\log} = \frac{\log(SUV+1)}{\log(26)}$. During evaluation, normalized images were inverted to the original scale. Position encoding follows PatchDiffusion [28], tagging each patch with normalized (-1 to 1) x, y, z coordinates. Demographics (age, height, weight)/100, sex (0 = female, 1 = male) are concatenated as additional channels of input volume.

For the model implementation, we extend the EDM2 framework [12,13] to the 3D domain for conditional PET/CT synthesis by integrating 3D convolutions. Our modified 3D U-Net captures spatial dependencies across all axes, ensuring anatomical fidelity while incorporating demographic attributes to enhance realism. The training was performed on 16,252K images with an accumulated batch size of 2048, 64 base channels, and an initial learning rate of 0.017 decayed over 35,000 batches, requiring 72 h on an NVIDIA DGX A100 system (4×40 GB A100 GPUs for each model). A two-stage approach is employed: a global context model using a $56 \times 56 \times 96$ input and a super-resolution model refining patches of the same size cropped from high-resolution volumes. For seams among the patches, advanced artifact solutions can be utilized such as MultiDiffusion [3].

In the model testing, for both diffusion and 3D flow-matching models [18,19], the number of sampling steps was set to 35 for global synthesis and 100 for super-resolution. All models were evaluated on a single 40 GB A100 GPU. The global synthesis stage required a peak of 2 GB GPU memory and approximately 30 s per sample, whereas the super-resolution stage used up to 24 GB GPU memory and took about 6 min per sample. Note that super-resolution was performed on partial volumes of size $224 \times 224 \times 96$, which were subsequently concatenated along the z-axis to reconstruct the full high-resolution output. Codes can be found at https://github.com/siyeopyoon/TotalGen.

Performance evaluation of the proposed cascaded 3D diffusion framework was conducted through both task-oriented quantitative analyses. Organ volumes

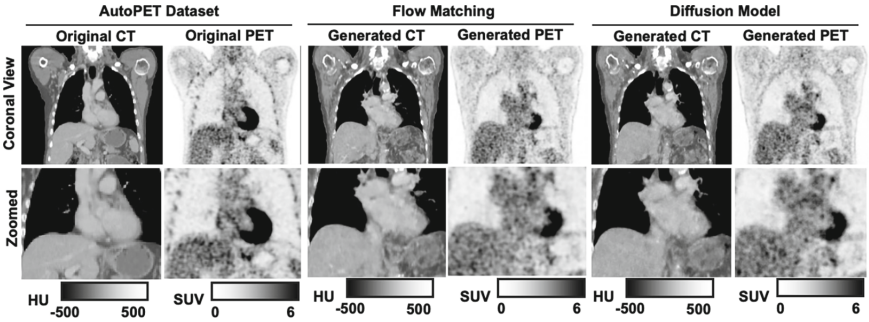


Fig. 3. Representative slices from the AutoPET (left) compared with synthetic CT/PET generated by the flow-matching (middle) and the diffusion model (right).

were measured using segmentation masks obtained from both original and generated CT images via TotalSegmentator [29], while organ SUV_{mean} and SUV_{max} were computed using the corresponding liver, heart, and kidney masks. Notably, failures or poor-quality segmentations were not observed in this study.

3 Results

Figure 2 shows representative CT, 18-F FDG PET, and the corresponding surface volume rendering of synthetic images. The synthetic PET/CT volumes exhibit visually realistic anatomical and metabolic features that are consistent with variations in demographic inputs. In particular, the generated images capture subtle morphological differences such as variations in adipose tissue distribution and heterogeneous PET signals across subjects.

Table 1 presents a quantitative comparison between the real (AutoPET) and synthetic datasets across multiple demographic groups. In general, synthetic PET/CT volumes showed strong concordance with real data at the group level. The mean SUV values in the synthetic PET images were within a close margin of the real data, and most subgroups did not show statistically significant differences ($p > 0.05$) in liver and kidney volumes. However, female groups exhibited significant offsets in heart volume (30 mL, 6.9% difference) due to segmentation variability at the boarder of blood pool and myocardium. Similarly, a significant underestimation of heart SUV_{max} was observed in the male group.

In our experiments, the flow-matching model produced synthetic PET/CT images with lower variability in quantitative metrics compared to the AutoPET data, as evidenced by a reduced standard deviation in organ volumes and SUV measurements. Flow matching using ODE formulations tends to yield less diverse samples compared to stochastic diffusion models, mainly because the deterministic integration in ODE approaches can constrain variability [25]. In contrast, diffusion models that rely on stochastic sampling can better capture and preserve the natural variability of the data, while the resolution enhancement stage

Table 1. Results of Organ-wise Volume and Standard Uptake Values (mean, peak) measured in the AutoPET and synthetic datasets of 18-F FDG PET/CT. * indicates p-value <0.05 compared to AutoPET. Data are Mean ± Std (Mean difference %).

Test cohort demographics						
	Male (N = 108)			Female (N = 92)		
Age (years)	58 ± 17			59 ± 15		
Heights (cm)	177 ± 7			165 ± 7		
Weights (kg)	83 ± 15			76 ± 19		
Measurement in Liver						
Method	AutoPET	Flow	Diffusion	AutoPET	Flow	Diffusion
Volume(L)	1.76 ± 0.44	1.78 ± 0.31	1.78 ± 0.36	1.55 ± 0.37	$1.67 \pm 0.35^*$	1.62 ± 0.33
	–	(1.1%)	(1.3%)	–	(7.6%)	(4.7%)
SUV_{mean}	2.26 ± 0.42	$1.96 \pm 0.26^*$	2.18 ± 0.34	2.34 ± 0.33	$2.15 \pm 0.22^*$	2.38 ± 0.40
	–	(−13.4%)	(−3.9%)	–	(−8.6%)	(1.5%)
SUV_{max}	7.29 ± 4.28	$5.27 \pm 1.83^*$	6.58 ± 3.02	7.03 ± 4.57	$5.41 \pm 1.67^*$	7.22 ± 3.76
	–	(−27.7%)	(−9.7%)	–	(−23.1%)	(2.8%)
Measurement in Heart						
Volume(L)	0.68 ± 0.15	$0.74 \pm 0.10^*$	0.71 ± 0.11	0.55 ± 0.09	$0.63 \pm 0.09^*$	$0.58 \pm 0.09^*$
	–	(8.3%)	(3.8%)	–	(15.8%)	(6.9%)
SUV_{mean}	2.48 ± 1.03	2.68 ± 0.52	2.85 ± 1.17	3.07 ± 1.39	3.07 ± 0.51	3.04 ± 1.17
	–	(7.8%)	(14.5%)	–	(0.2%)	(−0.8%)
SUV_{max}	10.20 ± 5.92	$13.72 \pm 4.71^*$	$13.96 \pm 7.25^*$	12.20 ± 7.46	$15.26 \pm 4.59^*$	13.69 ± 7.51
	–	(34.5%)	(36.8%)	–	(25.%)	(12.2%)
Measurement in Kidneys						
Volume(L)	0.30 ± 0.08	$0.33 \pm 0.05^*$	0.35 ± 0.07	0.30 ± 0.06	0.30 ± 0.05	0.30 ± 0.06
	–	(−6.9%)	(−2.3%)	–	(3.0%)	(3.0%)
SUV_{mean}	2.50 ± 0.41	$2.20 \pm 0.29^*$	2.47 ± 0.42	2.70 ± 0.45	$2.48 \pm 0.17^*$	2.71 ± 0.38
	–	(−12.0%)	(−0.9%)	–	(−8.1%)	(0.4%)
SUV_{max}	11.20 ± 5.24	$8.90 \pm 2.77^*$	10.63 ± 4.04	11.83 ± 5.55	11.39 ± 3.58	11.79 ± 3.63
	–	(−19.7%)	(−5.0%)	–	(−3.8%)	(−0.4%)

yielded outputs with similar quality (Fig. 3). Overall, shape metrics across demographic groups did not differ significantly, confirming that the synthetic images preserve anatomical structure while reflecting realistic metabolic activity.

4 Conclusion

In this work, we introduced a cascaded 3D diffusion framework for the synthesis of PET/CT images directly from demographic variables. Our approach leverages a two-stage process, where an initial conditional diffusion model generates a coarse, low-resolution anatomical framework, and a subsequent super-resolution diffusion model refines this output to recover fine metabolic and

structural details. This division of the synthesis task allows our method to effectively capture global anatomical structures while also ensuring that local, high-frequency details are faithfully reproduced. Quantitative evaluation on the AutoPET dataset demonstrates that the synthetic images closely replicate key clinical metrics, including organ volumes and standardized uptake values (SUV), across a range of demographic groups. The strong agreement observed in liver, kidney, and heart measurements indicates that the framework is capable of producing anatomically accurate and metabolically realistic images. The results of our study highlight the potential of this framework to serve as a reliable and scalable alternative to traditional imaging phantoms. By reducing the dependency on large, annotated imaging datasets, our method provides a novel solution for data augmentation, digital twin simulations, and virtual clinical trials. Future work will refine the synthesis pipeline to address residual artifacts in specific cohorts and evaluate its generalizability to other imaging modalities. We will also adapt the framework for multi-sequence MRI datasets and develop a text-conditioned pathology generation module.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akrout, M., et al.: Diffusion-based data augmentation for skin disease classification: impact across original medical datasets to fully synthetic images. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (eds.) MICCAI 2023. LNCS, vol. 14533, pp. 99–109. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-53767-7_10
2. Anderson, B.D.: Reverse-time diffusion equation models. Stoch. Proc. Appl. **12**(3), 313–326 (1982)
3. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: fusing diffusion paths for controlled image generation. Proc. Mach. Learn. Res. **202**, 1737–1752 (2023)
4. Chung, H., Ye, J.C.: Score-based diffusion models for accelerated MRI. Med. Image Anal. **80**, 102479 (2022)
5. Gatidis, S., et al.: A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. Sci. Data **9**(1), 601 (2022)
6. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
7. Hamamci, I.E., et al.: GenerateCT: text-conditional generation of 3D chest CT volumes. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV 2024. LNCS, vol. 15137, pp. 126–143. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72986-7_8
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
9. Hu, Q., Li, H., Zhang, J.: Domain-adaptive 3D medical image synthesis: an efficient unsupervised approach. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 495–504. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_47

10. Jiang, C., et al.: PET-Diffusion: unsupervised PET enhancement based on the latent diffusion model. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14220, pp. 3–12. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43907-0_1
11. Kadry, K., Gupta, S., Nezami, F.R., Edelman, E.R.: Probing the limits and capabilities of diffusion models for the anatomic editing of digital twins. npj Digit. Med. **7**(1), 1–12 (2024)
12. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Adv. Neural. Inf. Process. Syst. **35**, 26565–26577 (2022)
13. Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., Laine, S.: Analyzing and improving the training dynamics of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24174–24184 (2024)
14. Kazeminia, S., et al.: GANs for medical image analysis. Artif. Intell. Med. **109**, 101938 (2020)
15. Khader, F., et al.: Denoising diffusion probabilistic models for 3D medical image generation. Sci. Rep. **13**(1), 7303 (2023)
16. Kim, J., Park, H.: Adaptive latent diffusion model for 3D medical image to image translation: multi-modal magnetic resonance imaging study. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 7604–7613 (2024)
17. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)
18. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
19. Lipman, Y., et al.: Flow matching guide and code. arXiv preprint arXiv:2412.06264 (2024)
20. Pan, S., et al.: Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model. Med. Phys. **51**(4), 2538–2548 (2024)
21. Pianykh, O.S., et al.: Continuous learning AI in radiology: implementation principles and early applications. Radiology **297**(1), 6–14 (2020)
22. Pinaya, W.H., et al.: Generative AI for medical imaging: extending the MONAI framework. arXiv preprint arXiv:2307.15208 (2023)
23. Pinaya, W.H., et al.: Brain imaging generation with latent diffusion models. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (eds.) DGM4MICCAI 2022. LNCS, vol. 13609, pp. 117–126. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-18576-2_12
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
25. Schusterbauer, J., et al.: FMBoost: boosting latent diffusion with flow matching. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV 2024. LNCS, vol. 15119, pp. 338–355. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73030-6_19
26. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
27. Wang, T., Yang, X.: Take CT, get pet free: AI-powered breakthrough in lung cancer diagnosis and prognosis. Cell Rep. Med. **5**(4) (2024)
28. Wang, Z., et al.: Patch diffusion: faster and more data-efficient training of diffusion models. arXiv preprint arXiv:2304.12526 (2023)

29. Wasserthal, J., et al.: TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol. Artif. Intell. **5**(5), e230024 (2023)
30. Yoon, S., et al.: High-resolution 3D CT synthesis from bidirectional x-ray images using 3d diffusion model. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–4. IEEE (2024)
31. Yoon, S., et al.: Volumetric conditional score-based residual diffusion model for PET/MR denoising. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15007, pp. 754–763. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72104-5_72



Causality-Driven Spatio-Temporal Generator for Multi-phase Contrast-Enhanced CT Synthesis

Qikui Zhu[1(✉)], Hao Wu[1], Yanyan Zhang[2], and Shuo Li[2]

[1] Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, USA
qikuizhu@163.com

[2] Department of Computer Science, Case Western Reserve University, Cleveland, OH, USA
shuo.li11@case.edu

Abstract. Synthesizing multi-phase contrast-enhanced CT (CE-CT) images is clinically significant, as it can mitigate clinical risks such as radiation exposure and allergic reactions to contrast agents. However, existing methods treat multi-phase synthesis as separate tasks, failing to maintain the inter-phase dependencies and consistency between synthesized multi-phase CE-CT images. Moreover, the limited variability in CT intensity distributions makes it challenging to capture subtle variations in multi-phase imaging. For the first time, we propose a novel Causality-driven Spatio-temporal Generator (CSGen) for synthesizing multi-phase CE-CT imaging through three key novelties: 1) Using a novel phase-causality to creatively exploit the multi-phase variation content for driving the multi-phase CE-CT synthesizing, addressing the challenge of capturing multi-phase discriminative features through one model. 2) Introducing a new Spatio-temporal Transformer to establish the spatio-temporal correlation between multi-phase CE-CT images for leveraging multi-phase inter- and intra-dependencies and improving synthesis quality. 3) Multi-phase adversarial learning is designed for enhancing multi-phase discriminative feature learning. Experimental results (mean PSNR: 31.15, mean SSIM: 0.9066, mean NMAE: 3.17) demonstrate that CSGen outperforms state-of-the-art synthesis methods, and, for the first time, successfully synthesizes multi-phase CE-CT images.

Keywords: Multi-phase CE-CT synthesis · Causality-driven · Spatio-Temporal Transformer · Multi-phase adversarial learning

1 Introduction

Multi-phase contrast-enhanced CT (CE-CT) imaging is essential for diagnosing liver tumors [1,2]. However, in clinical practice, multi-phase CE-CT imaging is performed by injecting chemical contrast agents (CAs), which introduces several

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 110–119, 2026.
https://doi.org/10.1007/978-3-032-04947-6_11

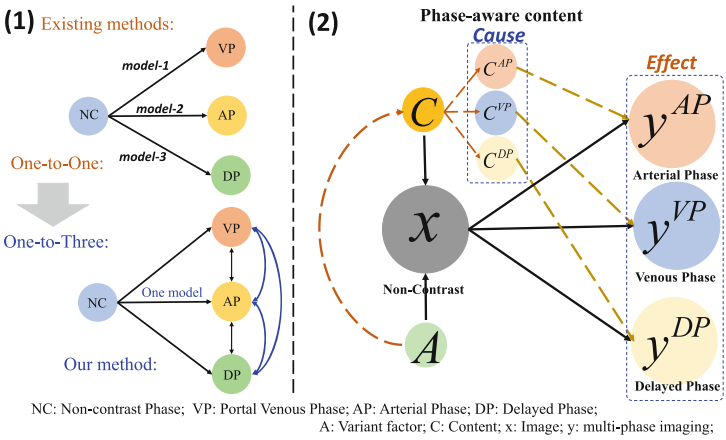


Fig. 1. 1) Existing methods treat multi-phase synthesis as separate tasks, failing to ensure consistency across the synthesized multi-phase images. 2) We use a phase-causality to exploit the multi-phase variation content and formulate multi-phase CE-CT image synthesis as a sequence synthesis task.

inefficiencies and risks, such as prolonged examination time, radiation exposure [7,16], and the potential for allergic reactions [9]. If multi-phase CE-CT imaging can be synthesized using non-contrast CT (NCCT) imaging without the administration of chemical CAs, it has the potential to reduce costs and eliminate associated dangers. This would leverage the full capabilities of CT imaging, leading to a significant impact on clinical applications and providing greater value for patients [17].

Although existing state-of-the-art methods [10,12,15,18,19] achieved remarkable performance in image synthesis, these methods pay all attention to single modality and can not exploit the correlation between multi-phase images to maintain the inter-phase dependencies and consistency (Fig. 1.1). Furthermore, these methods struggle to address the unique challenges associated with synthesizing multi-phase images. Specifically, 1) simultaneously learning multi-phase variation information from one input for these methods is challenging. The contents of CT imaging enhanced by CAs in different phases are different, coronary arteries, aorta in portal venous phase; liver, pancreas, kidney in the arterial phase; and kidney tumor in the delayed phase, simultaneously learning and classifying these contents is challenging. 2) Different from dense sequences (e.g.video), while multi-phase CE-CT imaging retains a consistent overall structure across phases, the content differs considerably, posing a significant challenge to capturing inter-phase dependencies [3]. What's more, the inherent intricacies of CT images also amplify the challenge [8]. Specifically, 1) CT imaging exhibits low sensitivity to CAs, making it challenging to obtain an accurate feature representation of CAs [5]. 2) CT imaging exhibits low contrast in soft-tissue structures, which makes extracting discriminative tissue representations for different phases challenging. These challenges have resulted in the synthesis of multi-phase contrast-enhanced CE-CT images remaining largely unexplored.

We propose a novel Causality-driven Spatio-temporal Generator (CSGen) with three key novelties for synthesizing multi-phase CE-CT images. Specifically, *1):* using a novel phase-causality to exploit the multi-phase variation content for driving the multi-phase CE-CT synthesizing (Fig. 1.2), addressing the challenge of capturing multi-phase discriminative features through one model. Unlike previous approaches that directly intervene from inputs to outputs, our phase-causality builds a causal graph to represent the dependency among multi-phase images, variant factors (e.g. CAs), and semantic contents and conducts counterfactual inference over the causal graph to exploit the intrinsic causality of the discrepancy between the multi-phase images. More specifically, it can drive generator in exploiting the multi-phase variation content for learning multi-phase CA-responding discriminative content from each phase. *2):* A new cross-phase Spatio-temporal Transformer (StTransformer) is proposed, establishing the spatio-temporal correlation between multi-phase CE-CT images. Specifically, StTransformer leverages the dependencies within multi-phase CE-CT images from both spatial and temporal perspectives, which empowers the model to harness content and structural correlations across multi-phase CE-CT imaging. Additionally, *3):* a multi-phase adversarial learning (MpAL) is proposed. Rather than focusing solely on the differences in feature distributions between synthetic and real images, MpAL leverages multi-phase information to enhance the discriminator's ability to distinguish among multiple phases, further encouraging the generator to learn discriminative features across all phases. Experimental results on three-phase CE-CT imaging synthesis show that CSGen significantly improves the quality of synthetic multi-phase CE-CT images and outperforms the state-of-the-art methods.

Our main contributions include:

- For the first time, multi-phase CE-CT imaging synthesis through a single model has been achieved. A single, phase-interaction methodology that eliminates the need for chemical CAs in synthesizing multi-phase CE-CT images.
- CSGen, for the first time, converts multi-phase CE-CT imaging synthesis to sequence synthesis, establishing the spatio-temporal correlation between multi-phase CE-CT images, addressing the limitation in exploiting correlation of multi-phase images.
- Phase-causality, for the first time, is used in synthesizing multi-phase CE-CT imaging. Advanced than previous approaches that directly intervene from inputs to outputs, it focuses on the causal inference between cross-phase feature representation space, allowing for more flexible and adaptive interventions.
- StTransformer has the advantage of establishing cross-phase spatio-temporal correlation, which addresses the challenge of Transformer in exploiting multi-phase spatial and temporal dependencies.

2 Method

Our Causality-driven Spatio-temporal Generator (CSGen) integrates causal inference with multi-phase adversarial learning, formulating multi-phase CE-

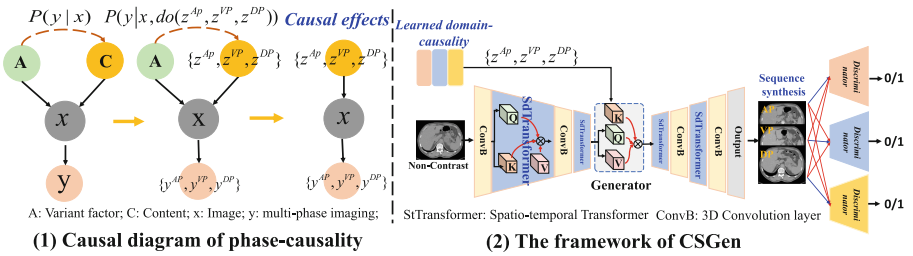


Fig. 2. 1) The causal diagram of our phase-causality. 2) The framework of CSGen.

CT image synthesis as a sequence synthesis task (Fig. 2.2). Specifically, given a NCCT image, CSGen employs a 3D generator—composed of stacked convolutional blocks and a Spatio-temporal Transformer—guided by learned phase causality to generate the multi-phase CE-CT sequence, including Portal Venous Phase (VP), Arterial Phase (AP), and Delayed Phase (DP).

2.1 Phase-Causality for Learning Phase-Specific Content

According to [11], the causal relationship between NCCT and CE-CT images can be defined as follows: the multi-phase CE-CT images $\{y^{VP}, y^{AP}, y^{DP}\} \in R^{C\times H\times W}$ are generated based on two independent factors: contrast agent: $\boldsymbol{A}$ and content: $\boldsymbol{C}$ (Fig. 2.1). Here, $\boldsymbol{A}$ represents the injected CAs, while $\boldsymbol{C}$ represents the tissue-specific response observable in NCCT. Factor $\boldsymbol{A}$ modulates the pixel intensities of various tissues over different time intervals, thereby yielding multi-phase CE-CT images. Building on this causal inference, we implement our phase-causality learning by extracting the phase-specific content latent code $\{z^{AP}, z^{VP}, z^{DP}\}$. Inspired by latent autoencoders [4,14], we employ an adversarial latent autoencoder to encode the phase-specific content latent. The adversarial latent autoencoder consists an encoder (E), a decoder (G), a discriminator (D), they learn to represent images with codes from a learned, discrete codebook $\mathcal{Z} = \{z^{AP}, z^{VP}, z^{DP}\}$. More precisely, we approximate a given image y^i by $\hat{x}^i = G(z^i)$. We obtain z^i using the encoder $\hat{z}^i = E(y^i)$ and a subsequent element-wise quantization $q(.)$ or each spatial code $\hat{z}^i$ onto its closet codebook z^i:

$$z^i = q(\hat{z}^i) := (\arg\min \left\|\hat{z}^i - z^i\right\|) \in R^{n\times h\times w} \tag{1}$$

To enable the codebook to learn discriminative feature representations from each phase, a discriminator oversees the learning process. This end-to-end codebook training is guided by the following loss function:

$$\mathcal{L}_{VQ}(E, G, \mathcal{Z}) = \|y^i - \hat{y^i}\|^2 + \|\text{sg}[E(y^i)] - z^i\|_2^2 \tag{2}$$

$$Loss = \arg\min_{E,G,\mathcal{Z}} \max_{D} \mathbb{E}_{x\sim q(x)} \left[\mathcal{L}_{VQ}(E, G, \mathcal{Z}) + \mathcal{L}_{GAN}(\{E, G, \mathcal{Z}\}, D)\right], \tag{3}$$

where, L_{GAN} is adversarial loss, $\|\text{sg}[E(y^i)] - z^i\|_2^2$ denotes commitment loss. To improve the discrimination ability, three independent discriminators are used for

three-phase CE-CT imaging evaluation. After model training, we can obtain the phase-specific content latent.

2.2 Spatio-temporal Transformer Establishes Cross-Phase Correlation

To exploit spatio-temporal correlations among multi-phase images, the Spatio-temporal Transformer (StTransformer) employs two key modules: (1) Cross-phase Attention Module (CAM) and (2) Inter-phase Attention Module (IAM). CAM explores both content and structural correlations across multi-phase CE-CT images. Meanwhile, IAM establishes spatial dependencies within each phase by utilizing inter-phase variance information, which captures the distribution of CAs and extracts distinctive feature representations from multi-phase CE-CT images. Given a sequence feature map $\mathrm{X} \in \mathbb{R}^{C\times D\times H\times W}$, where C is the number of channels, D, H, and W represent the spatial dimension. The CAM projects X into sequences $[Q_D, K_D, V_D] \in R^{(CHW)\times D}$, and performs attention calculation from the temporal aspect. Afterward, IAM projects Y_D into new sequences $[X_Q; X_K; X_V] \in R^{\mathrm{D}\times(CHW)}$, and computing the spatial dependencies within each phase:

$$Y_D = \mathrm{softmax}\left(\frac{Q_D K_D^T}{d_D}\right) V_D, \quad X = \mathrm{softmax}\left(\frac{X_Q X_K^T}{d}\right) X_V. \tag{4}$$

where d is a learnable scaling parameter. The two modules address long-range dependencies in both spatial and temporal dimensions, facilitating the exploration of inter- and intra-phase relationships within multi-phase images.

2.3 Multi-phase Adversarial Learning for Discriminative Training

To endow the generator with greater discriminative power, CSGen uses multi-phase information for adversarial learning. Specifically, all of target multi-phase CE-CT images that consist of cross-phase feature distribution are used as negative samples for training. The training loss for multi-phase adversarial learning (MpAL) is:

$$L_{MpAL} = \sum_{\tilde{y}^i, y^j} L_{ce}(y^j, 0)\mathbb{1}[\tilde{y}^i \neq y^j] + L_{ce}(\tilde{y}^i, 0) + L_{ce}(y^i, 1) \tag{5}$$

where $\mathbb{1}$ is the binary indicator denoting whether the pixel belongs to same class, L_{ce} is cross entropy loss function, $\tilde{y}_i$, y_i represents the pixel in synthetic multi-phase CE-CT images $\{\tilde{y}^{VP}, \tilde{y}^{AP}, \tilde{y}^{DP}\}$ and CE-CT images $\{y^{VP}, y^{AP}, y^{DP}\}$, respectively.

2.4 Model Training

In summary, the pre-learned phase-specific content latent from phase-causality is initially injected into the generator through self-attention (Fig. 2.2). Under the supervision of MpAL, the generator is supervised by

$$Loss = \alpha L_{pixel-wise} + \beta L_{perceptual} + \lambda L_{MpAL} \tag{6}$$

Table 1. Quantitative results of multi-phase CE-CT imaging synthesis evaluated from holistic (top) and tumor (bottom) perspectives.

Method	Mapping	Holistic evaluation											
		PSNR ↑				SSIM ↑				NMAE ↓			
		AP	VP	DP	Mean	AP	VP	DP	Mean	AP	VP	DP	Mean
CycleGAN [15]	One-to-One	29.39	28.93	28.63	28.98	0.8833	0.8802	0.8788	0.8808	3.68	3.92	3.81	3.80
AGAN [6]	One-to-One	29.05	29.16	29.11	29.11	0.8731	0.8666	0.8731	0.8709	4.01	3.92	3.91	3.95
CyTran [13]	One-to-One	29.25	29.36	28.92	29.18	0.8345	0.8718	0.8624	0.8562	4.39	3.97	4.17	4.18
CSGen (our)	One-to-Three	**31.24**	**31.34**	**30.87**	**31.15**	**0.9074**	**0.9081**	**0.9044**	**0.9066**	**3.14**	**3.13**	**3.25**	**3.17**
Method	Mapping	Tumor evaluation											
		AP	VP	DP	Mean	AP	VP	DP	Mean	AP	VP	DP	Mean
CycleGAN [15]	One-to-One	29.46	29.24	28.65	29.12	0.8884	0.8867	0.8844	0.8865	3.58	3.74	3.71	3.68
AGAN [6]	One-to-One	29.14	29.27	29.22	29.21	0.8787	0.8726	0.8792	0.8768	3.95	3.82	3.80	3.86
CyTran [13]	One-to-One	29.33	29.43	29.01	29.26	0.8377	0.8772	0.8682	0.8610	4.32	3.86	4.06	4.08
CSGen (our)	One-to-Three	**31.27**	**31.43**	**30.91**	**31.20**	**0.9125**	**0.9133**	**0.9101**	**0.9120**	**3.02**	**3.00**	**3.12**	**3.05**

$$L_{perceptual}(\tilde{y}^i, y^i) = \left\| VGG(\tilde{y}^i) - VGG(y^i) \right\|_1 \tag{7}$$

where, α, β and λ are weight coefficients. The $L_{pixel-wise}$ computes the difference between synthetic multi-phase CE-CT images $\{\tilde{y}^{VP}, \tilde{y}^{AP}, \tilde{y}^{DP}\}$ and target multi-phase CE-CT images $\{y^{VP}, y^{AP}, y^{DP}\}$ at the pixel level, where L_1 loss is used. The perceptual loss $L_{perceptual}$ is utilized to compute the difference between synthetic and true multi-phase CE-CT images in higher feature representations extracted from pre-trained VGG network.

$$L_{perceptual}(\tilde{y}^i, y^i) = \left\| VGG(\tilde{y}^i) - VGG(y^i) \right\|_1 \tag{8}$$

3 Experiments

3.1 Dataset and Implementation

A total of 92 well-registered multi-phase CT volumes are used for evaluation. Each volume has four phases including a non-contrast phase and three CA-enhanced phases, VP, AP, and DP. During training, the voxel intensity was normalized within each subject to a scale of $[-1, 1]$ via division by the maximum intensity. The training, validation, and test sets were randomly split in a 7:1:2 ratio. Our model is implemented using PyTorch and trained end-to-end with Adam optimization. During training, the learning rate is initially set to 0.0001 and decreased by a weight decay of 1.0×10^{-6}. The weight coefficients α, β, λ are set to 1.0, 1.0, 0.5 respectively. The performance is evaluated from **1) Holistic aspect**: evaluating the synthesis quality from a global view; and **2) Local aspect**: evaluating the sensitivity to tumor by using tumor labels. The peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) metrics, and normalized mean absolute error (NMAE) between the synthesized and true CE-CT images are measured.

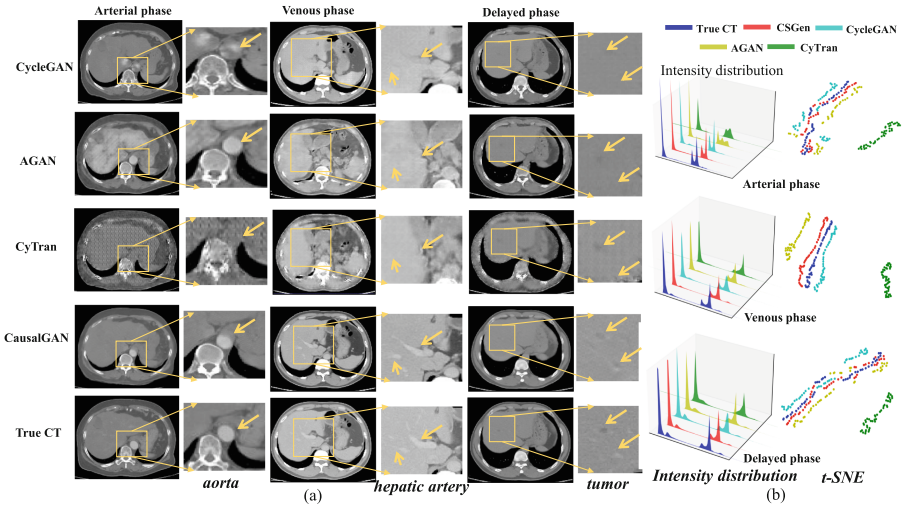


Fig. 3. Experimental results demonstrate that our model can more accurately enhance hepatic artery, aorta, and liver tumors than existing methods. b) The intensity distribution and t-SNE visualization of synthesis CE-CT images.

3.2 Comparison with State-of-the-Art Methods

Table 1 shows the quantitative result of CSGen and state-of-the-art methods from both holistic and liver tumor perspectives. In all three phases of CE-CT imaging, our CSGen exhibited superior performance in terms of PSNR, SSIM, and NMAE compared to other methods. Additionally, from the quantitative result of synthetic tumors, CSGen also achieved the best quantitative result. Notably, all compared methods employ a one-to-one mapping approach. These experimental results demonstrate that our CSGen has higher sensitivity to CA-aware feature distribution and are able to successfully synthesize multi-phase CE-CT images. Figure 3(a) clearly shows that our CSGen consistently enhances these CA-related organs including hepatic artery, aorta, and tumors across all phases, which provides compelling evidence for the effectiveness of our model in the synthesis of multi-phase CE-CT images. The feature distribution (Fig. 3(b)) of synthesized CE-CT images reveals that the intensity and feature distribution (features extracted by VGG) of CE-CT images synthesized by our CSGen is notably closer to true CE-CT images compared to other methods, which highlights the effectiveness of our approach in leveraging cross-domain discriminative features and proves that existing methods have low sensitivity in the CAs responding areas.

3.3 Effect Analysis of CSGen

Table 2 lists the quantitative results of various modules. StTransformer assists the baseline to obtain 1.96, 2.18, and 1.08 holistic improvements and 1.91, 1.08,

Table 2. Quantitative results of synthesized CE-CT imaging.

Method				Holistic evaluation											
				PSNR↑				SSIM (%)↑				NMAE↓			
Baseline	StTransformer	PC	MpAL	AP	VP	DP	Mean	AP	VP	DP	Mean	AP	VP	DP	Mean
✓				28.39 ± 3.00	28.36 ± 2.88	27.88 ± 2.67	28.21	87.98 ± 5.86	88.04 ± 5.74	86.87 ± 5.54	87.63	4.24 ± 1.43	4.25 ± 1.44	4.43 ± 1.46	4.31
✓	✓			30.31 ± 4.37	30.27 ± 4.41	29.94 ± 4.50	30.17	89.08 ± 5.62	90.32 ± 5.81	90.02 ± 5.84	89.81	3.25 ± 1.56	3.18 ± 1.70	3.27 ± 1.72	3.23
✓		✓		30.44 ± 4.26	30.44 ± 4.28	29.97 ± 4.29	30.28	87.43 ± 5.80	90.35 ± 5.84	89.83 ± 5.78	89.20	3.30 ± 1.55	3.11 ± 1.67	3.27 ± 1.67	3.23
✓			✓	28.76 ± 2.99	28.74 ± 3.12	28.34 ± 2.94	28.61	86.27 ± 2.99	89.34 ± 5.12	88.58 ± 4.91	88.06	3.72 ± 1.33	3.62 ± 1.37	3.76 ± 1.35	3.70
✓	✓	✓	✓	31.24 ± 5.23	31.34 ± 5.37	30.87 ± 5.32	31.15	90.74 ± 6.06	90.81 ± 6.22	90.44 ± 6.29	90.66	3.14 ± 1.69	3.13 ± 1.83	3.25 ± 1.88	3.17
				Tumor evaluation											
Baseline	StTransformer	PC	MpAL	AP	VP	DP	Mean	AP	VP	DP	Mean	AP	VP	DP	Mean
✓				28.44 ± 2.90	28.45 ± 2.77	27.97 ± 2.60	28.29	88.54 ± 5.39	88.62 ± 5.30	87.51 ± 5.02	88.23	4.18 ± 1.36	4.18 ± 1.31	4.33 ± 1.35	4.23
✓	✓			30.29 ± 4.17	30.29 ± 4.27	30.01 ± 4.38	30.20	89.56 ± 4.88	90.78 ± 5.14	90.59 ± 5.07	89.31	3.15 ± 1.41	3.09 ± 1.53	3.15 ± 1.56	3.13
✓		✓		30.41 ± 3.99	30.50 ± 4.08	30.04 ± 4.18	30.32	87.97 ± 4.96	90.89 ± 5.12	90.42 ± 4.99	89.76	3.19 ± 1.38	2.98 ± 1.48	3.15 ± 1.49	3.11
✓			✓	28.97 ± 2.81	28.95 ± 2.90	28.51 ± 2.78	28.81	87.05 ± 4.23	90.11 ± 4.31	89.37 ± 4.06	88.84	3.54 ± 1.17	3.44 ± 1.18	3.59 ± 1.20	3.52
✓	✓	✓	✓	31.27 ± 5.04	31.43 ± 5.22	30.91 ± 5.15	31.20	91.25 ± 5.39	91.33 ± 5.59	91.01 ± 5.50	91.20	3.02 ± 1.55	3.00 ± 1.64	3.12 ± 1.69	3.05

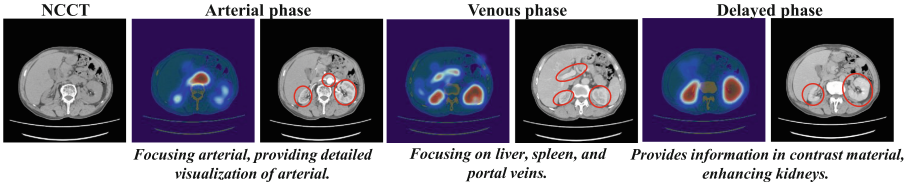


Fig. 4. The heatmap generated by phase-causality on three phases.

and 1.10 local improvements on mean PSNR, mean SSIM, and mean NMAE, respectively. These improvements reveal the advantage of StTransformer in utilizing the spatio-temporal correlation of multi-phase CE-CT images. The phase-causality (PC) assists the baseline to obtain a 2.07, 1.57, and 1.08 holistic improvement and obtain a 2.03, 1.53, and 1.12 local improvement on mean PSNR, mean SSIM, and mean NMAE PSNR, respectively. Those improvements demonstrate that phase-causality enhanced the generator's ability to learn CA-responding content information. Furthermore, the learned phase-causality contents are illustrated in Fig. 4, which reveals its aligns with discriminative features specific to each phase. In AP phase, the arterial vasculature is captured and enhanced, revealing intricate details of arterial structures and early enhancement within organs. In VP phase, features of the portal venous system are emphasized. In DP phase, the excretory characteristics of the contrast material become more apparent, particularly enhancing kidney visualization. These experimental results underscore the excellent ability of our phase-causality in learning CA-responding content from each CE-CT image. Additionally, MpAL assists the model to obtain a 0.40, 0.43, and 0.61 holistic improvement, and 0.52, 0.61 and 0.71 local tumor improvement on mean PSNR, mean SSIM, and mean NMAE, respectively. These improvements also prove the effectiveness of MpAL in capturing global and local discriminative contextual information.

4 Conclusion

Our CSGen is the first to successfully synthesize three-phase CE-CT imaging from NCCT using a single model. CSGen uses three novel mechanisms: 1) phase-causality mechanism; 2) spatio-temporal Transformer; 3) multi-phase adversarial learning to address the challenge of three-phase CE-CT synthesis. Experiments prove that the synthesized three-phase CE-CT imaging by CSGen is equivalent to real CE-CT imaging and outperforms state-of-the-art synthesis methods.

Acknowledgments. The National Institutes of Health supported this work under grants R01-HL167199, R01-HL165218, R01-HL165218-01A1W1, and R44-HL156811. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdelhalim, I., Abou El-Ghar, M., Dwyer, A., Ouseph, R., Contractor, S., El-Baz, A.: A new non-invasive AI-based diagnostic system for automated diagnosis of acute renal rejection in kidney transplantation: analysis of ADC maps extracted from matched 3D iso-regions of the transplanted kidney. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15012, pp. 390–398. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_37
2. Bae, H., et al.: Radiomics analysis of contrast-enhanced CT for classification of hepatic focal lesions in colorectal cancer patients: its limitations compared to radiologists. Eur. Radiol. **31**(11), 8786–8796 (2021)
3. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12873–12883 (2021)
5. Holalkere, N.S., Sahani, D.V., Blake, M.A., Halpern, E.F., Hahn, P.F., Mueller, P.R.: Characterization of small liver lesions: added role of MR after MDCT. J. Comput. Assist. Tomogr. **30**(4), 591–596 (2006)
6. Hu, T., et al.: Aorta-aware GAN for non-contrast to artery contrasted CT translation and its application to abdominal aortic aneurysm detection. Int. J. Comput. Assist. Radiol. Surg., 1–9 (2022)
7. Huang, W., et al.: LIDIA: precise liver tumor diagnosis on multi-phase contrast-enhanced CT via iterative fusion and asymmetric contrastive learning. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15009, pp. 394–404. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_38
8. Kim, A., et al.: S-SYNTH: knowledge-based, synthetic generation of skin images. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15003, pp. 734–744. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72384-1_69
9. Marckmann, P., et al.: Nephrogenic systemic fibrosis: suspected causative role of gadodiamide used for contrast-enhanced magnetic resonance imaging. J. Am. Soc. Nephrol. **17**(9), 2359–2362 (2006)

10. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
11. Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., Blundell, C.: Representation learning via invariant causal mechanisms. arXiv preprint arXiv:2010.07922 (2020)
12. Nie, D., et al.: Medical image synthesis with deep convolutional adversarial networks. IEEE Trans. Biomed. Eng. **65**(12), 2720–2730 (2018)
13. Ristea, N.C., et al.: CyTran: a cycle-consistent transformer with multi-level consistency for non-contrast to contrast CT translation. Neurocomputing **538**, 126211 (2023)
14. Tschannen, M., Bachem, O., Lucic, M.: Recent advances in autoencoder-based representation learning. arXiv preprint arXiv:1812.05069 (2018)
15. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
16. Zhu, Q., Wang, Y., Zhu, S., Du, B.: Partial consistent adversarial unified framework for unsupervised non-contrast CT cross-domain adaptation and segmentation. Pattern Recogn. **165**, 111638 (2025)
17. Zhu, Q., Wentland, A.L., Li, S.: Contrast-aware network with aggregated-interacted transformer and multi-granularity aligned contrastive learning for synthesizing contrast-enhanced abdomen CT imaging. IEEE Trans. Comput. Imaging (2025)
18. Zhu, Q., Zhu, S., Du, B., Wang, Y.: Cross-domain distribution adversarial diffusion model for synthesizing contrast-enhanced abdomen CT imaging. Pattern Recognit., 111695 (2025)
19. Zhu, X., Zhang, W., Li, Y., O'Donnell, L.J., Zhang, F.: When diffusion MRI meets diffusion model: a novel deep generative model for diffusion MRI generation. In: Linguraru, M.G., et al. (eds.) 2024. LNCS, vol. 15002, pp. 530–540. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72069-7_50



Contrastive Disentanglement Learning Framework for Multi-lead Wearable ECG Denoising

Yue Zhang[1,2,3], Chenyu Zhao[1,2,3], Wen Zhang[1,2,3], Jinliang Wang[4], Jun Guo[5], Wei Yang[1,2,3(✉)], and Qianjin Feng[1,2,3(✉)]

[1] School of Biomedical Engineering, Southern Medical University, Guangzhou, China
[2] Guangdong Provincial Key Laboratory of Medical Image Processing, Guangzhou, China
weiyanggm@gmail.com, fengqj99@smu.edu.cn
[3] Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Guangzhou, China
[4] CardioCloud Medical Technology (Beijing) Co., Ltd., Beijing, China
[5] Department of Cardiology, Chinese PLA General Hospital, Beijing, China

Abstract. Electrocardiogram (ECG) denoising enhances the clarity of noisy signals while preserving or even improving diagnostic performance. Most existing single-lead denoising algorithms require a preliminary noise assessment across all 12 leads, discarding clean leads and denoising only the noisy leads. In this paper, a novel disentanglement learning denoising network is proposed for 12-lead wearable ECG that directly processes 12-lead ECG, denoising noisy leads while preserving clean leads. Specifically, the proposed network takes both raw ECG and its corresponding simulated noisy ECG as inputs, disentangling them into noise codes and signal content codes. To ensure consistency between the content codes from two inputs, a discriminator is introduced. Additionally, considering that clean leads within the same ECG can provide valuable information for denoising noisy leads, a lead encoder is designed to extract lead specific features from the raw ECG. A contrastive loss is then applied between the features of noisy and clean leads to optimize the model. The results demonstrate that our method achieves superior denoising performance across two different lead system test datasets. Furthermore, evaluations on an ST-segment change multi-label classification task indicate that the denoised ECG improve diagnostic AUC and AUPRC. Furthermore, our model can be used into remote wearable ECG diagnostic workflows, providing preliminary noise reduction to assist cardiologists in subsequent clinical assessments.

Keywords: Multi-lead wearable ECG · Contrastive disentanglement learning · Denoising

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 120–129, 2026.
https://doi.org/10.1007/978-3-032-04947-6_12

1 Introduction

Electrocardiogram (ECG) is a cost-effective and non-invasive cardiac examination. It reflects the health status of the cardiovascular system and serves as an essential tool for diagnosing cardiovascular diseases. However, ECG is frequently affected by noises, especially during home self-monitoring due to improper operation or patient movement. Noisy ECG pose challenges for interpretation and hinders accurate diagnosis by cardiologists. To avoid the inconvenience of repeated data collection and to enhance the visual clarity of noisy ECG while preserving diagnostic performance, an effective denoising algorithm for wearable 12-lead ECG is required.

Previous research has made many attempts to enhance noisy ECG. Classical methods such as digital filtering remove noise through frequency domain transformations [1,10]. However, these methods perform poorly when ECG is severely noised. With the increasing availability of datasets and computational resources, machine learning and deep learning have gained popular in ECG denoising [2,8,18]. Various generative models have been investigated, employing supervised denoising by introducing simulated noise to clean ECG. Notable examples include generative adversarial networks (GANs) [20], and autoencoders [4], all of which have demonstrated promising results in enhancing ECG signal quality [6,12,16].

Despite the remarkable performance of deep learning ECG denoising methods, certain limitations hinder their practical application. Most methods are designed for single-lead ECGs [17], but clinical and wearable ECG devices commonly employ 12 leads configurations [7,9,15]. Notably, noise does not affect all leads uniformly, some leads may be noisy while others remain clean, as shown in Fig. 1(a). Consequently, single-lead methods require prior labeling of noisy leads, increasing the workload for cardiologists [3]. Moreover, it is worth exploring whether the inherent contrastive relationships between clean and noisy leads within the same ECG can improve denoising performance.

In this study, we propose a contrastive representation disentanglement framework for 12-lead wearable ECG denoising, termed CD-ECGNet. This framework takes 12-lead ECG as input and removes noise from noised leads while preserving clean leads. It reduces the workload for cardiologists by avoiding prior identification of noise leads. We assume a noisy ECG consists of a noise code (e.g., baseline wander) and a signal content code (e.g., cardiac information). To achieve effective denoising, these two codes are disentangled in the latent space, and the clean ECG is reconstructed using only the signal content while discarding the noise code. Since paired clean and noisy ECG are unavailable, simulated noise is introduced into clean ECG leads. Both the original ECG and the simulated ECG take the same disentanglement process in the latent space. To ensure that the signal content code remains noise-invariant between the original and simulated ECG, a domain discriminator is introduced, and adversarial learning is employed [21].

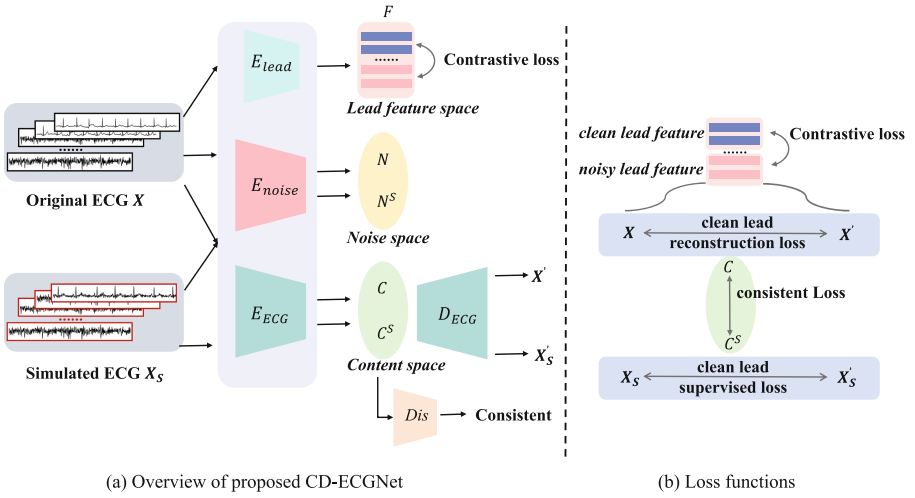
Moreover, clean leads can provide auxiliary supervision to facilitate the denoising of noisy leads during training. To leverage the contrastive relationships

between clean and noisy leads, we also introduce a lead encoder to capture inter-lead features. A contrastive loss is incorporated to encourage the model to learn the differences between noisy and clean leads, further enhancing denoising performance.

2 Methods

An overview of our proposed CD-ECGNet learning framework is shown in Fig. 1(a). Let X and X_S denote the original ECG and simulated ECG, respectively. The X may contain a mixture of clean and noisy leads, or consist entirely of clean or noisy leads. The X_S is generated by artificially adding noise into clean leads. $X^{'}$ and $X_S^{'}$ represent the denoised version of X and X_S. The final objective of this study is to learn a mapping from X to $X^{'}$.

The proposed CD-ECGNet consists of three encoders, a decoder, and a discriminator. Encoders E_{noise} and E_{ECG} are designed to disentangle the noise code and the signal content code. The E_{lead} extracts lead-specific features. Decoder D_{ECG} reconstructs the clean ECG using the signal content code. Discriminator Dis ensure the disentangled signal content remains consistent across two inputs.



(a) Overview of proposed CD-ECGNet (b) Loss functions

Fig. 1. Overview of proposed (a) CD-ECGNet and (b) loss functions for 12-lead ECG denoising.

2.1 Representation Distanglement and ECG Denoising

A possible approach for 12-lead ECG denoising is to introduce a clean-noisy lead classifier and only process noise leads. In this study, we aim to leverage clean leads to assist in denoising noisy leads during model training. Since representation disentanglement learning seeks to model latent space representations

such that they can be disentangled into multiple independent latent factors, we naturally extend this idea to ECG denoising [14].

The wearable ECG dataset lacks paired ECG. Instead, cardiologists have only annotated which leads contain noise. Since unsupervised ECG reconstruction performs poorly, we introduce simulated noise augmentation. During training, the model processes two types of inputs: X (e.g., in Fig. 1(a), leads I and II are clean), used to compute contrastive loss; X_S where clean leads are artificially noised, used to compute supervised loss.

As shown in Fig. 1(a), X is input into E_{noise}, E_{ECG} and E_{lead}, X_S is input into E_{noise} and E_{ECG}:

$$c = E_{ECG}(X), c_S = E_{ECG}(X_S) \tag{1}$$

$$n = E_{noise}(X), n_S = E_{noise}(X_S) \tag{2}$$

$$F = E_{lead}(X) \tag{3}$$

where c and c_S represent the signal content codes of the X andX_S, n and n_S denote corresponding noise codes. The E_{lead} performs convolutional along the signal length dimension . The extracted feature F has 12 channels, corresponding to the 12 ECG leads.

Given c and c_S, we can reconstruct a noise-free ECG:

$$X^{'} = D_{ECG}(c), X_S^{'} = D_{ECG}(c_S) \tag{4}$$

Since simulated noise may differ from wearable noise, we introduce a discriminator to ensure that the E_{ECG} can extract the same signal content code from the ECG with real wearable noise X and the ECG with simulated noise X_S. The Dis enforce similarity between n and n_S, helps model learn to separate noise from cardiac information, making the representations invariant to variations in noise types.

2.2 Loss Function

The loss function in CD-ECGNet as shown in Fig. 1(b).

1) Supervised loss: the original ECG X^c contains 12 channel, with clean lead X^{clean} and noise X^{noise}. The denoised $X_S^{'}$ corresponds to the $X_S^{'clean}$ and $X_S^{'noisy}$. The clean lead supervised loss is:

$$\mathcal{L}_{sup} = \left\| X^{clean}, X_s^{'clean} \right\|_1 \tag{5}$$

2) Reconstruction loss: this loss ensures that clean leads are not affected by the denoising. The denoised $X^{'}$ corresponds to the $X^{'clean}$ and $X^{'noisy}$. The clean lead reconstruction loss is:

$$\mathcal{L}_{rec} = \left\| X^{clean}, X^{'clean} \right\|_1 \tag{6}$$

3) Contrastive Loss: ECG consists of clean and noisy leads, it is clear that the clean feature F^{clean} and noisy feature F^{noisy} should be different. It is benefit from recognizing this difference when optimizing the model. Therefore, we introduce a contrastive loss to measure the similarity between clean and noisy features.

$$\mathcal{L}_{cont} = 1 - log(\frac{exp(comsim(\frac{F^{clean}}{\sigma}))}{exp(cossim(\frac{F^{clean}}{\sigma}, \frac{F^{noisy}}{\sigma}))}) \tag{7}$$

where $comsim(\frac{F^{clean}}{\sigma})$ represents the average cosine similarity for clean leads, $cossim(\frac{F^{clean}}{\sigma}, \frac{F^{noisy}}{\sigma})$ represents the average cosine similarity between clean leads and noisy leads, σ is the temperature coefficient in the contrastive loss.

4) Consistent loss: similar to GANs, we use adversarial learning to measure the distance between c and c_S, encourages consistency between c and c_S. We treat c and c_S as two classes with labels l and l_S, and adopt the Dis that discriminates c from c_S and outputs a value between l and l_S. The min-max game between the generator and the discriminator ensures that the learned features are consistent across X and X_s.

$$\mathcal{L}_{cons} = (Dis(c) - l)^2 + (Dis(c_S) - l_S)^2 \tag{8}$$

The overall objective is:

$$E, D = \min_{E,D} \max_{Dis} \{\mathcal{L}_{sup} + \mathcal{L}_{rec} + \mathcal{L}_{cont} + \lambda \mathcal{L}_{cons}\} \tag{9}$$

where E and D represent encoders and decoder, λ is a hyperparameter that controls the importance of the consistent loss, in this study, λ is 0.5.

2.3 Network Architecture

1) Encoder: The E_{ECG} and E_{noise} adopt a multi-scale 1D convolutional network, incorporating squeeze-and-excitation layers and residual connections [11]. Given an input tensor $x \in \mathbb{R}^{12 \times l}$, where l represents signal length and 12 denotes the number of leads, a permutation transformation is first applied to rearrange the input as $x \in \mathbb{R}^{l \times 12}$. Then 1D convolutional kernels of size 1 in lead-wise feature extractor E_{lead} progressively reducing the temporal dimension while preserving the number of channels. This design maintains lead-specific characteristics, preventing noise and clean leads from interfering with each other.
2) Decoder: The D_{ECG} is an adaptation of the 1-D U-Net up-sampling process. Each upsampling operation contains two convolutional layers with a kernel size of 3 and a 1D linear interpolation step [22].

3 Dataset

The data in this study contains a private wearable 12-lead ECG dataset and the publicly PTB-XL clinical 12-lead ECG dataset [19].

The wearable ECG used for developing our model was collected by the CardioCloud Medical Technology (Beijing) Company using a CONX CC1612 ECG device in a Mason-Likar system. The devices are widely used for daily health monitoring, self-checkups, and by rural doctors for assisted diagnosis. The dataset comprises 6,224 15 s ECG with a sampling frequency 500 Hz. The experiments follow a five-fold hold-out, with the dataset partitioned into training, validation, and test sets in an 8:1:1 ratio.

The PTB-XL was collected in hospital settings using the Wilson lead system, contains 21,837 10s ECG with a sampling frequency 500 Hz. The dataset used as an external test set. To ensure consistency in input length, these signals are zero-padded to 15 s.

To simulate noise conditions, we corrupt the ECG recordings using noise profiles from the Massachusetts Institute of Technology-Beth Israel Hospital Noise Stress Test Database (MIT-BIH NST) [13], which contains 3 half-hour real-world noise recordings. Electrodes were placed on limbs to avoid capturing visible ECG. We scale the noise amplitude by multiplying it with a random factor between 0.2 and 1.5.

4 Results and Analysis

4.1 Comparison with Existing Methods

In this section, we compared our CD-ECGNet against five representative from two main categories, two classical digital filter, FIR [10] and IIR [10]; three deep learning method, cGAN [20], Deep-Filter [16] and FCN-DAE [4]. To adapt the single-lead methods for our task, we modified the input and output channels to accommodate 12-lead ECG. As shown in Table 1, our method achieves improvements across multiple metrics in wearable testset. Table 2 presents the results on the PTB-XL testset, demonstrating that despite the differences in ECG devices and lead systems (wearable ECG using Mason-Likar system and PTB-XL using Wilson lead system), our approach outperforms others across all metrics, thereby highlighting its superior denoising performance and robustness.

Table 1. Wearable ECG testset evaluation results for different methods for random noise amplitude between 0.2 to 1.5.

Method	SSD (au)	PRD (%)	Cos sim (%)	SNR	RMSE
FIR [10]	380.04 ± 3.83	90.63 ± 0.17	78.99 ± 0.59	4.97 ± 0.07	0.37 ± 0.00
IIR [10]	378.92 ± 3.58	90.74 ± 0.29	79.10 ± 0.54	4.98 ± 0.07	0.37 ± 0.00
cGAN [20]	37.39 ± 1.62	41.11 ± 0.95	99.02 ± 0.07	18.26 ± 0.13	0.22 ± 0.00
Deep-Filter [16]	36.88 ± 3.10	38.96 ± 1.43	99.01 ± 0.09	17.99 ± 0.33	0.22 ± 0.01
FCNDAE [4]	67.17 ± 2.77	60.08 ± 1.57	98.28 ± 0.09	14.95 ± 0.12	0.26 ± 0.00
CD-ECGNet	$\mathbf{27.02 \pm 2.77}$	$\mathbf{34.51 \pm 1.65}$	$\mathbf{99.25 \pm 0.08}$	$\mathbf{18.81 \pm 0.39}$	$\mathbf{0.20 \pm 0.01}$

Table 2. PTB-XL dataset evaluation results for different methods for random noise amplitude between 0.2 to 1.5.

Method	SSD (au)	PRD (%)	Cos sim (%)	SNR	RMSE
FIR [10]	305.76 ± 5.62	90.48 ± 0.08	88.36 ± 0.19	4.73 ± 0.08	0.43 ± 0.00
IIR [10]	303.03 ± 7.40	90.44 ± 0.11	88.46 ± 0.25	4.77 ± 0.11	0.43 ± 0.00
cGAN [20]	27.82 ± 1.07	38.74 ± 0.83	99.10 ± 0.02	17.85 ± 0.06	0.23 ± 0.00
Deep-Filter [16]	24.83 ± 2.24	36.09 ± 2.04	99.16 ± 0.06	17.11 ± 0.32	0.23 ± 0.00
FCNDAE [4]	50.08 ± 1.89	56.68 ± 2.63	98.27 ± 0.05	14.16 ± 0.26	0.27 ± 0.00
CD-ECGNet	$\mathbf{23.52 \pm 1.33}$	$\mathbf{35.61 \pm 1.83}$	$\mathbf{99.21 \pm 0.03}$	$\mathbf{17.95 \pm 0.30}$	$\mathbf{0.22 \pm 0.00}$

4.2 Interpretative Visualization

Some visualized cases are given to analysis the performance of wearable ECG denoising. From the Fig. 2(a), it can be observed that CD-ECGNet effectively removes noise from the simulated ECG(blue signal). In Fig. 2(b), it is notable that the original clean signal (red signal) contains slight wearable noise, the denoised output (red signal) from our model appears even smoother than the original clean signal, demonstrates that our model can effectively denoise real-world wearable ECG signals.

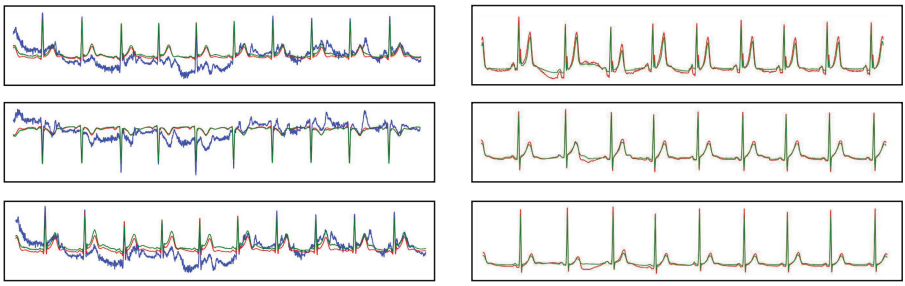


Fig. 2. Performance of the proposed method on the wearable dataset. The noisy leads by blue signal, the original clean lead by red signal, the denoised lead by green signal. (Color figure online)

4.3 Ablation Study

To evaluate the effectiveness of each module in our method, an ablation study is conducted using three variants: Baseline, Disentanglement Learning (DL), Disentanglement Learning with a Discriminator (DL+D). Our complete method is represented as CD-ECGNet, incorporates disentanglement learning and is optimized using both the consistent loss and lead contrastive loss. The results of the

ablation study are summarized in Table 3. It can be observed that disentanglement leaning significantly enhances denoising compare to the baseline. Further improvements are observed by the discriminator, highlighting the importance of enforcing content code consistency. The full model CD-ECGNet achieves the best overall performance by leveraging lead contrastive loss to enhance feature learning.

Table 3. Performance of ablation experiments on wearable testset.

Method	SSD (au)	PRD (%)	Cos sim (%)	SNR	RMSE
Baseline	51.65 ± 2.17	53.67 ± 1.91	98.70 ± 0.05	16.50 ± 0.20	0.24 ± 0.00
DL	40.88 ± 5.42	42.23 ± 2.98	98.92 ± 0.12	17.56 ± 0.59	0.22 ± 0.01
DL+D	30.89 ± 2.02	36.63 ± 1.02	99.14 ± 0.05	18.25 ± 0.35	0.21 ± 0.00
CD-ECGNet	$\mathbf{27.02 \pm 2.77}$	$\mathbf{34.51 \pm 1.65}$	$\mathbf{99.25 \pm 0.08}$	$\mathbf{18.81 \pm 0.39}$	$\mathbf{0.20 \pm 0.01}$

4.4 Impact of Denoising on Diagnosis

To assess CD-ECGNet enhances diagnostic performance, we conduct an imbalanced multi-label classification for ST-segment changes (STC)[14]. STC includes ST-segment elevation (STE) and depression (STD). The dataset consists of 4,000 ECG, including 180 with STE, 180 with STD, 180 with both, and 3,460 normal samples, positive samples account for only 1:11.

A five-fold cross-validation strategy is employed. In the 'w/o denoising' setting, the model is trained and evaluated on normalized ECG without denoising. In the 'with denoising' setting, ECG are first processed using CD-ECGNet before classification. Both experiments are conducted under identical conditions. As shown in Table 4. After denoising, AUPRC improves for minority classes, with STD increasing by 0.449% and STE by 2.837%.

Table 4. AUC and AUPRC performance for STC.

Method	AUC		AUPRC	
	STE	STD	STE	STD
w/o denoising	82.44 ± 1.62	82.89 ± 3.62	55.07 ± 4.54	50.28 ± 7.33
with denoising	$\mathbf{84.32 \pm 2.58}$	$\mathbf{83.75 \pm 3.86}$	$\mathbf{55.52 \pm 3.89}$	$\mathbf{53.12 \pm 4.83}$

5 Conclusion

We propose a 12-lead ECG denoising framework based on disentanglement learning, which disentangle ECG into noisy and content code, using clean leads to assist in denoising noisy leads. Our model is trained on wearable ECG and tested on both wearable and hospital datasets, demonstrating its effectiveness. Additionally, we evaluated diagnostic performance after denoising using the STC classification task, where our method improved diagnostic AUC and AUPRC. Our approach can be extended to remote wearable ECG applications, enhancing visual clarity and diagnostic performance. Since the simulated noise is based on the Wilson lead system, a domain shift exists between it and wearable ECG [5]. In the future, we aim to address this domain shift to enhance model's robustness.

Acknowledgments. This work was supported by a grant from the National Key R&D Program of China (No. 2018YFC2001203).

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Barati, Z., Ayatollahi, A.: Baseline wandering removal by using independent component analysis to single-channel ECG data. In: 2006 International Conference on Biomedical and Pharmaceutical Engineering, pp. 152–156. IEEE (2006)
2. Chen, M., et al.: Elimination of random mixed noise in ECG using convolutional denoising autoencoder with transformer encoder. IEEE J. Biomed. Health Inform. (2024)
3. Chen, M.: Multitask learning-based quality assessment and denoising of electrocardiogram signals. IEEE Trans. Instrum. Meas. **73**, 1–13 (2024)
4. Chiang, H.T., Hsieh, Y.Y., Fu, S.W., Hung, K.H., Tsao, Y., Chien, S.Y.: Noise reduction in ECG signals using fully convolutional denoising autoencoders. IEEE Access **7**, 60806–60813 (2019)
5. Du, M., Liang, K., Zhang, L., Gao, H., Liu, Y., Xing, Y.: Deep-learning-based metal artefact reduction with unsupervised domain adaptation regularization for practical CT images. IEEE Trans. Med. Imaging **42**(8), 2133–2145 (2023)
6. Gao, N., et al.: Frequency information enhanced half instance normalization network for denoising electrocardiograms. Biomed. Signal Process. Control **102**, 107225 (2025)
7. Hughes, J.W., et al.: Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. JAMA Cardiol. **6**(11), 1285–1295 (2021)
8. Jin, Y., Qin, C., Liu, J., Liu, Y., Li, Z., Liu, C.: A novel deep wavelet convolutional neural network for actual ECG signal denoising. Biomed. Signal Process. Control **87**, 105480 (2024)
9. Joo, J., Joo, G., Kim, Y., Jin, M.N., Park, J., Im, H.: Twelve-lead ECG reconstruction from single-lead signals using generative adversarial networks. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14226, pp. 184–194. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43990-2_18

10. Kumar, K.S., Yazdanpanah, B., Kumar, P.R.: Removal of noise from electrocardiogram using digital FIR and IIR filters with various methods. In: 2015 International Conference on Communications and Signal Processing (ICCSP), pp. 0157–0162. IEEE (2015)
11. Lai, J., et al.: Multi-expert ensemble ECG diagnostic algorithm using mutually exclusive–symbiotic correlation between 254 hierarchical multiple labels. npj Cardiovasc. Health **1**(1), 8 (2024)
12. Lomoio, U., Veltri, P., Guzzi, P.H., Liò, P.: Design and use of a denoising convolutional autoencoder for reconstructing electrocardiogram signals at super resolution. Artif. Intell. Med. **160**, 103058 (2025)
13. Moody, G.B., Muldrow, W., Mark, R.G.: A noise stress test for arrhythmia detectors. Comput. Cardiol. **11**(3), 381–384 (1984)
14. Neshatavar, R., Yavartanoo, M., Son, S., Lee, K.M.: CVF-SID: cyclic multi-variate function for self-supervised image denoising by disentangling noise from image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17583–17591 (2022)
15. Ribeiro, A.H., et al.: Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat. Commun. **11**(1), 1760 (2020)
16. Romero, F.P., Piñol, D.C., Vázquez-Seisdedos, C.R.: DeepFilter: an ECG baseline wander removal filter using deep learning techniques. Biomed. Signal Process. Control **70**, 102992 (2021)
17. Singh, P., Sharma, A.: Attention-based convolutional denoising autoencoder for two-lead ECG denoising and arrhythmia classification. IEEE Trans. Instrum. Meas. **71**, 1–10 (2022)
18. de Vries, I., de Jong, A., van Laar, J., Vullings, R., et al.: Deep learning for sparse domain Kalman filtering with applications on ECG denoising and motility estimation. IEEE Trans. Biomed. Eng. (2024)
19. Wagner, P., et al.: PTB-XL, a large publicly available electrocardiography dataset. Sci. Data **7**(1), 1–15 (2020)
20. Wang, X., et al.: An ECG signal denoising method using conditional generative adversarial net. IEEE J. Biomed. Health Inform. **26**(7), 2929–2940 (2022)
21. Yang, H., et al.: Unsupervised MR-to-CT synthesis using structure-constrained cycleGAN. IEEE Trans. Med. Imaging **39**(12), 4249–4261 (2020)
22. Zhang, Y., et al.: Abnormal recognition-assisted and onset-offset aware network for pathological wearable ECG delineation. Artif. Intell. Med. **157**, 102992 (2024)

CS^{2}C: Collaborative Spatial and Spectral Neural Clustering for Organelle Segmentation from Volumetric Electron Microscopy

Jimao Jiang and Yuru Pei(✉)

School of Intelligence Science and Technology, Key Laboratory of Machine Perception (MOE), State Key Laboratory of General Artificial Intelligence, Peking University, Beijing 100871, China
peiyuru@cis.pku.edu.cn

Abstract. Organelle segmentation is crucial for understanding the morphology of biological structures. Existing unsupervised methods leverage powerful feature extractors and clustering techniques to uncover organelle structures from volumetric electron microscopy images. However, these methods often struggle with noisy microscopy images and the computational complexity of numerical clustering. In this paper, we propose CS^{2}C, a novel collaborative spatial and spectral deep neural clustering framework, for multi-class organelle segmentation. The pillar of our approach is combining unsupervised deep spectral clustering and spatial clustering, which enhances a harmony of learned cluster assignments under the spatial and spectral superpixel-wise representation. Specifically, we adopt a masked autoencoder-based feature extractor to obtain powerful superpixel features, where spatial clustering is performed directly on these features. Beyond that, spectral clustering is applied in the spectral domain, naturally alleviating high-frequency perturbations in the image features. The entire framework is trained end-to-end using a combination of clustering loss and consistency regularization between spatial and spectral clustering. Extensive experiments demonstrate that our method outperforms state-of-the-art unsupervised methods on known benchmarks. Code is available at: https://github.com/JimaoJIANG/CS2C.

Keywords: Collaborative spatial and spectral clustering · Organelle segmentation · Volumetric electron microscopy

1 Introduction

Nanometer-level organelle segmentation plays a crucial role in understanding the intricate morphology and organization of cellular structures [11,16,20,22]. High-throughput imaging techniques, such as volumetric electron microscopy (VEM),

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 130–139, 2026.
https://doi.org/10.1007/978-3-032-04947-6_13

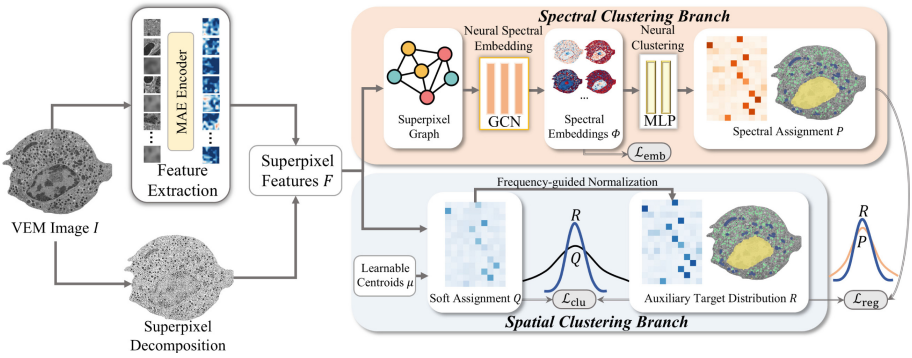


Fig. 1. Illustration of the proposed collaborative spatial and spectral neural clustering framework. When given a VEM slice image I, we conduct superpixel decomposition and adopts the a masked autoencoder (MAE) to extract superpixel features F. In the spatial clustering branch, we compute the soft cluster assignment probability Q and the axillary target distribution R. The cluster centroids are optimized by minimizing KL-divergence-based clustering loss $\mathcal{L}_{clu}$. In the spectral clustering branch, we feed the superpixel graph to the neural spectral embedding module for approximated spectral bases Φ, which are further used to infer the spectral clustering assignment matrix P via an MLP-based clustering module. We impose a consistency regularization $\mathcal{L}_{reg}$ of the spatial and spectral clustering. Our approach enables an end-to-end organelle segmentation with consistent cluster assignment across images.

provide high-resolution imaging of organelle structures with great efficiency. While manual segmentation remains the gold standard for a variety of downstream biological tasks, it is labor-intensive and relies heavily on expert knowledge. Transformer-based self-supervised learning techniques [4,8] have shown great potential in capturing both local and global spatial correlations, making them well-suited for robust representations of subcellular structures. The deep neural network-based representation learning facilitates scalable and automated structural segmentation from microscopy images [7,10,20].

Clustering has long been a powerful unsupervised segmentation approach for data grouping based on feature similarity, and can be performed in either spatial or spectral domains. Spatial clustering methods typically operate directly on extracted image features by identifying local pixel similarities. However, these methods are sensitive to high-frequency noise in microscopy images, often leading to suboptimal segmentation results. On the other hand, spectral clustering operates in the spectral embedding space and is more robust to high-frequency perturbations. However, numerical spectral embedding and clustering can be computationally expensive when dealing with large-scale graphs [6]. Furthermore, additional synchronization operations [14] are required to maintain consistent cluster assignments across images.

In this paper, we exploit collaborative spatial and spectral clustering learning and propose a novel unsupervised two-branch deep neural clustering framework,

CS^2C, for organelle segmentation, as shown in Fig. 1. Drawing inspiration from advances in self-supervised representation learning and deep embedding clustering [19], our framework leverages spatial clustering on rich semantic features extracted from the MAE model. In light of high frequency perturbations in microscopy image features, we exploit the deep spectral embedding and conduct clustering in the low-dimensional spectral embedding space, which is represented by the approximated spectral basis of the graph Laplacian matrix regarding the superpixel graph. We train the two-branch clustering model end-to-end by combining the clustering loss and the consistency regularization between spatial and spectral clustering assignments. Our approach promotes a unified solution to organelle segmentation in both spatial and spectral domains. Extensive experimental results demonstrate that our method outperforms state-of-the-art techniques on existing benchmarks. Our main contributions are as follows:

- We introduce an unsupervised collaborative neural spatial and spectral clustering framework for efficient organelle segmentation from VEM images.
- We present a simple yet effective two-branch design for unsupervised deep clustering, which mitigates high-frequency perturbations in microscopy image features and promotes consistent clustering in both spatial and spectral domains.
- We validate the effectiveness of our approach through comprehensive experiments, demonstrating superior performance in multi-class organelle segmentation compared to state-of-the-art methods.

2 Method

Given the input VEM images, the objective is to train a clustering model that takes a VEM slice I as input and outputs clustering assignments P for various organelle structures. As shown in Fig. 1, our framework follows the prevailing paradigm of feature extraction and clustering inference based on feature similarity. The core of our framework is harnessing a two-branch spatial-spectral clustering model for organelle segmentation. The spatial clustering branch performs clustering directly on the rich semantic superpixel features extracted by the MAE model, while the spectral clustering branch conducts neural spectral embedding and clustering in a low-dimensional space spanned by the approximated spectral basis of the graph Laplacian matrix. We introduce an end-to-end training strategy that combines a clustering loss with a consistency regularization between spatial and spectral clustering assignments. Our approach mitigates high-frequency perturbations in microscopy image features and encourages spatial-spectral consistency in clustering.

Superpixel Features. Given the fine granularity of organelle structures, the patches used in the vision transformer model must be small enough to capture detailed shapes and appearances. However, small patches increase computational complexity in feature extraction [8]. To address this, we convert the

high-resolution microscopy image into a more compact representation via superpixel decomposition [1] instead of using regular patches. We employ MAE-based feature extraction [8], which has demonstrated effectiveness in handling fine-grained, repetitive subcellular structures [10]. We up-sample the feature embedding from the pre-trained MAE encoder to obtain l-channel features $F \in \mathbb{R}^{m \times n \times l}$ for a VEM image with a resolution of $m \times n$. The superpixel feature is then defined as the concatenation of mean and standard deviation of the pixels within each superpixel.

2.1 Collaborative Spatial-Spectral Neural Clustering

The core of our framework consists of two parallel clustering models: a spatial model and a spectral model. They predict clustering assignments from superpixel features, leveraging the spectral clustering's resilience to high-frequency perturbations and the alignment with centroid-based probability distributions in the original feature space.

Spectral Clustering. In the spectral branch, organelle segmentation is formulated as a cut on the superpixel graph within the spectral embedding space. Traditional independent per-graph spectral clustering methods often overlook cross-image relationships, leading to inconsistent cluster assignments. To address this, we propose a deep neural spectral clustering model that leverages its generalization capabilities to ensure consistent labeling across images. We generate a superpixel graph using MAE features, where the affinity matrix $A \in \mathbb{R}^{n_s \times n_s}$ is defined with the RBF kernel, and $a_{ij} = \exp \frac{-\|F_i - F_j\|_2^2}{\kappa}$ for $1 \leq i, j \leq n_s$. κ is the kernel's variance, and n_s is the number of superpixels. We further reduce the bandwidth and eliminate weakly correlated values as [2], and $A \leftarrow \max\left(A - \frac{\max A}{\alpha}, 0\right)$. Where α controls the strength of the repulsion forces that diminish weak connections.

We utilize a Graph Convolutional Network (GCN)-based spectral embedding module to approximate the spectral basis $\Phi \in \mathbb{R}^{n_s \times u}$ of the graph Laplacian matrix L, avoiding eigenvector switching and sign flipping problems in the computationally expensive numerical eigendecomposition. u denotes the number of approximated spectral bases. The spectral embedding is optimized using two constraints: (1) the diagonalization of the symmetric Laplacian matrix L, and (2) the orthogonality of the approximated spectral basis Φ, as [18]. The spectral embedding loss:

$$\mathcal{L}_{emb} = \sum_{i \neq j} \left[(\phi_i^T L \phi_j)^2 + \phi_i^T \phi_j \right], \tag{1}$$

where ϕ_i and ϕ_j are the i-th and j-th column vectors of Φ. The eigendecomposition of the symmetric Laplacian L satisfies $\Lambda = \Phi^T L \Phi$, where Λ is a diagonal matrix containing the eigenvalues. The loss ensures that off-diagonal values of $\Lambda = \Phi^T L \Phi$ approach zero and enforces the orthogonality of Φ.

A neural clustering module f_θ is introduced to assign cluster labels based on the spectral embedding, eliminating the need for label synchronization like

Hungarian matching [2,14]. The cluster assignment probability matrix $P = \text{softmax}(f_\theta(\Phi))$ defines a cut on the superpixel graph using the perturbation-resistant spectral embedding.

Spatial Clustering. In the spatial branch, clustering is directly applied to the MAE features. Inspired by deep embedding clustering methods [19], we optimize learnable cluster centroids by minimizing the distance between a soft cluster assignment and an auxiliary target distribution. The soft cluster assignment matrix Q is computed using a Student's t-distribution kernel [12], and $q_{i,j} = \frac{(1+\|F_{s,i}-\mu_j\|/\tau)^{-(\tau+1)/2}}{\sum_{j'}(1+\|F_{s,i}-\mu_{j'}\|/\tau)^{-(\tau+1)/2}}$. $\mu \in \mathbb{R}^{k\times 2l}$ represents the learnable centroids, initialized using k-means, and τ controls the degrees of freedom. The auxiliary target distribution R is derived by frequency-guided normalization of Q, and

$$r_{i,j} = \frac{q_{i,j}^2 / \sum_k q_{k,j}}{\sum_{j'} \left(q_{i,j'}^2 / \sum_k q_{k,j'} \right)}. \tag{2}$$

We optimize the cluster centroids by minimizing the Kullback-Leibler (KL) divergence between the soft assignment Q and the target distribution R. The clustering loss is defined as:

$$\mathcal{L}_{clu} = \text{KL}(R\|Q) = \sum_i \sum_j r_{i,j} \log \frac{r_{i,j}}{q_{i,j}}. \tag{3}$$

Minimizing the loss $\mathcal{L}_{clu}$ refines the centroids μ, improving cluster purity and focusing on high-confidence assignments.

Spatial-Spectral Consistency Regularization. Spectral clustering is known to be robust against high-frequency perturbations in image features, owing to the low-pass nature of the graph Laplacian's eigenbasis approximation. Spatial clustering, when applied to MAE features, is effective for iterative refinement of soft cluster assignments. To exploit the strengths of both clustering methods, we propose minimizing the distance between their respective probabilistic assignment matrices. Specifically, we use the KL divergence to measure the dissimilarity between the spectral clustering matrix P and the spatial clustering matrix R. The spatial and spectral consistency regularization loss $\mathcal{L}_{reg}$ is defined as:

$$\mathcal{L}_{reg} = \text{KL}(P\|R) = \sum_i \sum_j p_{i,j} \log \frac{p_{i,j}}{r_{i,j}}. \tag{4}$$

Unlike cross-entropy loss, KL divergence updates the model more gently [3], helping to prevent abrupt shifts in data representations.

2.2 Training Loss

The overall training loss is a weighted combination of the neural spectral embedding loss $\mathcal{L}_{emb}$, the clustering loss $\mathcal{L}_{clu}$, and the spatial-spectral consistency regularization loss $\mathcal{L}_{reg}$.

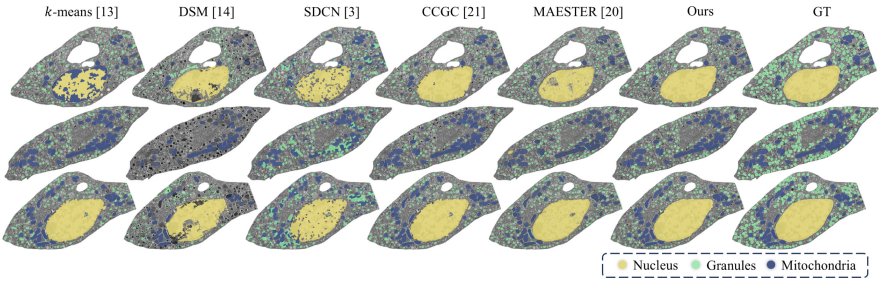


Fig. 2. Qualitative comparison of organelle segmentation by compared methods.

$$\mathcal{L} = \mathcal{L}_{emb} + \gamma_1 \mathcal{L}_{clu} + \gamma_2 \mathcal{L}_{reg}, \tag{5}$$

where γ_1 and γ_2 are hyperparameters that balance the regularized spatial and spectral clustering. Minimizing the combined loss function $\mathcal{L}$ optimizes both the neural spectral and spatial clustering modules simultaneously.

In the online inference process, given a VEM image I, first, the MAE-based feature extractor generates superpixel features, which are then used to construct the superpixel graph. Second, the spectral clustering module takes the weighted superpixel graph as an input and assigns each superpixel to a cluster, producing the clustering assignment matrix P and achieving the organelle segmentation.

3 Experiments

Datasets and Metrics. We evaluate the proposed CS^{2}C on the BetaSeg [9, 15], which consists of four preprocessed primary mouse pancreatic islet β cell volumes. The first three volumes are used for training, and the last one for testing as [20]. We divide the VEM slices into superpixels with a physical size of around 5 µm squared area using the SLIC [1] algorithm, with the compactness set to 0.2. During training, we use slices with a resolution of 560×560, containing approx. 3000 superpixels. We evaluate the performance using Dice Similarity Coefficient (DSC) and Intersection over Union (IoU), which measure the consistency of the predicted segmentation with the ground truth.

Implementation Details. We use the MAE based feature extractor [20], with the feature channel number l of 192. κ in RBF kernel, α in affinity matrix, and τ in Student's t-distribution kernel are set to 2, 4 and 1, respectively. We retain $u = 12$ spectral bases and set the cluster number k to 14 according to the number of organelle types empirically. The hyperparameters γ_1 and γ_2 are set to 1 in the loss function $\mathcal{L}$. The spectral embedding module includes three linear GCN layers with 384×96, 96×48, and 48×12 weight matrices. The MLP-based clustering module has two fully connected layers with dimensions of 14×14 and 14×8. We implement the proposed CS^{2}C using the PyTorch toolkit on a PC with an NVIDIA RTX 2080Ti GPU. We use the Adam optimizer with a momentum of 0.9 and 0.999. The learning rate is set to 0.01 for the first 3,000

Table 1. Segmentation results regarding the DSC and IoU by compared methods and variants without spatial (`spa`) or spectral (`spe`) branches. We report the output clustering, CS^2C_{spe} and CS^2C_{spa}, from the spectral and spatial branches. (sp: superpixel)

	Feature Genre	Use Graph	Nucleus		Granules		Mitochondria		*Average*	
			DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑
k-means [13]	sp	No	0.927	0.864	0.442	0.284	0.754	0.605	0.708	0.584
Classical SC [17]	sp	Yes	0.903	0.823	0.365	0.224	0.769	0.625	0.679	0.557
DSM [14]	patch	Yes	0.848	0.735	0.347	0.210	0.743	0.591	0.646	0.512
DeepCUT [2]	patch	Yes	0.861	0.756	0.333	0.200	0.694	0.531	0.629	0.495
FastDGC [5]	sp	Yes	0.651	0.482	0.484	0.319	0.658	0.490	0.597	0.431
SDCN [3]	sp	Yes	0.902	0.821	0.442	0.284	0.665	0.498	0.669	0.534
CCGC [21]	sp	Yes	0.957	0.917	0.452	0.292	0.835	0.717	0.748	0.642
MAESTER [20]	pixel	No	0.943	0.892	0.556	0.385	0.778	0.636	0.759	0.638
w/o `spe`	sp	No	0.952	0.909	0.575	0.404	0.824	0.701	0.784	0.671
w/o `spa`	sp	Yes	0.954	0.913	0.573	0.401	0.861	0.756	0.796	0.690
CS^2C_{spa}	sp	No	0.958	0.919	0.601	0.430	0.838	0.722	0.799	0.690
CS^2C_{spe} (Ours)	sp	Yes	**0.966**	**0.935**	**0.603**	**0.431**	**0.872**	**0.772**	**0.814**	**0.713**

iterations and then reduced to 0.0001 for 2,000 iterations. The mini-batch size is set to 1. The training takes approx. 6 h. In the online testing process, the feature extraction, spatial clustering, and spectral clustering of a 1082×545 image take 2.714 s, 0.005 s, and 0.002 s.

Experimental Results. We compare our approach to state-of-the-art deep clustering methods, including DSM [14], DeepCUT [2], FastDGC [5], SDCN [3], CCGC [21], and MAESTER [20], as well as methods that apply classical k-means [13] and spectral clustering [17] on MAE-based superpixel features. To ensure a fair comparison, we use the same superpixel features and graphs for compared deep clustering methods [3,5,21] as ours. Table 1 and Fig. 2 present quantitative and qualitative results for organelle segmentation. Our method consistently outperforms all baselines. The compared methods rely on deep neural network-based feature extraction, such as DINOViT [2,14] and MAE [20]. Clustering is performed either on individual pixels [20], patch graphs [2,14], or superpixel graphs as in our approach. Pixel-level clustering [20] generally outperforms patch-based clustering [2,14], though it incurs a higher computational cost during both training and testing. Notably, compared to MAESTER [20], which uses k-means on pixel-level MAE features, our method uses a compact superpixel graph that reduces computational time by more than a factor of 137, while achieving a DSC boost of 0.023 (nucleus), 0.047 (granules), and 0.094 (mitochondria). Figure 3(b) shows the effect of Gaussian noise levels from 1% to 8% on organelle segmentation. Across all noise levels, the performance drop of MAESTER [20] is approximately twice as large as that of our method. Superior performances of our approach can be attributed to spectral embedding and the collaborative learning strategy, mitigating negative impacts of noisy features.

Table 2. Organelle segmentation accuracy when using different cluster number k.

k	Nucleus		Granules		Mitochondria		*Average*	
	DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑	DSC↑	IoU↑
8	0.940	0.887	0.243	0.139	0.645	0.476	0.609	0.500
10	0.948	0.900	0.181	0.099	0.666	0.495	0.597	0.498
12	**0.968**	**0.938**	0.568	0.397	0.869	0.768	0.802	0.701
14	0.966	0.935	**0.603**	**0.431**	**0.872**	**0.772**	**0.814**	**0.713**
16	0.964	0.931	0.584	0.412	0.870	0.770	0.806	0.704
18	0.965	0.933	0.513	0.345	0.871	0.771	0.783	0.683

Our method excels by harmonizing spatial and spectral clustering, offering performance gains over approaches using k-means or neural clustering networks directly on superpixel features [3,5,21], as well as deep spectral clustering methods [14,17]. This makes our CS^{2}C SOTA among unsupervised baselines, and demonstrates the power by collaborative neural spatial and spectral clustering with generalization capacity and consistent clustering across images.

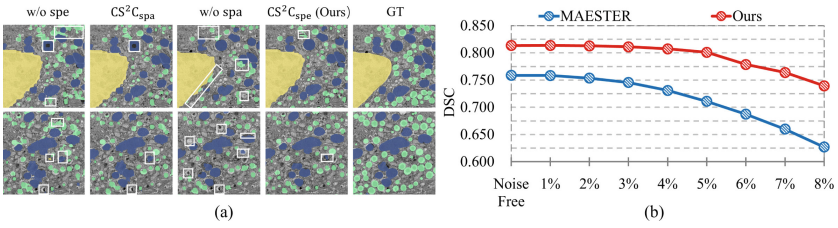


Fig. 3. (a) Organelle segmentation by variants of the proposed method without the spectral branch (`spe`) or the spatial branch (`spa`). We illustrate the output clustering, CS^{2}C$_{\texttt{spe}}$ and CS^{2}C$_{\texttt{spa}}$, from the spatial and spectral branches. Errors are white blocked. (b) Organelle segmentation accuracy when confronted with noisy VEM images.

Ablation Study. We conduct an ablation study to assess the contributions of the spatial and spectral clustering branches as shown in Table 1 and Fig. 3(a). We compare our full framework with two variant models: (1) w/o `spe`, which removes the spectral branch and uses only $\mathcal{L}_{clu}$ for training. This setup combines DEC [19] with frozen MAE features. (2) w/o `spa`, which removes the spatial branch and relies on $\mathcal{L}_{emb}$ and k-means supervision for training. The results demonstrate that both the spatial and spectral branches contribute to final clustering. Specifically, the spatial and spectral branches yield IoU gains of 0.016 and 0.012, respectively. The spatial branch also improves structural segmentation along nucleus boundaries, helping to prevent label omissions (Fig. 3(a)). Overall, the proposed CS^{2}C, which uses both spatial and spectral branches in parallel, achieves consistent organelle segmentation aligned with the ground truth.

Parameter Analysis. Table 2 demonstrates the impact of cluster number k on subcellular segmentation performance. The results show that performance is insensitive with moderate cluster numbers. However, both excessively small and large cluster numbers have limitations. With too few clusters, granules and mitochondria become difficult to distinguish, while too many clusters hinder the capture of fine-grained granules, resulting in reduced performance.

4 Conclusion

We presented a collaborative spatial-spectral neural clustering approach for organelle segmentation, which integrates neural clustering in both the spatial and spectral domains. Our method introduces a novel two-branch clustering framework that leverages the robustness of spectral clustering to high-frequency perturbations in image features, while also utilizing spatial clustering for iterative refinement of cluster assignments. We demonstrate the effectiveness of our approach with state-of-the-art accuracy in organelle segmentation.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China under Grant 62272011 and Beijing Natural Science Foundation 7232337.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 2274–2282 (2012)
2. Aflalo, A., Bagon, S., Kashti, T., Eldar, Y.C.: DeepCut: unsupervised segmentation using graph neural networks clustering. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 32–41 (2022). https://api.semanticscholar.org/CorpusID:254564022
3. Bo, D., Wang, X., Shi, C., Zhu, M., Lu, E., Cui, P.: Structural deep clustering network. In: Proceedings of the Web Conference 2020, pp. 1400–1410 (2020)
4. Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9630–9640 (2021). https://api.semanticscholar.org/CorpusID:233444273
5. Ding, S., et al.: Towards faster deep graph clustering via efficient graph autoencoder. ACM Trans. Knowl. Discov. Data **18**(8), 1–23 (2024)
6. Efroni, O., Ginzburg, D., Raviv, D.: Spectral teacher for a spatial student: spectrum-aware real-time dense shape correspondence. In: 2022 International Conference on 3D Vision (3DV), pp. 1–10. IEEE (2022)
7. Han, H., Dmitrieva, M., Sauer, A., Tam, K.H., Rittscher, J.: Self-supervised voxel-level representation rediscovers subcellular structures in volume electron microscopy. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1873–1882 (2022). https://api.semanticscholar.org/CorpusID:250980726

8. He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15979–15988 (2021). https://api.semanticscholar.org/CorpusID:243985980
9. Heinrich, L., et al.: Whole-cell organelle segmentation in volume electron microscopy. Nature **599**, 141–146 (2021). https://api.semanticscholar.org/CorpusID:238421373
10. Kraus, O., et al.: Masked autoencoders for microscopy are scalable learners of cellular biology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11757–11768 (2024)
11. Lu, Z., et al.: Long-term intravital subcellular imaging with confocal scanning light-field microscopy. Nat. Biotechnol., 1–12 (2024)
12. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
13. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press (1967)
14. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: Deep spectral methods: a surprisingly strong baseline for unsupervised semantic segmentation and localization. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8354–8365 (2022). https://api.semanticscholar.org/CorpusID:248811034
15. Müller, A., et al.: 3D FIB-SEM reconstruction of microtubule–organelle interaction in whole primary mouse β cells. J. Cell Biol. **220**(2) (2021)
16. Seal, S., et al.: Cell painting: a decade of discovery and innovation in cellular imaging. Nat. Methods, 1–15 (2024)
17. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000)
18. Sun, D., Pei, Y., Zhang, Y., Xu, T., Wang, T., Yan Zha, H.: Dense correspondence of deformable volumetric images via deep spectral embedding and descriptor learning. Med. Image Anal. **82**, 102604 (2022). https://api.semanticscholar.org/CorpusID:251947036
19. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning, pp. 478–487. PMLR (2016)
20. Xie, R., Pang, K., Bader, G.D., Wang, B.: MAESTER: masked autoencoder guided segmentation at pixel resolution for accurate, self-supervised subcellular structure recognition. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3292–3301 (2023). https://api.semanticscholar.org/CorpusID:261080761
21. Yang, X., et al.: Cluster-guided contrastive graph clustering network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 10834–10842 (2023)
22. Zhou, D., et al.: RepMode: learning to re-parameterize diverse experts for subcellular structure prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3312–3322 (2023)

CT-Based Hippocampus Segmentation with Dual-Decoder Network (HDD-Net)

Wonjun Son[1], Ji Young Lee[2], Sung Jun Ahn[3], and Hyunyeol Lee[1(✉)]

[1] School of Electronic and Electrical Engineering and IEDT, Kyungpook National University, Daegu, Republic of Korea
hyunyeollee@knu.ac.kr

[2] Department of Radiology, Seoul St. Mary's Hospital, Catholic University College of Medicine, Seoul, Republic of Korea

[3] Department of Radiology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea

Abstract. The hippocampus in the brain performs a pivotal role for memory formation, spatial navigation, and emotional regulation. Its volume and morphology are known to change with the progression of neurodegenerative diseases such as Alzheimer's disease. Hence, hippocampal atrophy serves as a key biomarker for early diagnosis and monitoring of such diseases. Whereas MRI has been predominantly employed in that regard due to its excellent soft-tissue contrast, CT-based segmentation of the structure has been relatively far less explored because the modality results in ambiguous boundaries between brain subregions. This study aims to address this technical challenge, achieving accurate segmentation of the hippocampus on CT images. To this end, we develop a deep learning model, termed 'Hippocampus Dual Decoder Network (HDD-Net)', characterized by the following four major components: 1) parallel, dual decoders that segment the hippocampal region and its boundaries, respectively, 2) a single, shared encoder in which features combined across multiple blocks are refined via attention, 3) a feature fusion module (FFM) that performs inter-decoder featural supplements, and 4) a cross loss to jointly optimize segmentation and edge predictions. HDD-Net was validated using both internal and external datasets, with its performance assessed using Dice similarity coefficient (DSC) and intersection-over-union (IoU). Our model yielded DSC = 0.823 ± 0.03 and IoU = 0.701 ± 0.04, and DSC = 0.759 ± 0.07 and IoU = 0.617 ± 0.09 for internal and external test datasets, respectively, outperforming seven other SOTA methods. Furthermore, volumetric analysis revealed a good agreement between MRI- and CT-derived hippocampal masks. Our findings suggest feasibility of CT-based hippocampal segmentation via HDD-Net, as a cost-effective alternative to MRI. The implementation of HDD-Net is available at https://github.com/sonwonjun103/HDD_Net.

Keywords: Computed Tomography · Hippocampus Segmentation · Hippocampus Dual Decoder Network (HDD-Net) · Deep Learning

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 140–149, 2026.
https://doi.org/10.1007/978-3-032-04947-6_14

1 Introduction

The hippocampus in the brain plays a crucial role in memory formation, spatial navigation, and emotional regulation, and abnormalities in its volume and morphology linked to Alzheimer's disease, epilepsy, PTSD, schizophrenia, and depression [1–3]. Magnetic resonance imaging (MRI) has been regarded as the gold standard for hippocampal segmentation due to its excellent soft-tissue contrast. Nevertheless, high cost and long acquisition time make its application difficult in certain clinical settings, particularly with limited-resource environments. Computed tomography (CT) can be considered as a more accessible and cost-effective alternative to MRI. However, near-flat contrast across brain subregions in CT images renders hippocampal segmentation from the modality challenging.

Recent advances in CNN-based architectures like U-Net [4] and transformer-based models such as UNETR [5] and TransUNet [6] have significantly improved segmentation performance. Accordingly, deep learning (DL)-based medical image segmentation has been extensively explored [7, 8], yet a majority of which are on MRI [9–12]. One of very few studies attempting DL segmentation of the hippocampus from CT scans is AG-3D ResNet by Portal et al. [13]. The authors integrated attention mechanisms into a 3D ResNet [14] backbone as a means to enhance feature extraction, and showed feasibility of CT-based hippocampal segmentation using DL.

In this work, we were aimed to enhance performance of DL segmentation of hippocampus on CT head images. To achieve this goal, we conceived a novel DL model, termed "Hippocampus Dual Decoder Network (HDD-Net)", and evaluated its performance in reference to MRI-derived hippocampal labels.

2 Methods

The overall architecture of HDD-Net, shown in Fig. 1, is designed to address the challenges of hippocampal segmentation on CT images. It consists of four key components: dual-decoder, a shared encoder, feature fusion module (FFM), and a cross loss function. By leveraging its dual-decoder structure, the network takes preprocessed volumetric 3D CT images as input and produces two outputs: hippocampal masks and corresponding edges. Details in each component are described in the following subsections.

2.1 Dual Decoder

The dual decoder architecture enables HDD-Net to leverage both volume- and edge-specific information, thereby achieving performance elevation in hippocampal segmentation. It consists of a Segmentation Decoder, which predicts the hippocampal region, and an Edge Decoder, which extracts boundaries of the structure. It ensures that the model captures complementary aspects of the hippocampal structure, addressing the challenges posed by its small size and complex anatomy.

Both decoders share a same structure with different training parameters, comprising a series of up-sampling and convolutional layers to progressively restore spatial resolution. Each decoder block contains two convolutional layers with kernel size $3 \times 3 \times 3$ followed

by batch normalization and ReLU activation with an up-sampling layer applied at the end of each block. The filter sizes in the decoder layers progressively decrease, following a sequence of (512, 256, 128, 64), with a final kernel size $1 \times 1 \times 1$ convolution layer at the end that uses a reduced filter size of 32 and performs the convolution operation only once. A soft-max function is applied at the end of each decoder to generate normalized probability maps.

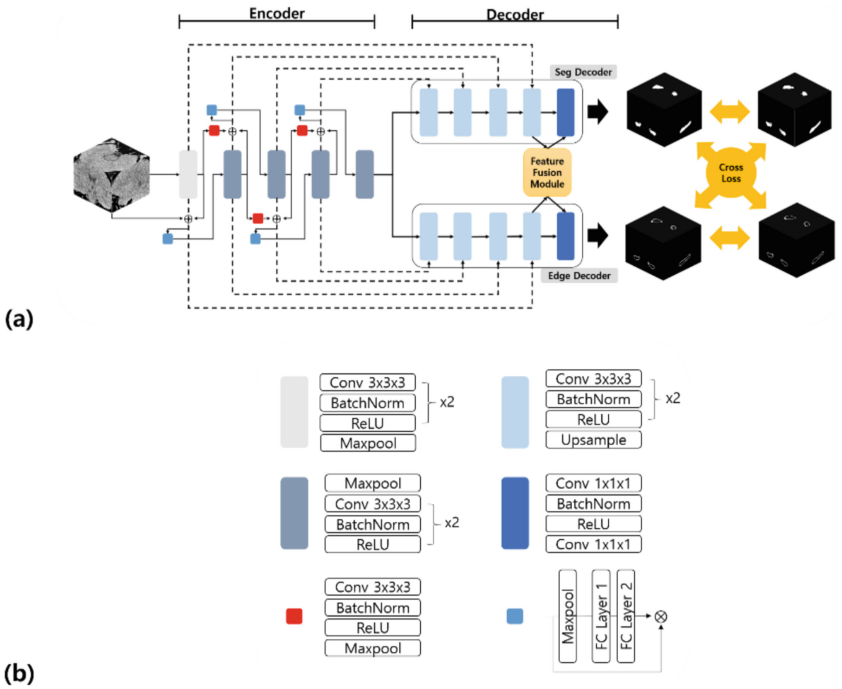


Fig. 1. The architecture of the proposed HDD-Net. (a) The network consists of dual decoders (Segmentation Decoder and Edge Decoder), a shared encoder, a feature fusion module (FFM), and a cross loss function. (b) Detailed structures of the convolutional blocks used in the encoder and decoders.

2.2 Shared Encoder

The shared encoder is responsible for extracting multi-scale spatial and contextual features from the input volumetric 3D CT images. It consists of five convolutional blocks with filter sizes set to (32, 64, 128, 256, 512).

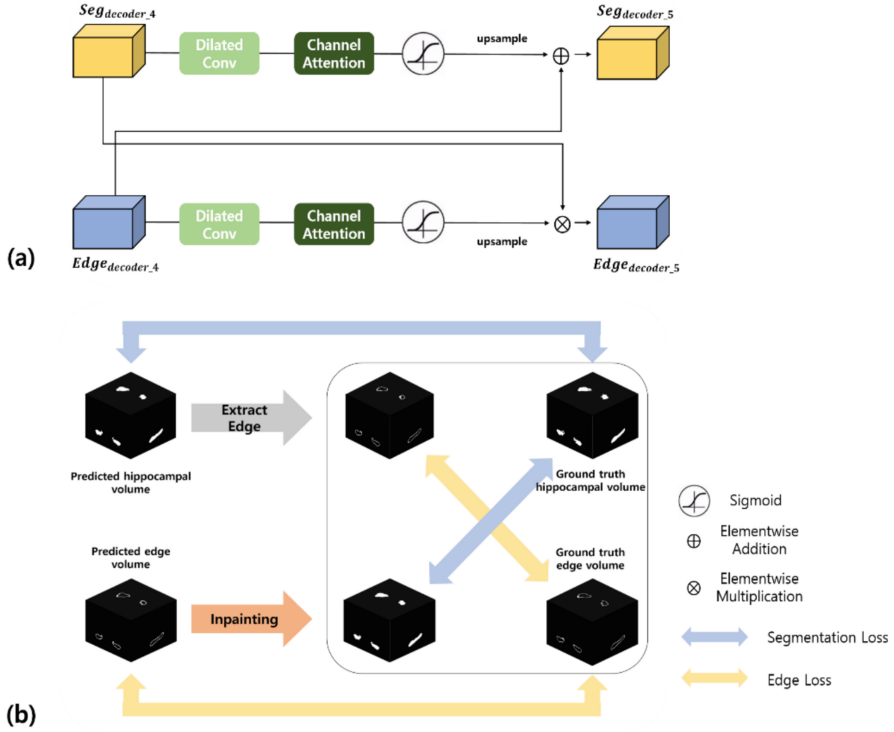


Fig. 2. (a) The feature fusion module (FFM) enables interaction between the Segmentation Decoder and Edge Decoder through dilated convolution, channel attention, and feature integration via elementwise operations. (b) The cross loss jointly optimizes volumetric and edge predictions by aligning them with the respective ground truths using segmentation and edge loss calculations.

The first block begins by processing the input using an initialization block (Init), which applies two $3 \times 3 \times 3$ convolution layers, each followed by batch normalization, ReLU activation, and max-pooling layer, produced the first-level feature map (E_1). From the second block onward, each block (O_{conv}) concatenates features from the outputs of the two preceding blocks after being downsampled (O_{reduce}) to match dimensions. A channel attention (O_{CA}) [15] is used to emphasize the most relevant channels in the concatenated features, improving feature quality. These refined features are then passed through max-pooling layer followed by two convolution layers with batch normalization and ReLU activation. Additionally, skip connections from each block to corresponding decoder blocks help retain spatial information during reconstruction.

Formally, the process can be described as:

$$E_1 = \text{Init}(\text{Input})$$

$$E_2 = O_{\text{conv}}(O_{\text{CA}}(\text{Concat}(\text{Input}, E_1)))$$

$$E_3 = O_{\text{conv}}(O_{\text{CA}}(\text{Concat}(O_{\text{reduce}}(E_1), E_2)))$$

$$E_4 = O_{conv}(O_{CA}(Concat(O_{reduce}(E_2),E_3)))$$

$$E_5 = O_{conv}(O_{CA}(Concat(O_{reduce}(E_3),E_4)))$$

where $E_i (i = 1, 2, 3, 4, 5)$ represents the feature map from each i th blocks.

2.3 Feature Fusion Module (FFM)

FFM enables interaction between the Segmentation Decoder and the Edge Decoder, as illustrated in Fig. 2(a). By combining the features extracted by the two decoders, the FFM provides a comprehensive representation that incorporates both volumetric and edge-specific information, leading to more precise segmentation results.

The fusion process begins with a dilated convolution [16] with kernel size $3 \times 3 \times 3$ (dilation $= 1, 3, 5, 7$), which expands the receptive field to capture multi-scale contextual information. This is followed by a channel attention [15] that dynamically recalibrates feature maps to prioritize the most relevant channels. To seamlessly integrate features between the decoders, the FFM performs targeted operations that enhance both volumetric and edge predictions. Specifically, the feature map from the Segmentation Decoder is combined with the FFM-processed edge feature map through elementwise multiplication, enabling the model to enhance edge predictions by leveraging volumetric features. Simultaneously, the feature map from the Edge Decoder is combined with the FFM-processed segmentation feature map using elementwise addition, allowing the model to refine volumetric predictions by incorporating detailed edge information.

This bidirectional flow in information ensures effective collaboration between the decoders, enabling model to capture both global structural details and fine-grained edges with high precision.

2.4 Cross Loss

The cross loss, depicted in Fig. 2(b), is a custom-designed function that jointly optimize the outputs of the Segmentation Decoder and the Edge Decoder, ensuring accurate prediction of both hippocampal volumes and edges. By integrating segmentation loss (L_{seg}) and edge loss (L_{edge}), the cross loss enables the model to focus on complementary aspects of segmentation accuracy. This dual optimization is achieved through a combination of dice loss and cross-entropy loss ($L_{dice\&CE}$), which together enhance both overlap-based and pixel-wise classification performance.

The segmentation loss measures the difference between the predicted hippocampal volume (Y_{P_V}) and the ground truth hippocampal volume (Y_{GT_V}). Additionally, it evaluates the consistency of in-painted predicted edge volume ($Y_{P_{E \to P_{V'}}}$) with the ground truth hippocampal volume (Y_{GT_V}), ensuring alignment between volumetric and boundary features:

$$L_{seg} = L_{dice\&CE}\left(Y_{P_V}, Y_{gt_V}\right) + L_{dice\&CE}\left(Y_{P_{E \to P_{V'}}}, Y_{gt_V}\right) \quad (1)$$

The edge loss ensures the consistency of predicted edge volumes (Y_{P_E}) with the ground truth edge volume $\left(Y_{GT_E}\right)$ and evaluates edges extracted from predicted volumes ($Y_{P_V \to P_{E'}}$):

$$L_{edge} = L_{dice\&CE}\left(Y_{P_E}, Y_{gt_E}\right) + L_{dice\&CE}\left(Y_{P_V \to P_{E'}}, Y_{gt_E}\right) \tag{2}$$

Combining Eq. (1) and Eq. (2), the cross loss (L_{cross}) is defined as:

$$L_{cross} = \alpha L_{seg} + \beta L_{edge} \tag{3}$$

where α and β are weighting factors. In this work, we set $\alpha = 1$ and $\beta = 1$.

3 Experiments and Results

3.1 Experiments

Datasets. We collected datasets from Gangnam Severance Hospital (GSH) and Seoul St. Mary's Hospital (SSMH) with approval of both institutional review boards. Informed consent from the patients was waived due to the study's retrospective nature. A total of 150 neurologically healthy individuals from GSH who underwent both brain CT and T1-weighted MRI [17] within three months were included, meeting specific imaging criteria. The GSH dataset was split into a training set (n = 120) and internal test set (n = 30), while an external validation set (n = 47) from SSMH was selected using the same criteria.

Data Preprocessing. CT and MRI images were converted from DICOM to NIFTI for efficient handling and processing of 3D images. CT images were registered to MRI using SPM12 [18], normalized to HU range (−20 to 100), and min-max scaled to [0, 1], MR images were processed with Freesurfer [19], yielding hippocampal masks as ground truth, from which edges were extracted. Finally, co-registered CT-MRI pairs along with ground-truth labels were center-cropped from (256, 256, 256) to (96, 128, 128).

Implementation Details. The proposed HDD-Net model implemented in this work comprises 100,195,112 learnable parameters, which were randomly initialized at the start of training. Hyperparameters we used during the training phase included a batch size of 4, a total of 150 epochs, and a learning rate of 0.0001 with the Adam optimizer. The network was implemented and trained using PyTorch 2.3.1 on two NVIDIA GeForce RTX 3090 Ti GPUs, each equipped with 24GB of memory. The implementation of HDD-Net is available at https://github.com/sonwonjun103/HDD_Net.

Evaluation. The performance of the proposed model was evaluated using two standard metrics: Dice similarity coefficient (DSC) and Intersection over Union (IoU). Additionally, Bland-Altman analysis was conducted to assess volumetric agreement between MRI-based ground truth and CT-based segmentation results.

3.2 Results

Qualitative Analysis. Figure 3 illustrates representative segmentation results for test subject. The first column shows CT image, while the second column shows the corresponding MR image. The final column shows the overlaid the ground truth hippocampal segmentation with the prediction map generated by HDD-Net. Additionally, the 3D renderings further validate the model's accuracy, as the predicted hippocampal segmentation maps closely resemble the hippocampus derived from MR images. This consistency across modalities emphasizes the models' capability to address the challenges of hippocampus segmentation in non-contrast CT images, producing results comparable to MRI-based ground truths.

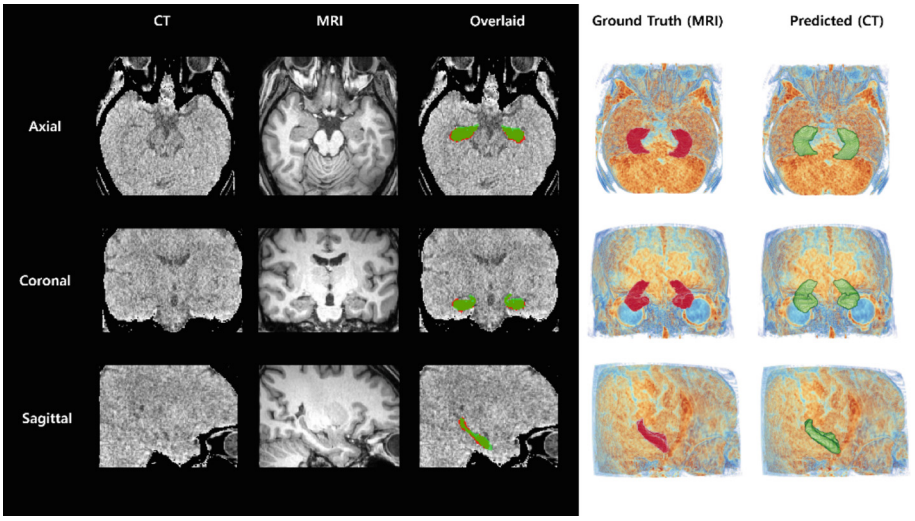


Fig. 3. Visualization of hippocampus segmentation results. The left section shows axial, coronal, and sagittal views of CT images, MRI ground truth, and overlaid segmentation results. (Red: ground truth volume, Green: overlaid volume, Light green: predicted volume). The right section provides 3D renderings of hippocampal volumes from the MRI ground truth and the predicted volume by the proposed model. (Color figure online)

Evaluation Scores. On internal datasets, the model achieved a mean DSC of 0.823 ± 0.03 and IoU of 0.701 ± 0.05. For external datasets, the model achieved a mean DSC of 0.759 ± 0.07 and IoU of 0.617 ± 0.09. To evaluate the performance of HDD-Net, a comparative study was conducted against widely-used U-Net based models in medical segmentation, summarized in Table 1. These models include the standard U-Net, it enhanced version such as Attention U-Net [21] and nnU-Net [22], as well as models integrating transformer blocks like 3D TransU-Net [6], Swin U-NetR [23], and U-NetR [5]. Additionally, the comparison included the AG-3D ResNet proposed by Portal et al. [13]. Our model outperformed these models across both internal and external datasets.

Table 1. Performance comparison of HDD-Net with U-Net based segmentation models.

Models	Internal (n = 30)		External (n = 47)	
	DSC	IoU	DSC	IoU
U-Net [20]	0.784 ± 0.04	0.647 ± 0.06	0.718 ± 0.07	0.565 ± 0.08
Attention U-Net [21]	0.775 ±0.04	0.635 ± 0.06	0.694 ± 0.08	0.538 ± 0.09
nnUNet [22]	0.769 ± 0.03	0.626 ±0.05	0.588 ± 0.06	0.419 ± 0.06
TransUNet [6]	0.773 ±0.03	0.632 ± 0.05	0.589 ± 0.05	0.420 ±0.05
UNETR [5]	0.656 ± 0.06	0.491 ± 0.06	0.508 ± 0.13	0.351 ± 0.11
Swin UNETR [23]	0.650 ± 0.07	0.485 ± 0.07	0.558 ± 0.14	0.400 ± 0.13
AG-3D ResNet [13]	0.762 ± 0.06	0.620 ± 0.07	0.673 ± 0.13	0.516 ± 0.11
HDD-Net (Ours)	**0.823 ± 0.03**	**0.701 ± 0.05**	**0.759 ± 0.07**	**0.617 ± 0.09**

Bland-Altman Analysis. To evaluate the agreement between MRI-based and DL-based CT segmentations, bland-altman analysis conducted for internal and external datasets. For the overall volumetric differences, it showed a mean difference 0.53 cm^3 /1.17 cm^3 for internal and external datasets, respectively (Fig. 4(a) and Fig. 4(b)), indicating good agreement between the two hippocampal volumes.

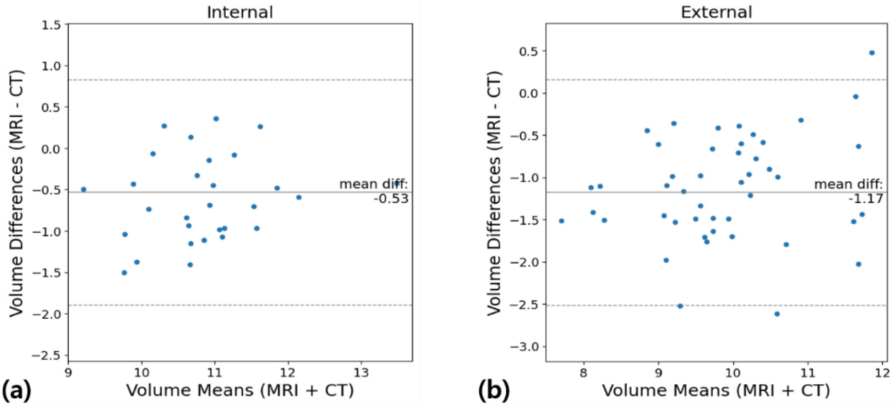


Fig. 4. Bland-Altman plots illustrating the differences in hippocampal volumetric measurements between MRI-based segmentation and DL-based segmentation on CT scans for (a) internal datasets (n = 30) and (b) external datasets (n = 47).

4 Discussion and Conclusions

HDD-Net is a novel dual-decoder deep learning model designed for hippocampal segmentation from CT images, addressing the limitations of MRI dependency in clinical applications. By integrating a FFM for bidirectional feature exchange and a cross loss for

joint optimization, HDD-Net achieves superior segmentation performance. The model outperforms existing U-Net and transformer-based methods on both internal and external datasets, demonstrating its effectiveness in CT-based hippocampal segmentation. Notably, this method enables hippocampal volumetric analysis in environments where MRI is limited or unavailable, broadening its clinical applicability.

In conclusion, HDD-Net presents a promising approach to hippocampal segmentation from CT, expanding accessibility to hippocampal volumetric analysis in MRI-limited environments. By eliminating the need for MRI in segmentation tasks, this method could enhance clinical workflows in Alzheimer's disease assessment and hippocampal avoidance whole-brain radiotherapy.

Acknowledgments. This work was supported by the IITP-Innovative Human Resource Development for Local Intellectualization program grant (IITP-2025-RS-2022-00156389, 50%), and the Commercialization Promotion Agency for R&D Outcomes (COMPA) grant (RS-2023-00304695, 50%), both funded by the Korean Government (Ministry of Science and ICT).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Dubois, B., et al.: Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. Lancet Neurol. **6**, 734–746 (2007)
2. Bonne, O., et al.: Longitudinal MRI study of hippocampal volume in trauma survivors with PTSD. Am. J. Psychiatry **158**, 1248–1251 (2001)
3. Bremner, J.D., Narayan, M., Anderson, E.R., Staib, L.H., Miller, H.L., Charney, D.S.: Hippocampal volume reduction in major depression. Am. J. Psychiatry **157**, 115–118 (2000)
4. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) MICCAI 2015. Part III, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
5. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
6. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Son, W.J., Ahn, S.J., Lee, J.Y., Lee, H.: Automated brain segmentation on computed tomographic images using perceptual loss based convolutional neural networks. Invest. Magn. Reson. Imaging **28**, 193–201 (2024)
8. Kim, J., Lee, H., Oh, S.S., Jang, J., Lee, H.: Automated quantification of total cerebral blood flow from phase-contrast MRI and deep learning. J. Imaging Inf. Med. **37**, 563–574 (2024)
9. Ataloglou, D., Dimou, A., Zarpalas, D., Daras, P.: Fast and precise hippocampus segmentation through deep convolutional neural network ensembles and transfer learning. Neuroinformatics **17**, 563–582 (2019)
10. Zeng, D., Li, Q., Ma, B., Li, S.: Hippocampus segmentation for preterm and aging brains using 3D densely connected fully convolutional networks. IEEE Access **8**, 97032–97044 (2020)
11. Chen, X., Peng, Y., Li, D., Sun, J.: DMCA-GAN: dual multilevel constrained attention GAN for MRI-based hippocampus segmentation. J. Digit. Imaging **36**, 2532–2553 (2023)

12. Xiao, Z., Zhang, Y., Deng, Z., Liu, F.: Light3DHS: a lightweight 3D hippocampus segmentation method using multiscale convolution attention and vision transformer. Neuroimage **292**, 120608 (2024)
13. Porter, E., et al.: Hippocampus segmentation on noncontrast CT using deep learning. Med. Phys. **47**, 2950–2961 (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2015)
15. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
16. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
17. Mugler, J.P., III., Brookeman, J.R.: Three-dimensional magnetization-prepared rapid gradient-echo imaging (3D MP RAGE). Magn. Reson. Med. **15**, 152–157 (1990)
18. Flandin, G., Friston, K.J.: Statistical parametric mapping (SPM). Scholarpedia **3** (2008)
19. Fischl, B.: FreeSurfer. Neuroimage **62**, 774–781 (2012)
20. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G., Wells, W. (eds.) MICCAI 2016. Part II, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
21. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
22. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: NnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**, 203–211 (2021)
23. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) BrainLes 2021. LNCS, vol. 12962, pp. 272–284. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08999-2_22



DCT-Net: Dual-Branch CT Reconstruction from Orthogonal X-Rays with Diffusion Model and Contrastive Learning

Zhiyu Zhang[1], Cong Shen[1(✉)], Jijun Tang[2], and Zhijun Liao[3]

[1] Tianjin University of Technology, Tianjin, China
congshen@email.tjut.edu.cn
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong, China
[3] Fujian Medical University, Fuzhou, Fujian, China

Abstract. Computed tomography (CT) reconstruction from X-ray images possesses significant advantages, including lower radiation exposure, reduced costs, and better accessibility than direct CT imaging. However, insufficient effective input samples caused by data volume under the moderate level or occlusion of partial soft tissues by skeletal structures in X-rays often hold back achieving high-quality image reconstruction. Additionally, contrasted with voxel-level differences, the texture and structure features are significant for image reconstruction. In virtue of these challenges, this study proposes an efficient approach named **D**ual-branch **CT Net**work (DCT-Net). It first integrates a conditional diffusion model for data augmentation, which mitigates data scarcity and achieves bone suppression. Subsequently, a dual-branch network in DCT-Net is leveraged to parallel process both augmented and raw data. In the framework, a perceptual loss based on high-level semantic features performs as the contrastive loss. Furthermore, it combines the voxel-level and adversarial losses to optimize the generator. However, the discriminator optimization only depends on the adversarial loss. Experimental results on two public datasets demonstrate that DCT-Net outperforms the state-of-the-art works, appearing to have promising potential among clinical applications.

Keywords: Computed tomography · X-rays · Dual-branch network · 3D reconstruction · Diffusion model · Contrastive learning

1 Introduction

Computed tomography (CT) generates cross-sectional images from multi-angle X-ray projections, providing visualization of bones and soft tissues with higher

This work is sponsored in part by National Natural Science Foundation of China (No. 62106175, 62020106004 and 92048301).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 150–160, 2026.
https://doi.org/10.1007/978-3-032-04947-6_15

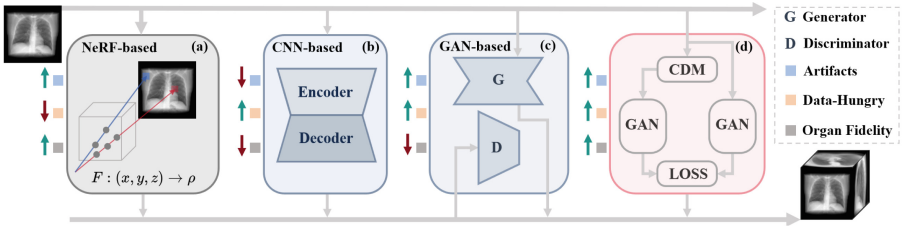


Fig. 1. Motivation for data augmentation in CT reconstruction. The arrows in different colors refer to the performance of the corresponding strategy. The red rounded rectangle refers to the data augmentation strategy in this study. (Color figure online)

resolution than X-ray images. However, considering its risky dose of radiation [1,2], low-dose scanning approaches could be viewed as the alternative manners [3–5], which still require hundreds of X-ray projections from a CT scanner. Recently, CT reconstruction from X-ray images has arisen, focusing on its effectiveness in cost-saving and low-dose imaging [6]. Generally, the encountered problems for CT reconstruction with X-rays could be categorized into the sparse view [7–11] and limited sample size. The prior usually adopts neural radiance fields (NeRF) to construct a continuous 3D volume representation. The latter generally employs a convolutional neural network (CNN) and generative adversarial network (GAN) to learn the mapping from 2D to 3D, as shown in Fig. 1. Neither (a), (b), nor (c) could handle the requirement for a large amount of input data, the reconstruction of organ details, and artifacts.

For the CNN-based methods in Fig. 1(b), they extract image features through multiple convolutional layers [12], and learn the mapping from X-rays to CT [13–16]. However, the local receptive fields of convolutional operations also limit the capturing of global 3D structural information, leading to artifacts or anatomical structures missing. Contrastively, NeRF employs a continuous function [9,17], which could produce relatively high-resolution results by leveraging X-rays taken from different angles [8,18]. NeRF estimates density values based on the light propagation in space and the interactions among surrounding tissues. Nevertheless, NeRF-based methods are data-hungry and highly rely on the amount of multi-view X-rays.

GAN utilizes adversarial training, which adopts a generator and a discriminator to compete against each other to produce more realistic images [19–24]. Additionally, GAN could effectively capture high-dimensional data distributions, making it well-suited for CT generation [20]. However, it often focuses on voxel-level restoration but ignores high-level semantic information. Moreover, due to the occlusion of soft tissues by the skeletal structure, the reconstruction of organ details could be viewed as a challenge.

Inspired by GAN and contrastive learning, this study proposes a refined framework named DCT-Net, which integrates a pre-trained conditional diffusion model (CDM) and contrastive loss as Fig. 1(d). Initially, the CDM based on the orthogonal X-rays is employed to generate bone-suppressed X-rays, which

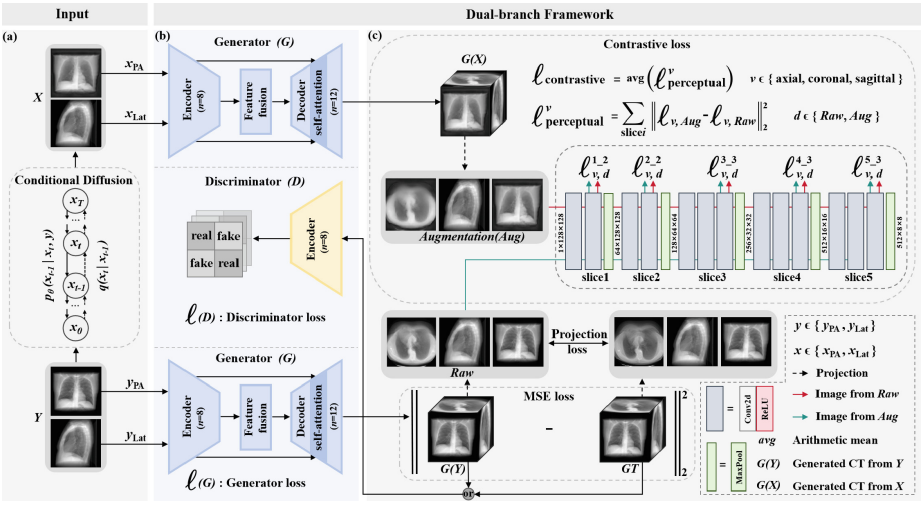


Fig. 2. The schematic of the DCT-Net framework. PA denotes the posterior-anterior view; Lat denotes the lateral view. (a) Generate bone-suppressed X-ray images from a pre-trained conditional diffusion model conditioned on input X-ray images. (b) Dual-generator reconstruction GAN processes the X-ray images. (c) Multi-path loss module.

reflect the positions and features of the internal organs [25]. Then, a dual-branch framework is designed to play a role in measuring the effect of data augmentation. Subsequently, to leverage the high-level semantic information of structure and texture embedded in the CT generated by the proposed dual-branch network, perceptual loss is applied as contrastive loss [26]. In addition, the final multi-path loss module also integrates reconstruction loss, projection loss, and adversarial loss to enhance the robustness and generalization ability. The main contributions of this work are listed below:

1. A conditional diffusion model is used in generating bone-suppressed X-rays to improve the visibility of the internal organ structures in CT and overcome data scarcity of input X-rays;
2. A dual-branch framework for CT reconstruction is proposed, which separately processes the original and bone-suppressed X-rays via a dual-generator reconstruction GAN, improving reconstruction accuracy and robustness;
3. A multi-path loss module is applied to balance the voxel-level details and high-level semantic features.

2 Method

The framework proposed in this study is named **D**ual-branch **CT Net**work (DCT-Net). As illustrated in Fig. 2, it leverages a pair of input orthogonal X-ray images $\langle y_{\mathrm{PA}}, y_{\mathrm{Lat}} \rangle$ to generate CT images.

2.1 Bone Suppression Conditional Diffusion Model

Due to the data scarcity of input X-rays and the occlusion of internal organs caused by bones, a CDM is employed for data augmentation, enabling the model to focus on the texture, structure, and location of internal organs. Given that x and y are the images in the image set X and Y, respectively. Then, DCT-Net generates bone-suppressed X-rays $X = \{x_{\mathrm{PA}}, x_{\mathrm{Lat}}\}$ from orthogonal X-rays $Y = \{y_{\mathrm{PA}}, y_{\mathrm{Lat}}\}$.

Diffusion Model. Diffusion models include forward and reverse processes [25, 27,28]. The forward process q gradually perturbs an initial data distribution $x_0 \sim q(x_0)$ using a predefined Markov chain, progressively adding Gaussian noise over T steps until $x_T \sim \mathcal{N}(0, \mathbf{I})$, which shown as (1):

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

where β_t denotes the noise scheduling coefficient at timestep t, controlling the noise injection rate; $\mathbf{I}$ denotes the identity matrix. The reverse process p_θ aims to recover the original data x_0 from the noise x_T, which is formulated as (2):

$$p_\theta\left(x_{t-1} \mid x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \Sigma_\theta\left(x_t, t\right)\right) \tag{2}$$

where μ_θ and Σ_θ denote the mean and covariance derived from a U-Net-based noise prediction network ϵ_θ parameterized by θ [29].

Conditional Diffusion Model. As the bone-suppressed X-ray x is generated according to the corresponding X-ray y, the reverse process is formulated as $p_\theta(x_{0:T} \mid y)$, with the original forward process q preserved. The conditional reverse transition could be (3):

$$p_\theta(x_{t-1} \mid x_t, y) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, y, t), \Sigma_\theta(x_t, y, t)\right) \tag{3}$$

During the training process, pairs $(x_0, y) \sim q(x_0, y)$ are sampled, and ϵ_θ learns to predict noise conditioned on y, where x_0 denotes the ground truth bone-suppressed X-ray. The joint distribution could be formulated as (4):

$$p_\theta(x_{0:T} \mid y) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, y) \tag{4}$$

2.2 Dual-Generator Reconstruction GAN

The GAN in DCT-Net uses two generators to process X and Y, of which the discriminator optimizes one of the generators that processes Y, shown as Fig. 2(b). Specifically, $\mathbf{p}_{\mathrm{PA}} \in \{x_{\mathrm{PA}}, y_{\mathrm{PA}}\}$ and $\mathbf{p}_{\mathrm{Lat}} \in \{x_{\mathrm{Lat}}, y_{\mathrm{Lat}}\}$. $G(X)$ and $G(Y)$ denote the generated CT from X and Y, respectively.

Generator. To address the limitation of capturing the latent relationship between two views with two encoders that increases the complexity of model [21–23], each generator in DCT-Net is based on an encoder-decoder, involving a feature fusion module γ and self-attention. Ying et al. proposed a skip connection

to adaptively transmit the low-level features during the encoding process [22]. To this end, the fused features of orthogonal X-rays could be defined as (5):

$$\mathbf{f}_{fusion} = \gamma(\mathbf{f}_1, \mathbf{f}_2) = \mathrm{FC}([\mathrm{EC}(\mathbf{p}_{\mathrm{PA}}), \mathrm{EC}(\mathbf{p}_{\mathrm{Lat}})]) \tag{5}$$

where $\mathrm{FC}(\cdot)$ and $\mathrm{EC}(\cdot)$ denote fully connected layer and encoder, respectively. $[\mathrm{EC}(\mathbf{p}_{\mathrm{PA}}), \mathrm{EC}(\mathbf{p}_{\mathrm{Lat}})]$ represents the concatenation of the features from $\mathbf{p}_{\mathrm{PA}}$ and $\mathbf{p}_{\mathrm{Lat}}$. For the decoder, self-attention is incorporated with each layer to enhance the capture of long-range dependencies in the reconstructed CT features, as shown in (6):

$$\mathrm{DC}_l = \mathrm{SA}\left(\mathrm{Deconv}\left(\mathrm{DC}_{l-1}\right)\right), \quad l = 1, 2, \ldots, L \tag{6}$$

where $\mathrm{Deconv}(\cdot)$ denotes deconvolution; $\mathrm{SA}(\cdot)$ represents self-attention; and DC_l denotes the output of the l-th decoding layer. The generated CT denotes as DC_L.

Discriminator. The discriminator in GAN outputs a scalar to determine the fidelity of an input sample [30]. This setting provides a unified judgment for the image without considering the significant bias on average for local areas. Hence, the output of the discriminator is constructed as (7), allowing the generator to obtain fine-grained feedback:

$$D(\cdot|Y) = \{d_{j,v}\}_{m \times n} \tag{7}$$

where '$\cdot$' denotes $G(Y)$ or the CT ground truth GT; $v \in \{axial, coronal, sagittal\}$; D represents the discriminator; m and n denote the patches of one CT horizontal and vertical slice, respectively; $d_{j,v}$ represents the probability of sub-region in j-th slice of $G(Y)$ from the sample distribution, as viewed from v.

2.3 Multi-path Loss Module

In DCT-Net, a multi-path loss module is designed to integrate perceptual loss, voxel-level loss, and adversarial loss.

Contrastive Loss. Regarding $G(X)$ is expected to suppress skeletal information, rather than enforcing a strict pixel-wise match with $G(Y)$, DCT-Net introduces the perceptual loss as the contrastive loss to measure high-level semantic and perceptual differences between them.

Specifically, $G(Y)$ and $G(X)$ are projected onto standard anatomical planes to obtain the *Raw* and *Aug* images, respectively. Suppose that $\phi_i^v(z)$ is the ReLU activation of the i-th layer in a network ϕ for an image z under view v. When i corresponds to a convolutional layer, $\phi_i^v(z)$ is a feature map of shape $C_i \times H_i \times W_i$. The perceptual loss between *Raw* and *Aug* on the same plane is calculated as (8):

$$\ell^v(Raw, Aug) = \sum_{i=1}^{N_{\mathrm{s}}} \frac{1}{C_i H_i W_i} \left\| \phi_i^v(Raw) - \phi_i^v(Aug) \right\|_2^2 \tag{8}$$

where N_S denotes the number of slices in network ϕ. $\phi_i^v(Raw)$ and $\phi_i^v(Aug)$ represent the activation values in i-th layer of ϕ, extracted from Raw and Aug on the projection plane with the same view v. Contrastive loss ℓ_{contr} is computed as the arithmetic mean of the perceptual loss across the three planes.

Voxel-Level Loss. It includes reconstruction loss ℓ_{recon} and projection loss ℓ_{proj}, ensuring voxel-wise consistency not only in 3D space but also across projected 2D slices. Reconstruction loss employs mean squared error (MSE) to minimize voxel-wise differences. Projection loss enforces consistency by computing the average L1 loss across axial, coronal, and sagittal planes.

Adversarial Loss. Following the LSGAN [31], the adversarial loss of DCT-Net is defined as (9):

$$\begin{aligned}\ell(D) = \frac{1}{2}\Big[&\mathbb{E}_{GT\sim p(\mathrm{CT})}(D(GT \mid Y) - 1)^2 \\ &+ \mathbb{E}_{Y\sim p(\mathrm{Xray})}(D(G(Y) \mid Y) - 0)^2\Big]\end{aligned} \tag{9a}$$

$$\ell(G) = \frac{1}{2}\mathbb{E}_{Y\sim p(\mathrm{Xray})}\left[(D(G(Y) \mid Y) - 1)^2\right] \tag{9b}$$

where $GT \sim p(\mathrm{CT})$ represents the real CT samples, and $Y \sim p(\mathrm{Xray})$ represents the input X-rays samples. Discriminator loss $\ell(D)$ comprises two terms to encourage it to classify GT and penalize it for classifying $G(Y)$. Generator loss function $\ell(G)$ encourages the discriminator to classify $G(Y)$ with high fidelity.

Overall Training Objective. It is defined as (10):

$$D^* = \arg\min_D \lambda_1 \ell(D) \tag{10a}$$

$$G^* = \arg\min_G \left[\lambda_1 \ell(G) + \lambda_2 \ell_{contr} + \lambda_3 \ell_{recon} + \lambda_4 \ell_{proj}\right] \tag{10b}$$

where λ_1, λ_2, λ_3, and λ_4 are hyper-parameters to adjust the weight of loss terms.

3 Experiments

3.1 Datasets and Experimental Settings

LIDC-IDRI Dataset. The dataset contains 1,018 lung CT samples [32]. In this study, 854 samples are used for training, 72 for validation, and 92 for testing.

CTspine1K Dataset. 784 CT images [33] are selected from this dataset and divided into 650 training samples, 60 validation samples, and 74 testing samples.

Implementations and Training Settings. In the experiments, the input X-rays synthesized by the digitally reconstructed radiographs technology are resized to 128×128 [34], and the output CT is resized to $128 \times 128 \times 128$. The implementation is based on the PyTorch framework with Adam optimizer, where the initial learning rate is set to 0.0002 and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$. Furthermore, due to the constraints of GPU memory, the batch size

Table 1. Experimental results and ablation study on LIDC-IDRI dataset.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)
PerX2CT [37]	27.248 ± 0.004	0.637 ± 0.005	0.253 ± 0.003
X2CT [22]	26.769 ± 0.040	0.622 ± 0.001	0.307 ± 0.002
RT-SRTS [38]	25.311 ± 0.010	0.570 ± 0.001	0.323 ± 0.003
X-Recon [23]	27.120 ± 0.001	0.619 ± 0.002	0.292 ± 0.004
SdCT-GAN [21]	27.054 ± 0.010	0.596 ± 0.003	0.322 ± 0.020
w/o Dual-branch	27.226 ± 0.030	0.620 ± 0.005	0.285 ± 0.001
w/o Recon-GAN	27.562 ± 0.001	0.648 ± 0.008	0.251 ± 0.004
w/o $\ell_{perceptual}$	27.602 ± 0.003	0.651 ± 0.003	0.283 ± 0.001
DCT-Net (ours)	$\mathbf{27.928 \pm 0.040}$	$\mathbf{0.703 \pm 0.001}$	$\mathbf{0.221 \pm 0.004}$

Table 2. Experimental results on CTspine1K dataset.

Method	PSNR (↑)	SSIM (↑)	LPIPS (↓)	Params (M)
PerX2CT [37]	25.575 ± 0.010	0.614 ± 0.001	0.282 ± 0.003	71.59
X2CT [22]	24.927 ± 0.001	0.616 ± 0.008	0.334 ± 0.001	73.15
RT-SRTS [38]	23.518 ± 0.020	0.523 ± 0.002	0.348 ± 0.005	546.26
X-Recon [23]	24.715 ± 0.010	0.594 ± 0.003	0.330 ± 0.001	101.74
SdCT-GAN [21]	25.124 ± 0.040	0.541 ± 0.003	0.336 ± 0.002	77.32
DCT-Net (ours)	$\mathbf{26.600 \pm 0.003}$	$\mathbf{0.691 \pm 0.010}$	$\mathbf{0.234 \pm 0.030}$	**60.70**

is set as 1. The training process is carried out for 100 epochs to ensure stable convergence based on the validation set. All the experiments are conducted on a workstation with an NVIDIA GeForce RTX 3090 Ti 24 GB GPU card.

Evaluation Metric. Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) are adopted in this work [35]. Moreover, learned perceptual image patch similarity (LPIPS) is introduced to focus on latent semantic perception [36].

3.2 Comparison with State-of-the-Art Methods

Table 1 presents the comparison results on the LIDC-IDRI dataset. DCT-Net outperforms other methods in PSNR, SSIM, and LPIPS. Compared with the current SOTA method PerX2CT, DCT-Net achieves improvements of +0.68 in PSNR, +0.066 in SSIM, and −0.032 in LPIPS. In particular, PerX2CT adopts slice-wise processing in CT reconstruction [37], ignoring cross-slice continuity required in 3D structural consistency. To evaluate the stability and generalization of DCT-Net and baseline methods, experiments on the CTspine1K pose more reconstruction challenges due to the lesions in the dataset. As shown in Table 2, DCT-Net achieves the best performance while maintaining the most

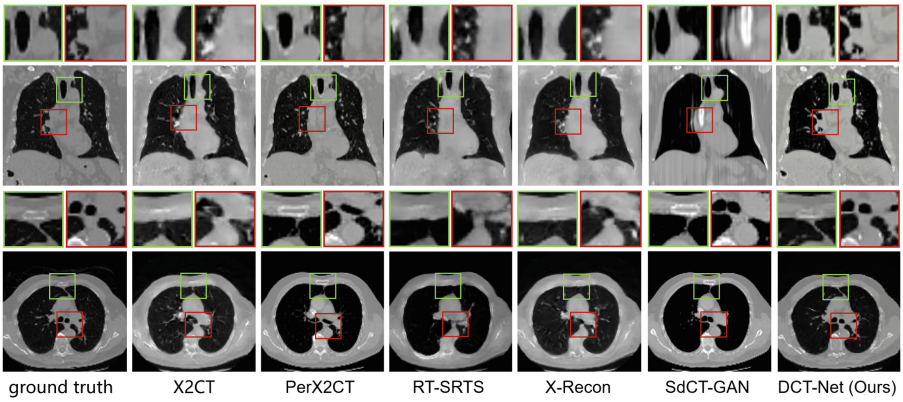


Fig. 3. Visual comparison of CT slices. The second and the last rows are the global CT slices, while the first and the third rows are the local details of global CT slices.

stable metrics with merely −1.328 PSNR and +0.013 LPIPS variation. Furthermore, DCT-Net requires the fewest parameters with only 60.7M, demonstrating higher computational efficiency.

Visualization results are shown in Fig. 3 where DCT-Net demonstrates superior reconstruction performance, particularly in preserving anatomical details such as lung structures, cardiac contours, and specific tracheal regions. This outcome is consistent with the results presented in Tables 1 and 2.

3.3 Ablation Study

To verify the effectiveness of the dual-branch network, dual-generator reconstruction GAN (Recon-GAN), and perceptual loss in DCT-Net, the ablation results are presented in Table 1.

Firstly, the bone suppression CDM failed without the dual-branch network component. While PSNR and SSIM decrease by 0.702 and 0.083, LPIPS increases by 0.064 compared with the combination. It indicates that bone-suppressed X-rays could conduct data cleaning for input data, reducing the interference of bone artifacts. Meanwhile, the dual-branch network could leverage the complementary information from the output CTs of the Recon-GAN. Secondly, Recon-GAN is replaced with the GAN proposed by Ying et al. [22]. Results show that it declines across all three evaluation metrics, verifying the significance of DCT-Net improvements to the generator and discriminator. Lastly, perceptual loss replaced with projection loss is conducted. LPIPS increases by 0.062 in Table 1, indicating that perceptual loss tends to capture high-level semantic information while avoiding optimizing only with voxel-level differences.

4 Conclusion

In this study, a network named DCT-Net is proposed to achieve CT reconstruction from orthogonal X-rays. A conditional diffusion model is firstly employed to generate bone-suppressed X-rays, overcoming data scarcity and the occlusion of internal organs by bones. Then, the augmented and original X-rays are fed into a dual-generator reconstruction GAN, which incorporates a feature fusion module and self-attention. Next, the perceptual loss is selected as the contrastive loss in DCT-Net, focusing on restoring information about internal organs. Experimental results on the LIDC-IDRI and CTspine1K datasets demonstrate that DCT-Net significantly outperforms existing benchmark methods. Future work will focus on data augmentation using self-supervised and semi-supervised methods for cases lacking samples.

Acknowledgements. The authors gratefully acknowledge the reviewers for their valuable comments and suggestions. We would also like to thank our colleagues.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Power, S., Moloney, F., Twomey, M., James, K., O'Connor, O., Maher, M.: Computed tomography and patient risk: facts, perceptions and uncertainties. World J. Radiol. **8**(12), 902–915 (2016)
2. Schmidt, C.: CT Scans: Balancing health risks and medical benefits. Environ. Health Perspect. **120**, A118–A121 (2012)
3. Zhang, Y., et al.: Side information-assisted low-dose CT reconstruction. IEEE Trans. Comput. Imaging **10**, 1080–1093 (2024)
4. Zhang, W., Huang, B., Chen, S., Xu, X., Wu, W., Liu, Q.: Low-rank angular prior guided multi-diffusion model for few-shot low-dose CT reconstruction. IEEE Trans. Comput. Imaging **10**, 1763–1774 (2024)
5. Jiang, J., Feng, Y., Xu, H., Zheng, J.: Low-dose CT reconstruction via optimization-inspired GAN. In: ICASSP, pp. 1–5 (2023)
6. Garg, Y., Seetharam, K., Sharma, M., Rohita, D.K., Nabi, W.: Role of deep learning in computed tomography. Cureus **15**(5), e39160 (2023)
7. Ding, C., Zhang, Q., Wang, G., Ye, X., Chen, Y.: Learned alternating minimization algorithm for dual-domain sparse-view CT reconstruction. In: MICCAI, vol. 14229, pp. 173–183. Springer, Cham (2023)
8. Cai, Y., Wang, J., Yuille, A., Zhou, Z., Wang, A.: Structure-aware sparse-view X-ray 3D reconstruction. In: CVPR, pp. 11174–11183 (2024)
9. Zha, R., Zhang, Y., Li, H.: NAF: neural attenuation fields for sparse-view CBCT reconstruction. In: MICCAI, vol. 13436, pp. 442–452. Springer, Cham (2022)
10. Cheng, Y., et al.: LIR-Net: learnable iterative reconstruction network for fan beam CT sparse-view reconstruction. IEEE Trans. Comput. Imaging **10**, 181–195 (2024)
11. Ayad, I., Larue, N., Nguyen, M.K.: QN-Mixer: a quasi-newton MLP-mixer model for sparse-view CT reconstruction. In: CVPR, pp. 25317–25326 (2024)

12. Shen, C., Wang, X., Tang, J., Liao, Z.: DETA-Net: a dual encoder network with text-guided attention mechanism for skin-lesions segmentation. In: ICIC, pp. 28–40 (2023)
13. Chen, H., et al.: Low-dose CT via convolutional neural network. Biomed. Opt. Express **8**(2), 679–694 (2017)
14. Jin, K., McCann, M., Froustey, E., Unser, M.: Deep CNN for inverse problems in imaging. IEEE Trans. Image Process. **26**(9), 4509–4522 (2017)
15. Dai, J., et al.: Volumetric tumor tracking from a single cone-beam X-ray projection image enabled by deep learning. Med. Image Anal. **91**, 102998 (2024)
16. Shao, H., et al.: Real-time liver tumor localization via a single X-ray projection using deep graph neural network-assisted biomechanical modeling. Phys. Med. Biol. **67**(11) (2022)
17. Rückert, D., Wang, Y., Li, R., Idoughi, R., Heidrich, W.: NeAT: neural adaptive tomography. ACM Trans. Graph. **41**(4) (2022)
18. Corona-Figueroa, A., Frawley, J., Taylor, S., Bethapudi, S., Shum, H., Willcocks, C.: MedNeRF: medical neural radiance fields for reconstructing 3D-aware CT-projections from a single X-ray. In: EMBC, pp. 3843–3848 (2022)
19. Yang, Q., et al.: Low-dose CT image denoising using GAN with Wasserstein distance and perceptual loss. IEEE Trans. Med. Imaging **37**(6), 1348–1357 (2018)
20. Cheng, S., Hong, Y., Chen, Q., Guo, J., Li, M., Zhang, Q.: CLCT-GAN: strong-weak contrastive learning for reconstructing CT images from radiographs. In: IJCNN, pp. 1–8 (2024)
21. Cheng, S., et al.: SdCT-GAN: reconstructing CT from biplanar X-rays with self-driven GAN. arXiv preprint arXiv:2309.04960 (2023)
22. Ying, X., Guo, H., Ma, K., Wu, J., Weng, Z., Zheng, Y.: X2CT-GAN: reconstructing CT from biplanar X-rays with GAN. In: CVPR, pp. 10611–10620 (2019)
23. Wang, Y., Wang, K., Zhuo, Y., Shi, W., Shan, F., Liu, L.: X-Recon: learning-based patient-specific high-resolution CT reconstruction from orthogonal X-ray images. arXiv preprint arXiv:2407.15356 (2024)
24. Zeng, P., et al.: Multi-modal long-short distance attention-based transformer-GAN for PET reconstruction with auxiliary MRI. IEEE Trans. Circuits Syst. Video Technol., 1–14 (2025). https://doi.org/10.1109/TCSVT.2025.3545911
25. Chen, Z., et al.: BS-Diff: effective bone suppression using conditional diffusion models from chest X-ray images. In: ISBI, pp. 1–5 (2024)
26. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
27. Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: WDM: 3D wavelet diffusion models for high-resolution medical image synthesis. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Mehrof, D., Yuan, Y. (eds.) DGM4MICCAI 2024. LNCS, vol. 15224, pp. 11–21. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-72744-3_2
28. Cui, J., et al.: MCAD: multi-modal conditioned adversarial diffusion model for high-quality PET image reconstruction. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15007, pp. 467–477. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72104-5_45
29. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

30. Liu, W., Ding, H.: Solving low-dose CT reconstruction via GAN with local coherence. In: MICCAI. vol. 14229. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43999-5_50
31. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: ICCV, pp. 2813–2821 (2017)
32. Armato, S.G.I., McLennan, G., Bidaut, L., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
33. Deng, Y., Wang, C., Hui, Y., et al.: CTSpine1K: a large-scale dataset for spinal vertebrae segmentation in computed tomography. arXiv preprint arXiv:2105.14711 (2021)
34. Milickovic, N., Baltast, D., Giannouli, S., Lahanas, M., Zamboglou, N.: CT imaging based digitally reconstructed radiographs and their application in brachytherapy. Phys. Med. Biol. **45**(10), 2787–2800 (2000)
35. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
36. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR, pp. 586–595 (2018)
37. Kyung, D., Jo, K., Choo, J., Lee, J., Choi, E.: Perspective projection-based 3D CT reconstruction from biplanar X-rays. In: ICASSP, pp. 1–5. IEEE (2023)
38. Zhu, M., et al.: RT-SRTS: angle-agnostic real-time simultaneous 3D reconstruction and tumor segmentation from single X-ray projection. Comput. Biol. Med. **173**, 108390 (2024)

EFMS-Net: Efficient Frequency-Enhanced Multi-scale Network for Ischemic Stroke Segmentation

Jie Yang[1], Shaowei Shen[2], Xuwei Fan[3], Ning Chen[4], Zhibin Gao[5], Lianfen Huang[2(✉)], and Yihong Zhan[6(✉)]

[1] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
[2] School of Informatics, Xiamen University, Xiamen, China
lfhuang@xmu.edu.cn
[3] College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou, China
[4] College of Control Science and Engineering, China University of Petroleum (East China), Qingdao, China
[5] Navigation Institute, Jimei University, Xiamen, China
[6] The First Affiliated Hospital of Xiamen University, School of Medicine, Xiamen University, Xiamen, China
yihongzhan31@163.com

Abstract. The combination of multi-modal medical imaging for ischemic stroke infarct segmentation is crucial for clinical treatment. However, existing methods often improve segmentation accuracy at the cost of efficiency, rendering them impractical for mobile health applications. To overcome this limitation, we integrate Mamba, a state-space model for long-sequence modeling, with convolutional operations to capture both global and local dependencies. To further enhance the feature representation, we incorporate multi-scale feature interaction and frequency-domain processing. As a result, we propose a novel Efficient Frequency-enhanced Multi-Scale Network (EFMS-Net) to achieve an optimal trade-off between segmentation accuracy, inference speed, and parameter efficiency. Extensive experiments on four datasets demonstrate the effectiveness and efficiency of EFMS-Net. We release a new dataset to promote further research in ischemic stroke infarct segmentation. The dataset is available on GitHub.

Keywords: Ischemic Stroke Infarct Segmentation · Efficiency · Mamba · Feature Interaction · Frequency-domain Processing

1 Introduction

Ischemic stroke occurs when blood supply to a specific area of the brain is partially or completely obstructed, resulting in high mortality and morbidity. Its

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 161–171, 2026.
https://doi.org/10.1007/978-3-032-04947-6_16

diagnosis and treatment heavily depend on timely intervention and neuroimaging techniques, particularly Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) [1]. Employing image segmentation techniques to identify and segment the infarct core can assist clinicians in making timely and accurate treatment decisions.

In recent years, CNN- and Transformer-based models have dominated medical image segmentation. SwinUNETR [2] uses a Swin Transformer encoder [3] and a multi-scale decoder for efficient 3D segmentation, while MedNeXt [4] improves accuracy by adjusting kernel size, channels, and layers. However, challenges such as limited receptive fields and the high computational cost of long-range modeling remain, making these models impractical for resource-limited environments like mobile health applications. UNETR++ [5] addresses these with an Efficient Paired Attention (EPA) block to reduce computational burden. However, its performance on high-frequency features, such as infarct regions with blurred boundaries, remains suboptimal. Preserving high-frequency features in CT and MRI images is particularly challenging due to variations in signal intensity, texture, and shape, as well as the loss of details in deeper neural network layers. These challenges highlight the need for improved edge-feature representation. MEA-Net [6] enhances boundary information with Edge Feature Extraction (EFE), while DE-Net [7] improves boundary delineation using a specialized loss function. However, both models still struggle with the computational efficiency challenges addressed by UNETR++.

These limitations have driven researchers to explore alternative models that balance accuracy, inference speed, and computational efficiency. One promising approach is to employ Mamba [8], a State Space Sequence Model (SSM), to efficiently model long-range dependencies with linear complexity. Building on this idea, several studies [9–11] have integrated CNNs with Mamba, combining the former's ability to capture detailed local features with the latter's strength in modeling global dependencies. Another promising approach, instead, leverages the frequency domain to extract discriminative features and enhance semantic representation [12–14]. We propose an Efficient Frequency-enhanced Multi-Scale Network (EFMS-Net) based on the typical U-Net structure [15], which is well-known for its excellent performance in medical image segmentation, and further design an efficient encoder-decoder architecture built upon it. Two novel blocks, Dual-branch Adaptive Frequency Fusion (DAFF) block and Laplacian-enhanced Multi-Scale Attention (LMSA) block, are incorporated into the encoder and skip connections of EFMS-Net.

Our primary contributions are as follows: (1) To improve segmentation in regions with fuzzy boundaries and low contrast, the DAFF block extracts global and local features via two parallel subnetworks. The block then applies the Discrete Cosine Transform (DCT) to obtain frequency representations of these features and utilizes the Adaptive Frequency Weighting (AFW) layer to adaptively adjust the weights of all frequency components. (2) The LMSA block employs direction-aware attention across multiple scales to enhance channel-spatial interactions and utilizes the parameter-free Laplacian of Gaussian oper-

ator to improve edge representation. (3) A new dataset for 3D segmentation of ischemic stroke is proposed, comprising 120 MRI scans annotated by experts.

2 Method

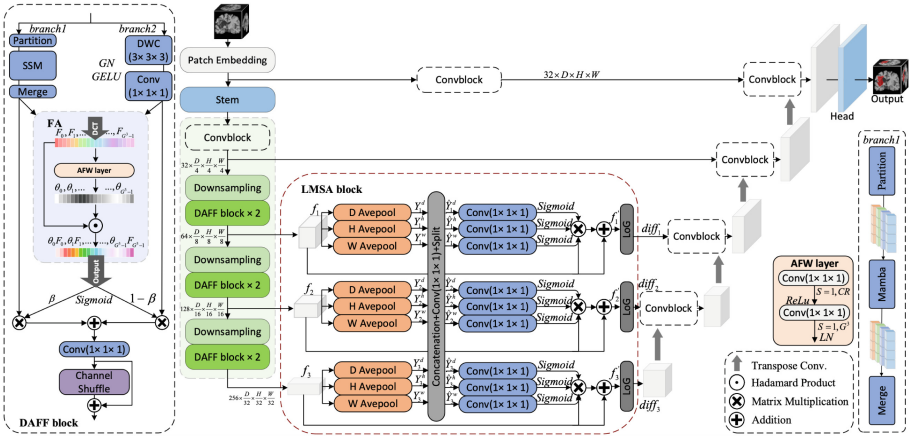


Fig. 1. The overview of the proposed EFMS-Net.

The overall architecture of EFMS-Net is illustrated in Fig. 1. The proposed model integrates DAFF and LMSA block to enhance multi-scale feature extraction and hierarchical information aggregation for stroke segmentation. The encoder begins with a stem layer that applies a $4 \times 4 \times 4$ convolution with a stride of 4 to obtain the initial feature map $f_0 \in \mathbb{R}^{C \times D/4 \times H/4 \times W/4}$ from the 3D input volume $I \in \mathbb{R}^{C \times D \times H \times W}$. This map is subsequently processed through three encoder stages, each consisting of two DAFF blocks and a downsampling layer. Additionally, the LMSA block are incorporated into the skip connections to enhance multi-scale feature interaction. The decoder gradually restores resolution through four stages by aggregating features from skip connections. Finally, transposed convolutions map the channels to segmentation targets, followed by a sigmoid activation to generate the binary segmentation mask.

Dual-Branch Adaptive Frequency Fusion Block: The DAFF block consists of two parallel branches. In the first branch, inspired by [16], we use cascaded $3 \times 3 \times 3$ depthwise convolutions followed by $1 \times 1 \times 1$ standard convolutions, with channel-wise GroupNorm and GELU activation applied between the convolutional layers. The second branch employs the bidirectional Mamba model [17], where the 3D feature maps are flattened into a sequence of 1D patch embeddings. These embeddings are then partitioned into s subsets with a step size of s to capture dependencies between adjacent sampling tokens. Finally, outputs

of the two branches, Mamba and $Conv_{casc}$, are converted into their frequency domain representation F_i using modified DCT [18], which is given by:

$$F_i = \sum_{d=0}^{N-1}\sum_{h=0}^{N-1}\sum_{w=0}^{N-1} x_{d,h,w} \cos\left(\frac{(2d+1)\mu\pi}{2N}\right)\cos\left(\frac{(2h+1)\nu\pi}{2N}\right)\cos\left(\frac{(2w+1)k\pi}{2N}\right) \quad \text{s.t. } i \in \{0,1,2,\ldots,G^3-1\} \tag{1}$$

where $x_{d,h,w}$ denotes the pixel intensity at position (d,h,w) within the 3D image block $x \in \mathbb{R}^{C\times N\times N\times N}$. The channel C is split into G^3 groups, and $\mu,\nu,k \in \{0,1,\cdots,G-1\}$ represent the frequency indices along the depth, height, and width dimensions of the block, respectively. When μ,ν and k in Formula 1 are zero, it's equivalent to Global Average Pooling (GAP), which captures only the lowest frequency component. The multi-spectral frequency vector F is given by:

$$F = cat([F_0, F_1, \cdots, F_{G^3-1}]) \tag{2}$$

The Frequency Attention (FA) mechanism is introduced to learn the importance of different frequency components. Its AFW layer is designed to generate learnable weights $\Theta \in \mathbb{R}^{G^3\times 1\times 1^3\times 1}$ for each component of F. Specifically, the AFW block employs two $1\times 1\times 1$ convolutional layers with ReLU and LayerNorm, whose convolutional weights are trainable parameters that adaptively adjust each frequency component. The output weights Θ are normalized along the channel dimension using a SoftMax function. F is then reshaped into the shape of $\mathbb{R}^{G^3\times \frac{C}{G^3}\times 1^3\times 1}$, where the weight of each group corresponds to a frequency component $F_i \in \mathbb{R}^{\frac{C}{G^3}\times 1^3\times 1}$. By performing element-wise multiplication between Θ and F, the layer re-weights frequency features and outputs a refined representation F', as follows:

$$\Theta = SoftMax(AFW(F)) \tag{3}$$

$$F' = \Theta F = cat(\theta_0 F_0, \theta_1 F_1, \ldots, \theta_{G^3-1} F_{G^3-1}) \tag{4}$$

where θ_i and F_i represent the i-th components of Θ and F, respectively. Instead of selecting only the top-k frequency components based on prior knowledge, as in FcaNet model [12], the proposed method captures a broader range of frequency components across the entire image block and dynamically reweights them.

The multi-frequency channel attention weights for global and local features are denoted as β and β', respectively. These weights adaptively modulate the features by multiplication, and the whole process can be formulated as follows:

$$\beta = Sigmoid\left(FA(Mamba(f)) + FA(Conv_{casc}(f))\right) \tag{5}$$

$$f' = \beta \times Mamba(f) + \beta' \times Conv_{casc}(f) \tag{6}$$

where $\beta, \beta' \in \mathbb{R}_{\geq 0}^{C\times 1\times 1\times 1}$ and $\beta + \beta' = 1$. The resulting feature maps f' are then fused through a $1 \times 1 \times 1$ convolution. Furthermore, a channel shuffle operation [19] is employed to enhance cross-channel communication among feature maps.

Laplacian-Enhanced Multi-scale Attention Block: To maintain the spatial consistency of multi-scale feature maps, LMSA block is applied to $\{f_l\}_{l=1}^{3}$, modeling the correlation between spatial and channel attention information. Specifically, LMSA block performs directional pooling along depth, height, and width dimensions to extract long-range dependencies while maintaining positional information in the orthogonal directions. The process is expressed as follows:

$$Y_l^d(d) = \frac{1}{H_l \times W_l} \sum_{0\leq i\leq H_l} \sum_{0\leq j\leq W_l} f_l(d_l, i, j) \tag{7}$$

$$Y_l^h(h) = \frac{1}{D_l \times W_l} \sum_{0\leq k\leq D_l} \sum_{0\leq j\leq W_l} f_l(k, h_l, j) \tag{8}$$

$$Y_l^w(w) = \frac{1}{D_l \times H_l} \sum_{0\leq k\leq D_l} \sum_{0\leq i\leq H_l} f_l(k, i, w_l) \tag{9}$$

where $Y_l^d \in \mathbb{R}^{C_l\times d_l\times 1\times 1}$, $Y_l^h \in \mathbb{R}^{C_l\times 1\times h_l\times 1}$ and $Y_l^w \in \mathbb{R}^{C_l\times 1\times 1\times w_l}$ represent the outputs of the l-th feature map along the depth, height, and width dimensions. Here, $d_l \in \{0, 1, ..., D_l - 1\}$, $h_l \in \{0, 1, ..., H_l - 1\}$, and $w_l \in \{0, 1, ..., W_l - 1\}$ are depth, height, and width indices. These outputs are adjusted to match the minimum channel dimension and then concatenated into $\mathbb{R}^{C_{min}\times(d_1+h_1+w_1+...+d_3+h_3+w_3)\times 1\times 1}$, followed by a shared $1 \times 1 \times 1$ convolution to facilitate layer-wise feature interaction. The resulting outputs are split along spatial dimensions back into $\hat{Y_l^d} \in \mathbb{R}^{C_l\times d_l\times 1\times 1}$, $\hat{Y_l^h} \in \mathbb{R}^{C_l\times 1\times h_l\times 1}$, and $\hat{Y_l^w} \in \mathbb{R}^{C_l\times 1\times 1\times w_l}$. Finally, a $1 \times 1 \times 1$ convolution followed by a sigmoid activation is applied to each $\hat{Y_l^d}$, $\hat{Y_l^h}$, and $\hat{Y_l^w}$ to get the corresponding attention weights. These weights are used to refine the original feature maps and a residual connection is introduced to preserve the original feature representations and mitigate overfitting. The process is expressed as follows:

$$f_l' = f_l + f_l \times \sigma(\text{Conv}(\hat{Y_l^d})) \times \sigma(\text{Conv}(\hat{Y_l^h})) \times \sigma(\text{Conv}(\hat{Y_l^w})) \tag{10}$$

The output f_l' is processed through Laplacian of Gaussian (LoG) [20], defined as $\text{LoG}(f) = \nabla^2(f * G)$, which smooths the features using a Gaussian filter to suppress high-frequency noise and then calculates the second-order derivative to capture edge information. However, directly computing the second-order derivative can be computationally expensive. To improve efficiency, we approximate the LoG by combining a Gaussian filter with $2\times$ upsampling and $2\times$ downsampling to extract high-frequency information. The formula is as follows:

$$diff_l = f_l - upsample(downsample(f_l * G)). \tag{11}$$

where G is a normalized 3D Gaussian kernel of size $5 \times 5 \times 5$, and $*$ denotes the convolution operation. Downsampling is performed by slicing the feature map to reduce its resolution, while upsampling restores the resolution by zero-padding and applying Gaussian convolution to enhance the feature details. Finally, the difference between the original and upsampled feature maps approximates the high-frequency components.

3 Experiments

3.1 Experimental Setup

We use three public datasets (ISLES'22 [21], ISLES'18 [22], ATLAS v2.0 [23]) and a self-constructed dataset for ischemic stroke lesion segmentation. The ISLES'22 dataset includes 400 multi-modal MRI scans (FLAIR, DWI, ADC) for pre- and post-operative cases. ISLES'18 has 156 records from 103 patients, featuring perfusion maps (CBF, CBV, MTT, Tmax) and CT images (excluding 4D CTP). ATLAS v2.0, one of the largest stroke imaging datasets, contains T1-weighted images from 1,271 patients. Our self-constructed dataset comprises DWI images from 120 patients, with in-plane dimensions ranging from 180×140 to 200×160, and 36 to 44 along the z-axis. All images were resampled to (3.0, 1.0, 1.0) mm voxel spacing and manually annotated by three clinicians using ITK-SNAP [24].

Table 1. Performance comparison of methods on the ISLES'18, ISLES'22, and ATLAS v2.0 datasets. The evaluation metrics are DSC (%), HD95 (mm), P (M), and FLOPS (G). Best values are in bold, and second-best values are underlined.

Methods	ISLES'18				ISLES'22				ATLAS v2.0			
	DSC	HD95	P	FLOPs	DSC	HD95	P	FLOPs	DSC	HD95	P	FLOPs
CNN-based												
nnUNet	60.5	10.1	29.9	84.9	79.4	10.9	30.8	566.0	61.1	23.0	31.2	479.0
UX-Net	55.6	15.0	51.1	300.5	80.0	9.9	51.1	3407.2	60.7	22.3	51.1	2884.5
MedNeXt	59.3	10.8	31.0	132.7	79.1	10.2	31.7	627.5	61.2	<u>20.1</u>	31.7	535.3
Transformer-based												
nnFormer	54.1	14.1	62.4	71.0	79.5	10.1	140.7	281.3	58.8	24.5	131.2	261.1
UNETR^{++}	58.6	10.5	31.1	<u>23.8</u>	80.3	9.9	46.5	**85.0**	59.1	23.1	42.6	**70.5**
Mamba-based												
U-Mamba	<u>61.4</u>	<u>8.3</u>	42.4	148.0	<u>80.4</u>	<u>8.7</u>	42.9	1165.7	<u>62.3</u>	23.3	42.8	990.6
LKM-UNet	56.5	11.1	62.7	187.8	78.4	11.1	64.0	1232.2	59.7	22.9	64.4	1047.0
LightM	53.6	12.4	<u>5.1</u>	40.3	78.7	9.1	<u>5.1</u>	<u>117.3</u>	59.3	23.6	<u>5.1</u>	<u>99.9</u>
Ours	**62.7**	**8.1**	**2.5**	**9.5**	**81.8**	**7.3**	**3.8**	224.6	**62.6**	**18.8**	**3.8**	187.8

Table 2. Memory and efficiency comparison of methods on ISLES'18, ISLES'22, and ATLAS v2.0 datasets. Evaluation metrics are TM (MiB), IM (MiB) and IT (case/s).

Methods	ISLES'18			ISLES'22			ATLAS v2.0		
	TM	IM	IT	TM	IM	IT	TM	IM	IT
nnUNet	1739	1169	**0.23**	8507	2279	1.62	6531	2085	2.92
UX-Net	3537	2499	0.44	32107	3667	11.73	23775	3105	21.74
MedNeXt	3001	2411	0.29	17341	2669	2.81	14761	2565	5.21
nnFormer	2764	2348	0.28	16645	2542	2.62	14217	2456	4.81
UNETR^{++}	2143	965	0.78	5629	1471	1.75	4699	**1359**	3.12
U-Mamba	3777	2333	0.46	30432	6283	7.34	22529	5459	13.65
LKM-UNet	8981	4841	0.86	40680	6973	10.13	35035	6573	18.72
LightM	7079	5289	0.54	33503	3817	12.95	29347	3584	24.45
Ours	**1401**	**917**	0.25	**5405**	**1403**	**1.59**	**4669**	1475	**2.91**

Table 3. Ablation study for different modules of EFMS-Net on the four datasets.

Modules	ISLES'18			ISLES'22			ATLAS v2.0			self-constructed		
	DSC	P	FLOPs	DSC	P	FLOPs	DSC	P	FLOPs	DSC	P	FLOPs
Baseline	58.1	4.7	9.9	78.5	12.2	227.4	58.5	12.1	190.1	37.3	4.6	55.7
DAFF only	60.6	2.4	9.3	80.8	3.7	224.4	61.1	3.7	187.7	40.4	2.4	52.4
LMSA w/o LoG	61.1	4.7	10.0	80.6	12.2	227.5	61.3	12.2	190.2	39.6	4.7	55.7
LMSA	61.5	4.7	10.1	80.9	12.2	227.6	61.8	12.2	190.3	40.2	4.7	55.8
DAFF + LMSA	62.7	2.5	9.5	81.8	3.8	224.6	62.6	3.8	187.8	40.9	2.5	52.5

For a fair comparison, all models were implemented in `nnUNetv2` [25] and trained with default settings for 1,000 epochs on the same preprocessed data. Performance was evaluated using 5-fold cross-validation. To reduce overfitting, data augmentation was applied. The proposed model, built with Monai, used a batch size of 2, a learning rate of 0.001, and the SGD optimizer, with a loss function combining soft dice and cross-entropy loss. Input patch sizes were $96 \times 160 \times 160$ (ISLES'22), $5 \times 224 \times 192$ (ISLES'18), $128 \times 128 \times 128$ (ATLAS v2.0), and $40 \times 192 \times 160$ (self-constructed dataset). All experiments ran on a single A100 80 GB GPU. Metrics included Dice Similarity Coefficient (DSC), 95th Percentile Hausdorff Distance (HD95), Number of Parameters (P), FLOPs, Training Memory (TM), Inference Memory (IM), and Inference Time (IT).

3.2 Experimental Results

Comparison with SOTA Methods: In this section, we compare our method with SOTA approaches on three public datasets. Specifically, we evaluate CNN-based methods [4,25,26], Transformer-based methods [5,27] and Mamba-based

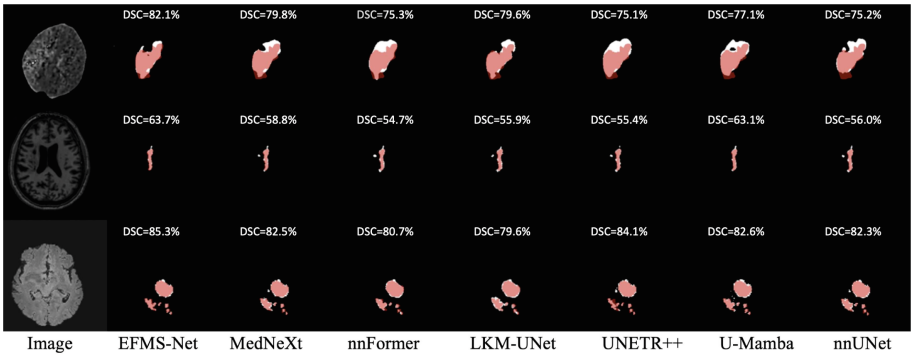


Fig. 2. Visualization of ischemic stroke infarct segmentation results. Row 1 shows ISLES'18 sample, Row 2 shows ATLAS v2.0 sample, and Row 3 shows ISLES'22 sample. Red regions indicate the ground-truth labels. (Color figure online)

methods [9–11], The experimental results are summarized in Table 1 and Table 2. EFMS-Net demonstrates the best segmentation accuracy, with an optimal balance across the number of parameters, memory usage, FLOPs, and inference speed on the three datasets. Although the segmentation accuracy of U-Mamba is only slightly behind EFMS-Net, its parameters, FLOPs, memory usage, and inference time are significantly worse. In the ISLES'18 dataset, EFMS-Net reduces parameters and computational complexity by 17.0× and 15.6×, respectively. In the ISLES'22 dataset, EFMS-Net uses 25,027 MiB and 4,880 MiB less memory in the training and inference stages and is about 4.6× faster than U-Mamba. While the FLOPs and inference memory of EFMS-Net are slightly higher than those of the lightweight model UNETR++ on ISLES'22 and ATLAS v2.0, EFMS-Net consistently demonstrates superior performance across all other metrics.

Visualization Results: We validate the effectiveness of the proposed method by visualizing the segmentation results on three datasets, comparing them with SOTA methods, as illustrated in Fig. 2. EFMS-Net demonstrates better alignment with ground-truth labels compared to SOTA methods. For cases with low contrast (Row 1) and small infarct cores (Row 3), it achieves more accurate delineation of infarct areas and shapes through effective integration of global and local dependencies and edge feature enhancement.

Ablation Study: We perform ablation studies on four datasets, including a self-constructed one, to explore the effects of different EFMS-Net modules, as shown in Table 3. Specifically, replacing the DAFF module with an equivalent number of Convblock modules shows that both DAFF and LMSA improve EFMS-Net's performance over the baseline, highlighting their effectiveness in capturing multi-scale features and enhancing edge details. The dual-branch design of depthwise separable convolution and Mamba greatly reduces computational cost compared to traditional convolutional operations, while FA further improves the accuracy

by adaptively weighting frequency components. LMSA outperforms DAFF by leveraging multi-scale channel-spatial interactions, and its LoG operator effectively preserves high-frequency details. Integrating these components, EFMS-Net achieves the best performance.

4 Conclusion

In this paper, we propose the Efficient Frequency-enhanced Multi-Scale Network (EFMS-Net), a novel approach to ischemic stroke infarct segmentation that strikes a balance between accuracy, inference speed, and computational efficiency. Specifically, we introduce the Dual-branch Adaptive Frequency Fusion (DAFF) block, which captures both global and local dependencies from frequency and spatial perspectives. Additionally, the Laplacian-enhanced Multi-Scale Attention (LMSA) block enables multi-scale interactive attention and employs the Laplacian of Gaussian (LoG) operator to enhance high-frequency details. Comparative experiments and ablation studies on four datasets demonstrate the superior performance of the proposed model in terms of segmentation accuracy and efficiency, especially in resource-constrained environments.

Acknowledgments. The work presented in this paper was partially supported by the 2024 National Natural Science Foundation of China (Grant No. 62371406) and Natural Science Foundation of Xiamen, China (Grant number 3502Z202473053).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Hui, C., Tadi, P., Suheb, M.Z., Patti, L.: Ischemic stroke. StatPearls [Internet] (2024)
2. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: SwinUNETR-V2: stronger swin transformers with stagewise convolutions for 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 416–426. Springer. Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_40
3. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
4. Roy, S., et al.: Mednext: Transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 405–415. Springer (2023)
5. Shaker, A.M., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: UNETR++: delving into efficient and accurate 3D medical image segmentation. IEEE Trans. Med. Imaging (2024)
6. Liu, H., et al.: MEA-Net: multilayer edge attention network for medical image segmentation. Sci. Rep. **12**(1), 7868 (2022)

7. Gu, R., Wang, L., Zhang, L.: DE-Net: a deep edge network with boundary information for automatic skin lesion segmentation. Neurocomputing **468**, 71–84 (2022)
8. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
9. Ma, J., Li, F., Wang, B.: U-Mamba: enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
10. Wang, J., Chen, J., Chen, D., Wu, J.: LKM-UNet: large kernel vision mamba UNet for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 360–370. Springer (2024)
11. Liao, W., Zhu, Y., Wang, X., Pan, C., Wang, Y., Ma, L.: LightM-UNet: mamba assists in lightweight UNet for medical image segmentation. arXiv preprint arXiv:2403.05246 (2024)
12. Qin, Z., Zhang, P., Wu, F., Li, X.: FcaNet: frequency channel attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 783–792 (2021)
13. She, D., Zhang, Y., Zhang, Z., Li, H., Yan, Z., Sun, X.: EoFormer: edge-oriented transformer for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 333–343. Springer (2023)
14. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: MedSegDiff-V2: diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 6, pp. 6030–6038 (2024)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. Springer (2015)
16. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
17. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)
18. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE Trans. Comput. **100**(1), 90–93 (2006)
19. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
20. Marr, D., Hildreth, E.: Theory of edge detection. Proc. Roy. Soc. London Ser. B Biol. Sci. **207**(1167), 187–217 (1980)
21. Hernandez Petzsche, M.R., et al.: ISLES 2022: a multi-center magnetic resonance imaging stroke lesion segmentation dataset. Sci. Data **9**(1), 762 (2022)
22. The ISLES Challenge 2018 website. https://www.isles-challenge.org/ISLES2018/
23. Liew, S.L., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Sci. Data **9**(1), 320 (2022)
24. Yushkevich, P.A., Gao, Y., Gerig, G.: ITK-SNAP: an interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3342–3345 (2016)
25. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)

26. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3D UX-Net: a large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076 (2022)
27. Zhou, H.Y., et al.: nnFormer: volumetric medical image segmentation via a 3D transformer. IEEE Trans. Image Process. **32**, 4036–4045 (2023)



Faster, Self-supervised Super-Resolution for Anisotropic Multi-view MRI Using a Sparse Coordinate Loss

Maja Schlereth[1(✉)], Moritz Schillinger[1], and Katharina Breininger[1,2]

[1] Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
maja.schlereth@fau.de

[2] Center for AI and Data Science (CAIDAS), Julius-Maximilians-Universität Würzburg, Würzburg, Germany

Abstract. Acquiring images in high resolution is often a challenging task. Especially in the medical sector, image quality has to be balanced with acquisition time and patient comfort. To strike a compromise between scan time and quality for Magnetic Resonance (MR) imaging, two anisotropic scans with different low-resolution (LR) orientations can be acquired. Typically, LR scans are analyzed individually by radiologists, which is time consuming and can lead to inaccurate interpretation. To tackle this, we propose *tripleSR*, a novel approach for fusing two orthogonal anisotropic LR MR images, to reconstruct anatomical details in a unified representation. Our multi-view neural network is trained in a self-supervised manner, without requiring corresponding high-resolution (HR) data. To optimize the model, we introduce a sparse coordinate-based loss, enabling the integration of LR images with arbitrary scaling. We evaluate our method on MR images from two independent cohorts. Our results demonstrate comparable or even improved super-resolution (SR) performance compared to state-of-the-art (SOTA) self-supervised SR methods for different upsampling scales. By combining a patient-agnostic offline and a patient-specific online phase, we achieve a substantial speed-up of up to ten times for patient-specific reconstruction while achieving similar or better SR quality. Code is available at https://github.com/MajaSchle/tripleSR.

Keywords: MR imaging · Super-Resolution · Multi-View

1 Introduction

Magnetic Resonance (MR) imaging allows the 3-D assessment of bone and soft tissue anomalies without ionizing radiation. To increase patient comfort and minimize motion artifacts, it is important to keep image acquisition times as low as reasonably possible [12,20]. One important factor that impacts acquisition time is the inter-slice (out-of-plane) resolution. Acquiring multiple anisotropic

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 172–182, 2026.
https://doi.org/10.1007/978-3-032-04947-6_17

3-D images with a high in-plane resolution can effectively reduce and split up scanning time, especially in populations who struggle to remain still [15]. Multi-view anisotropic acquisitions are already used in rheumatology and abdominal imaging [13]. Conventionally, the individually acquired images then need to be assessed separately, which is time consuming and error prone as lesions might only be visible in one of the scans, potentially leading to misinterpretations. Multi-view super-resolution (SR) enables the generation of a single highresolution (HR) image given several low-resolution (LR) images for a simpler diagnostic process [3].

In this work, we introduce a novel self-supervised MR SR method that combines pretraining on a small, anisotropic dataset and per-patient adaptation during inference. Incorporating multi-patient information enables the use of cross-patient similarities while maintaining applicability to clinical scenarios with limited data availability. Accordingly, we split our optimization process into two phases: The "offline phase" which constitutes patient-agnostic feature extraction, and the "online phase" for additional patient-specific adaptation and subsequent inference to generate the final SR image. Minimizing runtimes for the online phase is critical for efficient deployment in production settings, such as hospitals, where fast image processing is essential. Specifically, we incorporate two orthogonal anisotropic LR MR scans to form a joint representation that facilitates the generation of accurate SR images. Following clinical routine, where typically no HR data is available, we train our network only on the available LR data using a fully self-supervised approach, circumventing the need for HR images as reference for training. By integrating a sparse coordinate-based loss function, we are able to 1) use images with varying resolution for training and 2) allow for arbitrary-scale upsampling. For evaluation, we perform extensive evaluations on two publicly available datasets [1,8,16] and demonstrate that our proposed method generalizes across datasets and to unseen MR sequences. We specifically highlight that our approach speeds up the SR process substantially compared to state-of-the-art (SOTA) self-supervised SR methods while preserving image quality.

1.1 Related Work

Several methods have been proposed to improve the resolution of MR image data [2,9,10,23]. They can be divided into single-image super-resolution (SISR) and multi-image super-resolution (MISR) methods [10,22]. MISR utilizes different (anisotropic) views or several sequentially acquired LR scans by merging them to generate a higher-resolution image. For MR SR specifically, multiple learning-based methods have been proposed that utilize matching LR and HR data as reference for training SR models. Lyu et al. [5] performed MR SR by matching features of the target LR contrast and the available reference HR image of a different domain. They used a transformer with a texture-preserving branch and contrastive learning to enhance the textural details of the SR image. To overcome the issue of generalization of SR approaches to different scales, Tan et al. [14] proposed an arbitrary scale SR approach for brain MR images. They

combined a Weight Prediction Network with SRGAN and achieved comparable performance to SOTA SR techniques. However, their approach relies on a large training dataset with HR images and they performed evaluation only on 2-D slices. Similar works propose SISR methods for MR or Computed Tomography (CT) SR but still rely on ground truth HR data [11,19]. Recently, implicit neural representations (INRs) have become more prominent in the context of image SR. In the work of Wu et al. [18], an INR for MR image SR was proposed. This method allows implicit modeling of a continuous function of sparsely sampled data points and projects the learned intensity distribution onto a more densely sampled grid. They achieved high reconstruction performance for several upsampling scales but still relied on paired isotropic LR and HR data. Distinct from the previously presented methods, other approaches allow for reference-free training of SR network while integrating arbitrary-scale upsampling. Zhao et al. propose a reference free SISR for MR image data [23]. McGinnis et al. [7] used two anisotropic LR MR images of different MR sequences to perform INR-based SR. While achieving very good SR reconstruction, their approach requires patient-specific training in the order of 10 to 20 min per image, which may not be clinically feasible.

2 Methodology

Our model, *tripleSR* for self-supervised SR, takes N paired LR images I^i_{ax} and I^i_{cor} from a training set $T = \{I^i_{\text{ax}} \in \mathbb{R}^{h \times w \times \frac{d}{e_i}}, I^i_{\text{cor}} \in \mathbb{R}^{h \times \frac{w}{e_i} \times d}\}_{i=1}^{N}$ as inputs, with e_i being the upsampling scale of the i-th LR-image pair. In Fig. 1 the complete training process is shown. Building on the methodology of [18], a convolutional encoder extracts feature maps $V^i_{\text{ax}} \in \mathbb{R}^{h \times w \times \frac{d}{e_i} \times 128}$ and $V^i_{\text{cor}} \in \mathbb{R}^{h \times \frac{w}{e_i} \times d \times 128}$ which are subsequently used for reconstruction. We adapt a residual dense network (RDN) [21], which is specifically developed for image SR, by removing the upsampling layer and applying 3-D convolution.

Each element in the feature map corresponds to a voxel in the respective LR image. In the HR image, coordinates may lie between those of the LR counterparts, along with the corresponding voxel data. To obtain a feature vector at a corresponding high-resolution coordinate x_{HR}, we perform trilinear interpolation on the feature maps V^i_{ax} and V^i_{cor} and generate two HR feature maps with the shape $h \times w \times d \times 128$ each. From this, we sample the feature vectors $v^i_{\text{HR(ax)}}$ and $v^i_{\text{HR(cor)}}$ at a specific HR coordinate x_{HR}. The two extracted feature vectors $v^i_{\text{HR(ax)}}$ and $v^i_{\text{HR(cor)}}$ corresponding to the HR coordinate are concatenated and fed to the decoder. The decoder consists of eight fully connected layers, each followed by ReLU activation with a residual connection after the fourth layer. Each inner layer has 512 input and output features. The decoder outputs the predicted voxel intensity $\hat{I}$ of the respective spatial coordinate x_{HR}.

We split our SR process into two phases. The first part is a training phase on a disjoint dataset which allows patient-agnostic feature extraction. We call this the "offline phase" as it only needs to be performed once. The second part

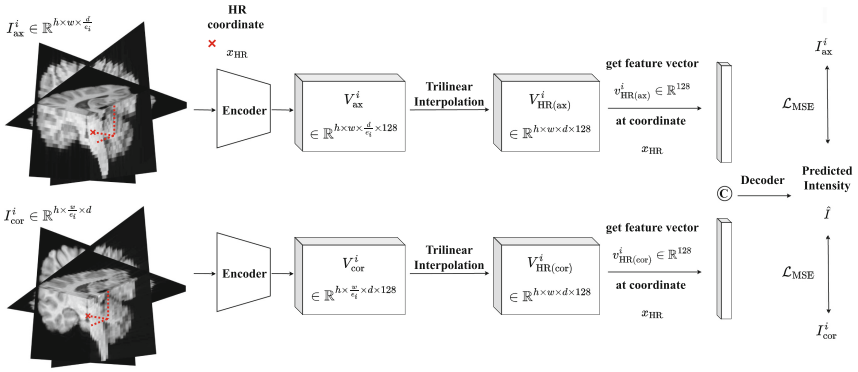


Fig. 1. Overview of the network architecture proposed in this study.

is an optional patient-specific fine-tuning and inference, which we refer to as the "online phase". Here, additional patient-specific details and anomalies can be learned. During both phases, no HR ground truth data is available. We optimize our model by using a loss function based on sparse continuous coordinates. Therefore, two orthogonal LR images of the same object are viewed as sparse representations of the same underlying instance. Instead of fixed voxel coordinates, we compute the loss for suitable locations in the continuous coordinate space where the coordinates are normalized to the range $[-1, 1]$. This enables the seamless SR of data with varying resolutions, as demonstrated in our experiments section. The applied loss computes the mean squared error ($\mathcal{L}_{\text{MSE}}$) of the voxel intensity of reference LR images and the voxel intensity of the predicted image $\hat{I}$ at matching coordinates M_{ax} and M_{cor}, where

$$\mathcal{M}_{\text{ax}} = \mathbf{C}_{\text{HR}} \cap \mathbf{C}_{\text{ax}} \text{ , and } \mathcal{M}_{\text{cor}} = \mathbf{C}_{\text{HR}} \cap \mathbf{C}_{\text{cor}}. \tag{1}$$

Here, $\cap$ represents the intersection of available normalized coordinates. Coordinates corresponding to regions without measurements in the LR image are excluded from the loss calculation.

3 Experiments

3.1 Datasets

To assess performance, we simulated LR images from HR images and compared the SR results to the corresponding HR. We emphasize that the HR data was used only as a ground truth for evaluation and was not used during model training. We used two publicly available datasets for our evaluation: The Brain Tumor Segmentation (BraTS) dataset [1,8], consisting of brain MR images with an isotropic resolution of $1 \times 1 \times 1\,\text{mm}^3$ and the HCP dataset [16], which includes 1200 MR images with an isotropic resolution of $0.7 \times 0.7 \times 0.7\,\text{mm}^3$. For each dataset and each MR sequence, we randomly selected 170 subjects (100 for training, 30 for validation, and 40 for final testing).

3.2 Preprocessing

For fair comparison, all images were registered using the NMI152 template with an isotropic resolution of 1 mm. Background areas were removed, and images were normalized to $[0, 1]$. To generate the anisotropic images in the *training dataset*, we randomly downsampled each image in both axial and coronal direction using a random downsampling scale in the range $[2, 4]$ once at the beginning of the training. We applied a commonly used downsampling method where HR images are downsampled by cropping in the frequency domain [4,6].

For evaluation, we generated two anisotropically sampled LR MR images in the axial and coronal dimension each with an in-slice resolution two and four times the between-slice resolution. As the LR images were simulated from one HR scan, they can be assumed to be rigidly registered. The axial images were resampled to a resolution of $1 \times 1 \times 2\,\text{mm}^3$ and $1 \times 1 \times 4\,\text{mm}^3$, while the coronal images were resampled to a resolution of $1 \times 2 \times 1\,\text{mm}^3$ and $1 \times 4 \times 1\,\text{mm}^3$. All LR image coordinates were normalized to a reference frame in the range $[-1, 1]$.

3.3 Implementation Details

During the offline phase, we trained our network for 35 epochs with a batch size B of 10 (lr $= 0.0001$, Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$), with hyperparameters tuned on the validation set. Training and evaluation were executed on an NVIDIA A40 with Python 3.9 and PyTorch 1.13.1. In each batch, B randomly cropped LR patches p of one image were selected, where B equals the training batch size. The axial and coronal image patches were of size $10 \times (e_i \times e_i \times 1)$ and $10 \times (e_i \times 1 \times e_i)$, respectively. To enable efficient batch processing for different upsampling scales, 8000 random samples are selected from each patch. Each sample corresponds to one HR voxel coordinate. For the online phase, we used the trained model from the offline phase. For the optional patient-specific training, the model undergoes 10 additional epochs on LR patches from a single patient before the final inference.

During the online phase, two different settings were evaluated. First, patient-specific online training was performed to update the model before subsequent inference. Second, inference without additional training was conducted, which is referred to "ours w/o FT". We quantitatively evaluated models using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [17], which assess pixel accuracy and perceptual quality. To ensure a fair comparison, we computed the metrics only within the brain region and excluded the background, as BISR is only trained on the brain area.

3.4 Reference Approaches

As we trained our models without HR ground truth data, we performed evaluations by comparing them with other unsupervised/self-supervised methods. We used cubic spline interpolation, a SISR method SMORE [23] and an INR MISR approach referred to as BISR [7]. Originally, [7] used two sequences as input

Table 1. Quantitative results for all MR sequences and SR methods on the BraTS and the HCP test set (trained and evaluated on the same MR sequence). Best results are bold, second best underlined. "Ours w/o FT" refers to results without online training.

		Resampling Scale	×2		×4	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
BraTS	T1 CE	cubic spline	33.794 ± 2.408	0.981 ± 0.006	29.93 ± 2.285	0.945 ± 0.01
		SMORE	37.939 ± 1.483	0.981 ± 0.003	30.678 ± 1.348	0.915 ± 0.009
		BISR	42.727 ± 2.576	0.993 ± 0.003	35.008 ± 2.183	0.963 ± 0.007
		Ours	**43.166 ± 2.105**	**0.994 ± 0.001**	**35.598 ± 1.813**	**0.966 ± 0.005**
		Ours w/o FT	41.358 ± 2.701	0.993 ± 0.002	35.531 ± 1.4655	0.964 ± 0.004
	T1	cubic spline	35.774 ± 3.745	0.991 ± 0.004	29.624 ± 3.629	0.962 ± 0.009
		SMORE	34.333 ± 2.64	0.982 ± 0.003	26.627 ± 1.817	0.906 ± 0.008
		BISR	**39.003 ± 3.984**	0.993 ± 0.006	31.078 ± 4.378	0.967 ± 0.017
		Ours	38.832 ± 4.121	**0.995 ± 0.002**	**31.169 ± 3.33**	**0.969 ± 0.006**
		Ours w/o FT	37.780 ± 4.061	0.993 ± 0.003	30.473 ± 3.632	0.965 ± 0.009
	T2	cubic spline	33.890 ± 1.876	0.986 ± 0.004	30.643 ± 2.168	0.959 ± 0.014
		SMORE	35.645 ± 1.388	0.984 ± 0.004	27.799 ± 1.217	0.91 ± 0.015
		BISR	**42.460 ± 1.853**	**0.997 ± 0.001**	**34.74 ± 1.661**	**0.976 ± 0.007**
		Ours	42.218 ± 1.750	**0.997 ± 0.001**	33.427 ± 1.640	0.973 ± 0.009
		Ours w/o FT	40.531 ± 2.190	0.995 ± 0.002	33.073 ± 1.412	0.969 ± 0.008
HCP	T1	cubic spline	29.820 ± 3.785	0.982 ± 0.006	22.341 ± 3.947	0.930 ± 0.023
		SMORE	30.603 ± 4.099	0.973 ± 0.007	22.533 ± 3.177	0.881 ± 0.019
		BISR	31.249 ± 4.574	0.985 ± 0.007	22.758 ± 4.636	0.931 ± 0.034
		Ours	**34.140 ± 5.410**	**0.990 ± 0.004**	**23.301 ± 5.274**	**0.938 ± 0.025**
		Ours w/o FT	29.950 ± 4.655	0.984 ± 0.007	22.523 ± 4.860	0.933 ± 0.026
	T2	cubic spline	30.561 ± 1.275	0.973 ± 0.004	27.336 ± 0.864	0.930 ± 0.008
		SMORE	31.444 ± 0.814	0.967 ± 0.004	25.817 ± 0.597	0.877 ± 0.010
		BISR	35.321 ± 0.858	0.985 ± 0.002	**29.862 ± 0.531**	**0.947 ± 0.004**
		Ours	**35.482 ± 0.877**	**0.986 ± 0.001**	29.372 ± 1.231	0.941 ± 0.006
		Ours w/o FT	34.841 ± 1.155	0.985 ± 0.002	29.681 ± 1.055	0.942 ± 0.006

and optimized the SR jointly. We adapted the approach to directly predict only one SR image instead of two images with different MR sequences. Cubic spline interpolation and the SISR method SMORE [23] are performed individually on the axial and coronal LR images of each patient. The coronal and axial metric values are then averaged.

4 Quantitative and Qualitative Results

We evaluated on 40 brain MR images each for T1, T1 contrast-enhanced (CE), and T2 sequences for BraTS and T1 and T2 for HCP which are trained either on images from the same sequence or from any of the other sequences (for 2× and 4×

Table 2. Duration of the offline and online phase for each method in minutes. Best, i.e., shortest times are highlighted in bold.

	Resampling Scale	×2		×4	
		offline	online	offline	online
BraTS T1	SMORE	0	9.09 ± 0.29	0	9.16 ± 0.32
	BISR	0	19.25 ± 1.32	0	10.59 ± 0.67
	Ours	130.99 ± 5.15	3.32 ± 0.31	130.99 ± 5.15	1.55 ± 0.15
	Ours w/o FT	130.99 ± 5.15	**1.94 ± 0.22**	130.99 ± 5.15	**0.73 ± 0.08**

Table 3. Quantitative results for different training settings with patient-specific online training. "Train on" refers to the used training data and "test on" refers to the specific test set. "Br" is used as abbreviation for BraTS.

Resampling Scale		×2		×4	
Train on	Test on	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Br T1 CE	Br T1 CE	43.166 ± 2.105	0.994 ± 0.001	35.598 ± 1.813	0.966 ± 0.005
Br T1	Br T1 CE	42.680 ± 2.435	0.994 ± 0.001	35.275 ± 1.818	0.963 ± 0.005
Br T2	Br T1 CE	42.762 ± 2.285	0.994 ± 0.001	35.476 ± 1.66	0.963 ± 0.005
Br T1	Br T1	38.832 ± 4.121	0.995 ± 0.002	31.169 ± 3.33	0.969 ± 0.006
Br T1 CE	Br T1	40.178 ± 4.388	0.995 ± 0.003	31.457 ± 3.382	0.970 ± 0.006
Br T2	Br T1	39.981 ± 4.259	0.995 ± 0.002	31.474 ± 3.245	0.969 ± 0.007
HCP T1	Br T1	39.410 ± 4.200	0.995 ± 0.002	31.450 ± 3.674	0.970 ± 0.006
Br T2	Br T2	42.218 ± 1.750	0.997 ± 0.001	33.427 ± 1.640	0.973 ± 0.009
Br T1 CE	Br T2	42.360 ± 1.962	0.997 ± 0.001	34.290 ± 1.359	0.976 ± 0.007
Br T1	Br T2	42.226 ± 1.958	0.997 ± 0.002	33.610 ± 1.467	0.973 ± 0.008
HCP T2	Br T2	42.456 ± 2.311	0.997 ± 0.002	33.548 ± 1.819	0.975 ± 0.009

upsampling). Table 1 summarizes the quantitative results for all SR approaches, including mean and standard deviation. We achieve higher or comparable PSNR and SSIM values on all MR sequences and upsampling scales. Using the additional patient-specific online training helps to improve performance in all cases. Still, without additional training (ours w/o FT), the performance is comparable to the best reference approach at approx. 10–15× faster inference times. Table 3 shows results obtained from testing on a dataset distinct from the one used for training. Even when the network is trained on a completely different dataset, the SR performance is comparable or even better compared to training on the same dataset. A reason for the better performance of the model trained on Br T1 CE for the BR T1 test data may be that the contrast-enhanced data potentially provides richer features and patterns, which allows the model to generalize to non-enhanced images.

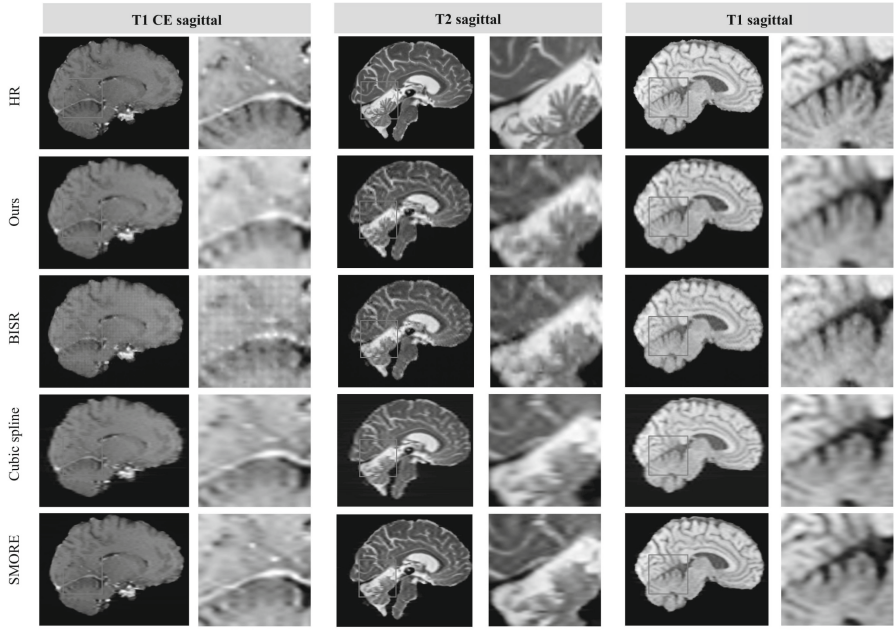


Fig. 2. Reference HR image and qualitative SR results for all MR sequences in the sagittal plane of the BraTS test set where no in-plane HR images are available.

The time required to generate a SR reconstruction for each method can be divided into offline and online phases. The results are shown in Table 2. Our model is pre-trained once offline which takes approximately 130 min. We average the time for the offline phase for three runs. During the online phase, our approach is approximately 10–15× faster when using the offline model w/o patient-specific fine-tuning and still 6× faster when fine-tuning for a specific patient, compared to the reference approach BISR.

In Fig. 2, qualitative examples of the BraTS dataset for each MR sequence and SR method can be seen in comparison to the actual HR image. The full image and a zoomed-in version can be seen for each sample. Overall, the SR results obtained using SMORE and cubic spline interpolation appear blurred or noisy. In some cases, BISR introduces noticeable blocking artifacts. In contrast, our method preserves accurate anatomical structures while minimizing noise.

5 Discussion and Conclusion

In this work, we presented a novel approach for MR image SR. In both quantitative and qualitative evaluations, our method produces HR images that match or surpass the quality of current reference approaches. A key strength lies in its substantially reduced patient-specific training time during the online phase, which is critical for translating such techniques into clinical practice. Notably,

after generating only 14 SR images with a resampling factor of 4 using BISR, the time invested in our offline phase is already offset. This efficiency gain enables a ten-fold increase in throughput for the generation of SR data. This could help to further drive acceptance of deep learning-supported diagnostic imaging in clinical workflows. Additionally, the sparse coordinate loss enables the integration of LR images with varying resolution scales, further enhancing the versatility and applicability of our approach. The patient-specific training allows strong adaptation and has lower risk of traditional overfitting due to the inherent inductive bias of the network.

Finally, the presented method demonstrates promising capabilities by consistently achieving high-quality results, regardless of whether training is performed on data from the same cohort or on a disjoint dataset. Further experiments are needed to validate our findings and demonstrate the benefits of using the reconstructed SR images in diagnostic assessments.

Acknowledgements. We gratefully acknowledge support by d.hip campus - Bavarian aim (Ma.S. and K.B.) and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117
2. Chen, Y., et al.: Brain MRI super resolution using 3D deep densely connected neural networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 739–742. IEEE, Piscataway, NJ (2018). https://doi.org/10.1109/ISBI.2018.8363679
3. Jia, Y., Gholipour, A., He, Z., Warfield, S.K.: A new sparse representation framework for reconstruction of an isotropic high spatial resolution MR volume from orthogonal anisotropic resolution scans. IEEE Trans. Med. Imaging **36**(5), 1182–1193 (2017). https://doi.org/10.1109/TMI.2017.2656907. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5534179/
4. Li, G., et al.: Rethinking multi-contrast MRI super-resolution: rectangle-window cross-attention transformer and arbitrary-scale upsampling. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21173–21183. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01941
5. Lyu, J., et al.: Multicontrast MRI super-resolution via transformer-empowered multiscale contextual matching and aggregation. IEEE Trans. Neural Netw. Learn. Syst. **PP** (2023). https://doi.org/10.1109/TNNLS.2023.3250491
6. Masutani, E.M., Bahrami, N., Hsiao, A.: Deep learning single-frame and multiframe super-resolution for cardiac MRI. Radiology **295**(3), 552–561 (2020). https://doi.org/10.1148/radiol.2020192173

7. McGinnis, J., et al.: Single-subject multi-contrast MRI super-resolution via implicit neural representations. In: Greenspan, H., et al. (eds.) MICCAI 2023, LNCS, vol. 14227, pp. 173–183. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43993-3_17
8. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
9. Nasrollahi, K., Moeslund, T.B.: Super-resolution: a comprehensive survey. Mach. Vis. Appl. **25**(6), 1423–1468 (2014). https://doi.org/10.1007/s00138-014-0623-4. https://link.springer.com/article/10.1007/s00138-014-0623-4#Sec3
10. Ooi, Y.K., Ibrahim, H.: Deep learning algorithms for single image super-resolution: a systematic review. Electronics **10**(7), 867 (2021). https://doi.org/10.3390/electronics10070867. https://www.mdpi.com/2079-9292/10/7/867#B6-electronics-10-00867
11. Peng, C., et al.: SAINT: spatially aware interpolation network for medical slice synthesis, pp. 7750–7759 (2020). https://openaccess.thecvf.com/content_CVPR_2020/html/Peng_SAINT_Spatially_Aware_Interpolation_NeTwork_for_Medical_Slice_Synthesis_CVPR_2020_paper.html
12. Riek, J.K., Tekalp, A.M., Smith, W.E., Kwok, E.: Out-of-plane motion compensation in multislice spin-echo MRI. IEEE Trans. Med. Imaging **14**(3), 464–470 (1995). https://doi.org/10.1109/42.414611
13. Shetty, A.S., et al.: Body MRI: imaging protocols, techniques, and lessons learned. Radiographics **42**(7), 2054–2074 (2022). https://doi.org/10.1148/rg.220025
14. Tan, C., Zhu, J., Lio', P.: Arbitrary scale super-resolution for brain MRI images. Artif. Intell. Appl. Innova. **583**, 165–176 (2020). https://doi.org/10.1007/978-3-030-49161-1_15. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7256400/
15. Thrower, S.L., et al.: The effect of slice thickness on contours of brain metastases for stereotactic radiosurgery. Adv. Radiat. Oncol. **6**(4), 100708 (2021). https://doi.org/10.1016/j.adro.2021.100708. https://pubmed.ncbi.nlm.nih.gov/34124413/
16. van Essen, D.C., et al.: The WU-Minn human connectome project: an overview. NeuroImage **80**, 62–79 (2013). https://doi.org/10.1016/j.neuroimage.2013.05.041. https://www.sciencedirect.com/science/article/pii/s1053811913005351
17. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/tip.2003.819861
18. Wu, Q., et al.: An arbitrary scale super-resolution approach for 3D MR images via implicit neural representation. IEEE J. Biomed. Health Inf. **27**(2), 1004–1015 (2023). https://doi.org/10.1109/JBHI.2022.3223106. https://arxiv.org/pdf/2110.14476
19. Yu, P., et al.: RPLHR-CT dataset and transformer baseline for volumetric super-resolution from CT scans. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 344–353. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_33
20. Zaitsev, M., Maclaren, J., Herbst, M.: Motion artifacts in MRI: a complex problem with many partial solutions. J. Magn. Reson. Imaging: JMRI **42**(4), 887–901 (2015). https://doi.org/10.1002/jmri.24850
21. Zhang, Y., et al.: Residual dense network for image super-resolution, pp. 2472–2481 (2018). https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_Residual_Dense_Network_CVPR_2018_paper.html

22. Zhao, C., et al.: Applications of a deep learning method for anti-aliasing and super-resolution in MRI. Magn. Reson. Imaging **64**, 132–141 (2019). https://doi.org/10.1016/j.mri.2019.05.038. https://www.sciencedirect.com/science/article/pii/S0730725X18306507
23. Zhao, C., et al.: SMORE: a self-supervised anti-aliasing and super-resolution algorithm for MRI using deep learning. IEEE Trans. Med. Imaging **40**(3), 805–817 (2021). https://doi.org/10.1109/TMI.2020.3037187

FDF-VQVAE: A Frequency Disentanglement and Fusion Learning Framework for Multi-sequence MRI Enhancement

Xinghe Xie[1], Luyi Han[2,3], Yue Sun[1], Chi Kin Lam[1], Jian Zheng[4], Tong Tong[5], Wei Ke[1], Chan-Tong Lam[1], and Tao Tan[1(✉)]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
taotan@mpu.edu.mo
[2] Radboud University Medical Centre, Nijmegen, Netherlands
[3] Department of Radiology, The Netherlands Cancer Institute, Amsterdam, Netherlands
[4] Department of Medical Imaging, Suzhou Institute of Biomedical Engineering, Chinese Academy of Sciences, Suzhou, China
[5] Fuzhou University, Fuzhou, China

Abstract. Multi-sequence magnetic resonance imaging (MRI) faces critical challenges in balancing accelerated acquisition and image quality: Rapid scanning typically induces degradation, including resolution reduction, increased noise, motion artifacts, and image blurring. While existing image enhancement models partially mitigate these issues, they often exhibit insufficient exploitation of complementary information across multi-sequence data. To address this issue, we propose an interpretable deep learning framework, FDF-VQVAE, for MRI image enhancement through frequency-domain feature disentanglement and fusion. Our framework constructs a dual-branch frequency-domain disentanglement module (DBFD) that precisely decouples high-frequency and low-frequency features of different sequences through parallel high-frequency feature and low-frequency feature extraction pathways. The multi-frequency-domain feature weighting mechanism (MFDFW) adaptively fuses the high and low-frequency features of different sequences. Finally, feature recombination and decoding achieve MRI enhancement through joint optimization. We conducted denoising, super-resolution, and deblurring experiments on the IXI dataset (546 subjects) with external validation on the BraTS2021 dataset (357 subjects). Experimental results demonstrate that our method significantly outperforms the state-of-the-art approaches in denoising, motion artifact removal, and super-resolution tasks. Our code is available at https://github.com/kkllxh/FDF-VQVAE.

Keywords: Multi-Sequence Brain MRI · Image Enhancement · Frequency-Disentanglement and Fusion · Interpretability

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 183–193, 2026.
https://doi.org/10.1007/978-3-032-04947-6_18

1 Introduction

Multi-sequence magnetic resonance imaging (MRI) has become an indispensable tool in clinical diagnosis, providing doctors with rich detailed tissue information [3,11–13], and is widely used in disease diagnosis [12], lesion segmentation [13], and treatment prognosis evaluation [2]. However, to improve scanning efficiency and save time, it is often necessary to reduce image resolution or increase the acceleration factor, which can lead to issues such as noise, motion artifacts, and image blurring [6]. To address these problems, deep learning-based MRI image enhancement techniques have gradually become an effective solution, attracting widespread attention.

In recent years, numerous deep learning-based methods have been applied to medical image enhancement [4,6,15,19]. However, these methods mainly focus on unimodal processing and fail to fully utilize the complementary information between multi-sequence MRI. Although some methods have attempted to leverage the complementary information across multi-sequence MRI, current approaches still have fundamental limitations. For instance, Dalmaz et al. [5] proposed a shared encoder architecture for multi-sequence feature extraction, but their framework inadequately decouples sequence-specific features, resulting in suboptimal feature discriminability. Furthermore, other methods [9,20] employing independent encoders for individual sequence feature extraction often struggle to effectively capture inter-sequence shared representations. The inability to concurrently learn sequence-specific distinctiveness and cross-sequence commonality significantly constrains their practical efficacy. To improve MRI image enhancement performance, it is crucial to effectively decouple and fuse this complementary information.

MRI images can be decoupled into high-frequency and low-frequency features [14,18]. The low-frequency features contain shared information about anatomical structures (such as organ morphology and tissue distribution), while the high-frequency features capture complementary information (such as texture details and edge features). Therefore, we propose an interpretable deep learning framework that can adaptively decouple and fuse high-frequency and low-frequency features from different MRI sequences, further synthesizing enhanced images. Our contributions can be summarized as follows: (1) We propose an interpretable network architecture capable of decoupling and multi-frequency-domian feature weighting mechanism (MFDFW) of high-frequency and low-frequency features from different sequences while quantifying the contribution of high-frequency and low-frequency features in each sequence. (2) We design a wavelet transform decoupling loss that guides the model to effectively decouple high-frequency and low-frequency features, thereby enhancing the feature extraction process. (3) We propose a dual-branch frequency-domain disentanglement module (DBFD) that better decouples high-frequency and low-frequency features from different sequences. (4) We validate the model through tasks such as super-resolution, denoising, and motion artifact removal, demonstrating its generalization ability on external datasets.

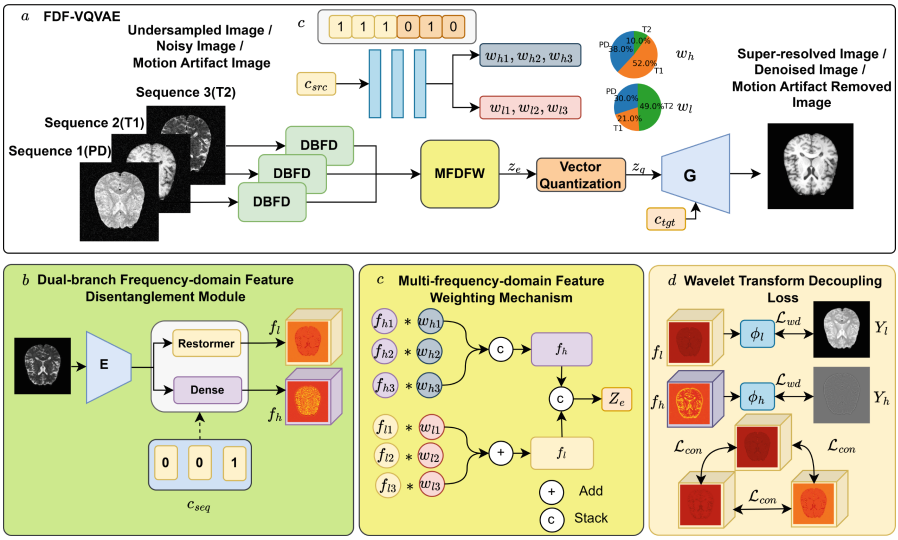


Fig. 1. The overview of the FDF-VQVAE framework.

2 Method

2.1 Frequency Disentanglement and Fusion Learning Framework

Figure 1 illustrates the overall architecture of the proposed FDF-VQVAE network. The network mainly consists of an encoder E, a dual-branch frequency-domain disentanglement module, multi-frequency-domian feature weighting mechanism, and a decoder G. The decoder G is composed of two hyper ConvNeXt blocks, a hyper convolution layer, and LeakyReLU. First, the input MRI sequence is processed by the encoder E to extract shallow features. Then, the dual-branch frequency-domain disentanglement module decouples the high-frequency and low-frequency features of all input sequences based on c_{seq}. Next, the weights for the high and low frequency features are generated through c_{src} to enable the effective fusion of multi-sequence features. Finally, the fused features Z_e are transformed into quantized features Z_q using vector quantization [16], which are then passed to the decoder G along with c_{tgt} to generate the target sequence. In addition, in this network, c is composed of c_{src} and c_{tgt}.

Dual-Branch Frequency-Domain Disentanglement Module. Define a set of N sequences MRI images with its corresponding simulated degraded image $\mathcal{I} = \{Y_i, X_i \mid i = 1, \ldots, N\}$, where $X_i = \mathcal{F}_d(Y_i)$. $\mathcal{F}_d$ represents the degradation function. We propose a dual-branch frequency-domain feature disentanglement module to extract high and low-frequency features $f_{\text{hi}}, f_{\text{li}} = \text{DBFD}(X_i, c_{\text{seq}})$, as shown in Fig. 1(b). Here, c_{seq} represents the one-hot encoding of the sequence. The module consists of an Encoder, a hyper Restormer block, and a hyper dense

block. The hyper Restormer module [17], where hyper convolutions [7] replace standard convolutions, is used to extract global features from MRI images. The hyper dense module [21], similarly using hyper convolutions instead of standard convolutions, is designed to extract local features from MRI images. This enables targeted disentangling for different MRI sequences. The encoder consists of two ConvNeXt blocks, a convolutional layer, LayerNorm, and LeakyReLU, aimed at extracting more complex high-dimensional features while ensuring computational efficiency, thereby enhancing feature representation ability.

Multi-frequency-Domian Feature Weighting Mechanism. The low-frequency features represent shared information, while the high-frequency features represent complementary information. Therefore, we integrate the low-frequency features using a weighted sum and the high-frequency features using weighted concatenation, followed by further fusion of the integrated high and low-frequency information. To enable the model to learn the optimal weights, we define a learnable parameter to predict the best weights $\omega_i \in \mathbb{R}^N$ from c_{src}.

$$\omega_i = \text{sigmoid}(c_{\text{src}} \cdot \omega + b) + \epsilon, \quad i \in \{low, high\} \tag{1}$$

where ω and b are the weights and biases of the fully connected layer. $\epsilon = 10^-5$ to avoid dividing 0. To simulate modality missing, during training, we randomly set c_{src} to 0 in the MFDFW mechanism and ensure that the output ω_i sums to 1.

$$\omega_i = \frac{\omega_i \cdot c_{\text{src}}}{\sum_{j=1}^{N} \omega_{ij} + \epsilon}, \quad i \in \{low, high\} \tag{2}$$

where $\cdot$ refers to the element-wise product, j represents the j^{th} weight in ω_i, N is the number of ω_i, $\epsilon = 10^{-5}$ prevent the denominator from being zero.

By using the Dual-Branch Frequency-Domain Disentanglement Module and the high-frequency and low-frequency weights ω_h and ω_l, the weighted fusion of high-frequency and low-frequency features from multiple sequences is achieved.

$$\begin{aligned} f_h &= [\omega_{h1} \cdot f_{h1}, \omega_{h2} \cdot f_{h2}, \omega_{h3} \cdot f_{h3}] \\ f_l &= \sum_{i=1}^{N} \omega_{li} \cdot f_{li} \\ f &= [f_h, f_l] \end{aligned} \tag{3}$$

where f_{hi} and f_{li} represent the high-frequency and low-frequency features of the i-th input sequence, respectively. $[\cdot, \cdot]$ presents channel-wise concatenation.

2.2 Loss Function

Wavelet Transform Decoupling Loss. To better decouple high-frequency and low-frequency information, we propose a wavelet transform decoupling loss.

First, we introduce a low-frequency consistency loss $\mathcal{L}_{\text{con}}$ to enforce the consistency of the low-frequency features extracted from different sequences. Using wavelet constraint loss $\mathcal{L}_{\text{wd}}$ to guide model to further decouple the high and low-frequency information, we use the wavelet-transformed high and low-frequency images Y_{hi} and Y_{li} of the original image to supervise the extracted features f_{hi} and f_{li}. We remap the decoupled features back to the image space and apply pixel-level loss for constraint. Wavelet transform decoupling loss consists of $\mathcal{L}_{\text{con}}$ and $\mathcal{L}_{\text{wd}}$.

$$\mathcal{L}_{\text{con}}(i,j) = \|f_{li} - f_{lj}\|_2^2 \tag{4}$$

where f_{li} and f_{lj} represent the low-frequency features extracted from the different input sequences.

$$\mathcal{L}_{\text{wd}}(f_{li}, Y_{li}, f_{hi}, Y_{hi}) = \mathcal{L}_{rec}(\phi_l(f_{li}), Y_{li}) + \mathcal{L}_{rec}(\phi_{hi}(f_{hi}), Y_{hi}) \tag{5}$$

where ϕ_l denotes the shared weight projection for low-frequency features, and ϕ_{hi} denotes the private weight projection for high-frequency features. The reconstruction loss, $\mathcal{L}_{rec}$, consists of the mean squared error (MSE) and structure similarity index measure (SSIM) loss.

Image Reconstruction Loss. To improve the enhancement effect, we impose constraints on the generated images at the image level, structural level, and semantic level. $\hat{Y}$ represents the model's output image.

$$\begin{aligned}\mathcal{L}_{\text{pixel}}(Y,\hat{Y}) = &\mathcal{L}_{\text{MSE}}(Y,\hat{Y}) + \mathcal{L}_{\text{SSIM}}(Y,\hat{Y}) + \mathcal{L}_{\text{Lap}}(Y,\hat{Y}) \\ &+ \mathcal{L}_{\text{per}}(Y,\hat{Y}) + L_{vq}(z_e, z_q)\end{aligned} \tag{6}$$

Here, $\mathcal{L}_{\text{MSE}}$ denotes the MSE loss, $\mathcal{L}_{\text{SSIM}}$ denotes the SSIM loss, $\mathcal{L}_{\text{Lap}}$ denotes the Laplacian loss, $\mathcal{L}_{\text{per}}$ denotes the perceptual loss based on a pre-trained VGG19 network, and $\mathcal{L}_{\text{vq}}$ denotes the commitment loss and vector quantization loss [16].

Total Loss. The loss function combines image reconstruction loss, wavelet transform decoupling loss, and low-frequency consistency loss to constrain the network. The implementation is as follows:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{pixel}}(Y,\hat{Y}) + \sum_{i=1}^{N} \mathcal{L}_{\text{wd}}(f_{li}, Y_{li}, f_{hi}, Y_{hi}) + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \mathcal{L}_{\text{con}}(i,j) \tag{7}$$

3 Experiments

3.1 Experimental Settings

Dataset and Evaluation Metrics. We conducted a study using brain MRI images from 568 subjects in the IXI dataset [10], which included three aligned sequences: T1, T2, and proton density (PD). We selected 340 subjects for training, 57 for validation, and 171 for testing. Additionally, we used brain MRI

images from 357 subjects in the BraTS2021 dataset [1] for external validation, which included T1 and T2 sequences. All MRI images underwent intensity normalization, with the intensity range standardized to [0, 1]. We simulated different low-quality images in the following ways: for the super-resolution task, we reduced the MRI image quality by applying 8x Fast Fourier Transform (FFT) undersampling to simulate low-resolution images; for the denoising task, gaussian noise with a mean of 0 and a standard deviation of 1 was added to the MRI images to simulate real-world noise interference; for motion artifact generation, we applied 6x cartesian acceleration undersampling to simulate image distortions caused by motion. To evaluate the synthesis performance, we used three common performance metrics: peak signal-to-noise ratio (PSNR), structure similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS). Using these three metrics, we conducted a comprehensive evaluation of the synthesized images from the perspectives of intensity, structure, and perception.

3.2 Implementation Details

The models are implemented with PyTorch and trained on the NVIDIA A40 GPU. The E and G are trained using the AdamW optimizer with an initial learning rate of 0.0001, betas = (0.9, 0.95), weight decay = 0.05, a batch size of 1, for 500 epochs.

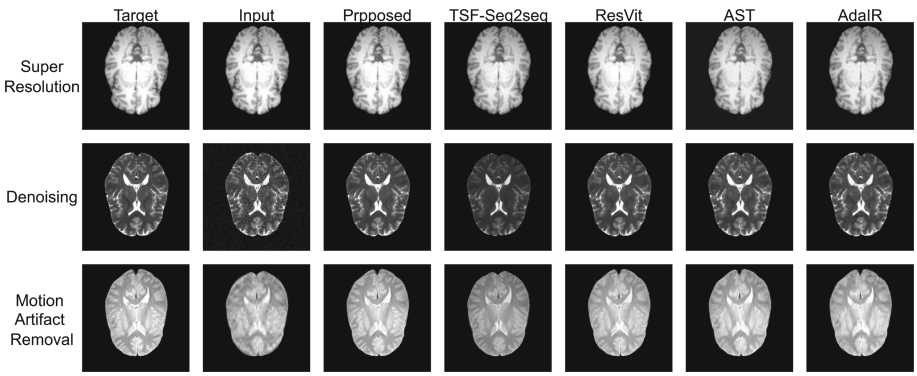


Fig. 2. The synthesized results of PD, T1, and T2 for our model and the comparison model in the tasks of super-resolution, denoising, and motion artifact removal.

3.3 Experimental Results

Comparative Experiment. We compared the proposed method with multi-sequence models (ResVit [5], TSF-seq2seq [8]) and single-sequence models (AST

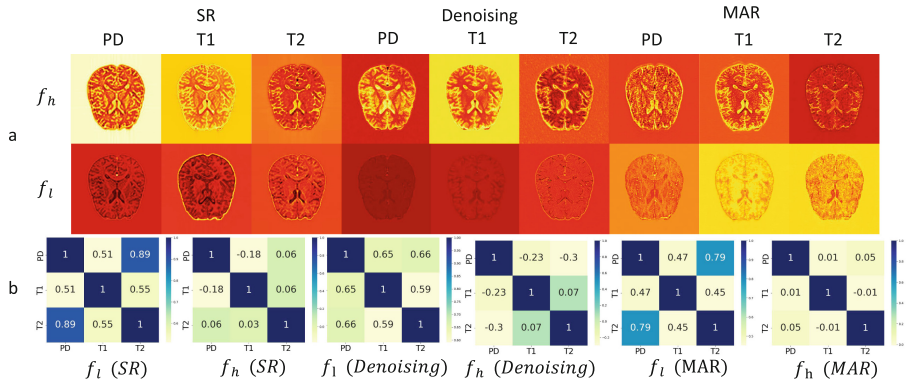


Fig. 3. (a) The visualization of high and low frequency features. (b) The CC between high-frequency features and the CC between low-frequency features in the three tasks.

Table 1. The results of super-resolution, denoising, and motion artifact removal on the IXI dataset and BraTS2021 dataset.

Dataset		IXI Dataset			BraTS2021 dataset (External Validation)				
Task name	Methods	SSIM ↑	PSNR (dB) ↑	LPIPS ↓	SSIM ↑	PSNR (dB) ↑	LPIPS ↓	Parameters (M)	GFLOPs (G)
SR	Our Model	0.972 ± 0.01	32.5 ± 1.8	0.039 ± 0.01	0.963 ± 0.01	31.2 ± 1.4	0.060 ± 0.02	0.3	157.2
	TSF-Seq2Seq	0.967 ± 0.01	31.6 ± 1.7	0.046 ± 0.04	0.948 ± 0.02	25.3 ± 2.4	0.090 ± 0.03	7.3	67.1
	AST	0.954 ± 0.01	31.3 ± 1.7	0.112 ± 0.03	0.949 ± 0.01	30.8 ± 1.7	0.145 ± 0.04	47.7	189.0
	AdaIR	0.957 ± 0.01	31.1 ± 2.1	0.117 ± 0.05	0.933 ± 0.02	30.8 ± 1.2	0.093 ± 0.02	28.7	344.9
	ResVit	0.969 ± 0.01	31.6 ± 2.0	0.076 ± 0.03	0.947 ± 0.01	26.6 ± 1.6	0.106 ± 0.03	11.3	167.6
Denoising	Our Model	0.975 ± 0.01	33.9 ± 1.1	0.028 ± 0.01	0.972 ± 0.01	32.4 ± 1.1	0.049 ± 0.02	0.3	157.2
	TSF-Seq2Seq	0.971 ± 0.01	33.4 ± 1.3	0.042 ± 0.01	0.953 ± 0.01	27.0 ± 2.7	0.085 ± 0.03	7.3	67.1
	AST	0.942 ± 0.01	32.6 ± 0.8	0.052 ± 0.01	0.910 ± 0.01	32.1 ± 0.7	0.067 ± 0.01	47.7	189.0
	AdaIR	0.951 ± 0.01	32.1 ± 1.0	0.050 ± 0.01	0.903 ± 0.01	31.4 ± 0.6	0.067 ± 0.02	28.7	344.9
	ResVit	0.972 ± 0.01	33.4 ± 1.3	0.042 ± 0.01	0.963 ± 0.01	30.8 ± 1.9	0.083 ± 0.02	11.3	167.6
Motion artifact removal	Our Model	0.973 ± 0.01	33.5 ± 1.4	0.026 ± 0.01	0.970 ± 0.01	32.1 ± 1.2	0.041 ± 0.01	0.3	157.2
	TSF-Seq2Seq	0.964 ± 0.01	33.3 ± 1.6	0.046 ± 0.03	0.940 ± 0.01	25.4 ± 2.4	0.100 ± 0.03	7.3	67.1
	AST	0.954 ± 0.01	31.2 ± 1.2	0.067 ± 0.01	0.928 ± 0.01	30.7 ± 1.2	0.075 ± 0.01	47.7	189.0
	AdaIR	0.946 ± 0.01	30.9 ± 1.2	0.068 ± 0.01	0.926 ± 0.02	30.9 ± 1.1	0.082 ± 0.02	28.7	344.9
	ResVit	0.966 ± 0.01	32.4 ± 1.9	0.047 ± 0.02	0.962 ± 0.01	31.2 ± 1.2	0.060 ± 0.02	11.3	167.6

[**?**], AdaIR [4]). Figure 2 shows the effect maps for super-resolution (SR), denoising, and motion artifact removal (MAR). Table 1 summarizes the performance of these models in these tasks. The proposed method outperforms all others, with multi-sequence models showing better results than single-sequence models, highlighting the benefit of using complementary information from multiple sequences. Our method uses only about 0.3 M parameters and around 157.2 GFLOPS of computation, achieving a superior parameterâĂŞcompute balance while maintaining high performance.

Ablation Study. To validate the rationality of each module in the model, we conducted ablation experiments on wavelet transform decoupling loss, DBFD, and MFDFW, with the results shown in Table 2. The wavelet transform decoupling loss effectively constrains the model's performance in high and low frequency feature extraction; the DBFD module effectively decouples high and low

Table 2. The results of the ablation experiments on the model conducted on the IXI dataset.

Dataset					IXI Dataset		
Task name		L_{wd}, L_{con}	MFDFW	DBFD	SSIM ↑	PSNR (dB) ↑	LPIPS ↓
SR	Our Model	✓	✓	✓	$\underline{0.972 \pm 0.01}$	$\underline{32.5 \pm 1.8}$	$\underline{0.039 \pm 0.01}$
	Our Model		✓	✓	0.958 ± 0.02	29.3 ± 1.6	0.067 ± 0.02
	Our Model	✓		✓	0.957 ± 0.01	29.8 ± 1.5	0.063 ± 0.01
	Our Model	✓	✓		0.936 ± 0.03	30.0 ± 1.7	0.079 ± 0.02
Denoising	Our Model	✓	✓	✓	$\underline{0.975 \pm 0.01}$	$\underline{33.9 \pm 1.1}$	$\underline{0.028 \pm 0.01}$
	Our Model		✓	✓	0.962 ± 0.01	30.4 ± 1.5	0.049 ± 0.01
	Our Model	✓		✓	0.963 ± 0.01	31.0 ± 1.9	0.043 ± 0.01
	Our Model	✓	✓		0.962 ± 0.01	30.8 ± 1.0	0.051 ± 0.01
Motion artifact removal	Our Model	✓	✓	✓	$\underline{0.973 \pm 0.01}$	$\underline{33.5 \pm 1.4}$	$\underline{0.026 \pm 0.01}$
	Our Model		✓	✓	0.955 ± 0.01	30.2 ± 1.5	0.048 ± 0.01
	Our Model	✓		✓	0.959 ± 0.01	30.6 ± 1.5	0.046 ± 0.01
	Our Model	✓	✓		0.956 ± 0.01	30.0 ± 1.3	0.052 ± 0.01

frequency features, enhancing the model's performance; and the MFDFW mechanism effectively fuses high and low frequency features across multiple sequences.

Visualization of Feature Extraction. As shown in Fig. 3, the visualization of high and low frequency features and their correlation coefficients (CC) across three tasks is presented. As shown in Fig. 3, the model effectively extracts both high and low frequency features in the MRI image.

Contribution of Low and High-Frequency Features. In the SR task, the low-frequency weight of the PD sequence (0.98) helps maintain the stability of anatomical structures, while the low-frequency weight of the T1/T2 sequences (0.01) mitigates information conflicts caused by inter-modality contrast differences. The high-frequency weight of the T2 sequence (0.58) enhances the texture of lesion boundaries, the high-frequency weight of the T1 sequence (0.39) improves the sharpness of the gray-white matter interface, and the high-frequency weight of the PD sequence (0.03) effectively suppresses interference caused by low signal-to-noise ratio. In the denoising task, the low-frequency weight of the T2 sequence (0.49) stabilizes pathological signals, the weight of the PD sequence (0.3) performs brightness correction, and the low-frequency weight of the T1 sequence (0.21) prevents the amplification of contrast noise. The high-frequency weight of the T1 sequence (0.52) preserves the details of the gray-white matter structure, the PD sequence (0.38) supplements anatomical texture, and the high-frequency weight of the T2 sequence (0.1) supplements other details.

In the MAR task, regarding low-frequency weights, the PD sequence (0.53) is primarily used to correct motion artifacts, the T1 sequence (0.26) assists in locating low-frequency structures at the gray-white matter boundary, and the T2 sequence (0.21) supplements the stability of fluid regions. In terms of high-frequency weights, the T2 sequence (0.49) helps to enhance true edges, the T1 sequence (0.46) aids in identifying the direction of artifacts, and the PD sequence (0.05) suppresses motion-related high-frequency noise.

External Validation. We conducted an external validation of our model and comparison models using the BraTS2021 dataset, which includes 357 subjects, to explore the model's generalization capability. Table 1 shows that the proposed method outperforms all other models in every task, demonstrating superior performance and excellent generalization ability.

3.4 Conclusion

In this study, we propose an interpretable network architecture designed to effectively extract high-frequency and low-frequency features from different sequences and adaptively adjust the fusion weights of these features according to specific task requirements. Additionally, we introduce a wavelet transform decoupling loss function to guide the model in extracting high-frequency and low-frequency features from MRI images. This method is suitable for multi-sequence synthesis tasks such as super-resolution reconstruction, denoising, and motion artifact removal. Experimental results on the IXI and BraTS2021 datasets demonstrate that our method effectively utilizes complementary information between sequences, significantly outperforming existing state-of-the-art techniques.

Acknowledgments. This work was supported in part by Macao Polytechnic University Grant (RP/FCA-08/2024), the grant from Science and Technology Development Fund of Macao (0004/2024/E1B1).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baid, U., et al.: The RSNA-ASNR-MICCAI brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 782–791 (2021)
3. Chen, J.H., Su, M.Y.: Clinical application of magnetic resonance imaging in management of breast cancer patients receiving neoadjuvant chemotherapy. Biomed. Res. Int. **2013**(1), 348167 (2013)

4. Cui, Y., Zamir, S.W., Khan, S., Knoll, A., Shah, M., Khan, F.S.: AdaIR: adaptive all-in-one image restoration via frequency mining and modulation. arXiv preprint arXiv:2403.14614 (2024)
5. Dalmaz, O., Yurt, M., Çukur, T.: ResViT: residual vision transformers for multimodal medical image synthesis. IEEE Trans. Med. Imaging **41**(10), 2598–2614 (2022)
6. Feng, C.-M., Yan, Y., Fu, H., Chen, L., Xu, Y.: Task transformer network for joint MRI reconstruction and super-resolution. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 307–317. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_30
7. Han, L., et al.: Synthesis-based imaging-differentiation representation learning for multi-sequence 3D/4D MRI. Med. Image Anal. **92**, 103044 (2024)
8. Han, L., et al.: Non-adversarial learning: vector-quantized common latent space for multi-sequence MRI. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15011, pp. 481–491. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72120-5_45
9. Han, L., et al.: An explainable deep framework: towards task-specific fusion for multi-to-one MRI synthesis. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14229, pp. 45–55. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43999-5_5
10. Jiang, D., Dou, W., Vosters, L., Xu, X., Sun, Y., Tan, T.: Denoising of 3D magnetic resonance images with multi-channel residual learning of convolutional neural network. Jpn. J. Radiol. **36**, 566–574 (2018)
11. Li, F., et al.: Biomed research international. Heart **24**(69), 8–74 (2015)
12. Mann, R.M., Cho, N., Moy, L.: Breast MRI: state of the art. Radiology **292**(3), 520–536 (2019)
13. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
14. Qu, L., Zhang, Y., Wang, S., Yap, P.T., Shen, D.: Synthesized 7T MRI from 3T MRI via deep learning in spatial and wavelet domains. Med. Image Anal. **62**, 101663 (2020)
15. Sadikov, A., Wren-Jarvis, J., Pan, X., Cai, L.T., Mukherjee, P.: Generalized diffusion MRI denoising and super-resolution using SWIN transformers. arXiv preprint arXiv:2303.05686 (2023)
16. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
17. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5728–5739 (2022)
18. Zhao, Z., et al.: CDDFuse: correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5906–5916 (2023)
19. Zhou, S., Chen, D., Pan, J., Shi, J., Yang, J.: Adapt or perish: adaptive sparse transformer with attentive feature refinement for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2952–2963 (2024)

20. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-Net: hybrid-fusion network for multi-modal MR image synthesis. IEEE Trans. Med. Imaging **39**(9), 2772–2781 (2020)
21. Zhu, Y., Newsam, S.: DenseNet for dense flow. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 790–794. IEEE (2017)



Fit Pixels, Get Labels: Meta-learned Implicit Networks for Image Segmentation

Kushal Vyas, Ashok Veeraraghavan, and Guha Balakrishnan(✉)

Rice University, Houston, TX 77005, USA
guha@rice.edu

Abstract. Implicit neural representations (INRs) have achieved remarkable successes in learning expressive yet compact signal representations. However, they are not naturally amenable to predictive tasks such as segmentation, where they must learn semantic structures over a distribution of signals. In this study, we introduce MetaSeg, a meta-learning framework to train INRs for medical image segmentation. MetaSeg uses an underlying INR that simultaneously predicts per pixel intensity values and class labels. It then uses a meta-learning procedure to find optimal initial parameters for this INR over a training dataset of images and segmentation maps, such that the INR can simply be fine-tuned to fit pixels of an unseen test image, and automatically decode its class labels. We evaluated MetaSeg on 2D and 3D brain MRI segmentation tasks and report Dice scores comparable to commonly used U-Net models, but with 90% fewer parameters. MetaSeg offers a fresh, scalable alternative to traditional resource-heavy architectures such as U-Nets and vision transformers for medical image segmentation.

Keywords: Medical Image Segmentation · Implicit Neural Representations · Meta-Learning · MRI

1 Introduction

Implicit neural representations (INRs) have demonstrated impressive performance in encoding complex natural signals. A typical INR $f_\theta(\cdot)$ is a multilayer perceptron (MLP) with parameters θ that maps a coordinate $\mathbf{x} \in \mathrm{R}^d$ to signal value $I(\mathbf{x}) \in \mathrm{R}^D$ and is iteratively fit to a specific signal. Because they offer continuous signal representations, excellent reconstruction performance, and implicit signal priors, INRs have gained significant attention in the computer vision community for compactly modeling large signals [13,16,17], and in the medical imaging community for inverse imaging tasks such as accelerated MRI [1], and sparse-view CT reconstruction [18,23]. Despite these successes, learned INR representations are highly specific to a given signal and to the way its parameters are initialized. As a result, unlike architectures such as U-Nets [15] and vision transformers [9], INR-produced features lack structural or semantic coherence, making them less used for tasks requiring learning over an image distribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 194–203, 2026.
https://doi.org/10.1007/978-3-032-04947-6_19

Results from recent studies suggest that this shortcoming of INRs may be fixable. In particular, by fitting an INR across multiple images starting from the same parameter intialization, the final parameters exhibit clear semantic, structural properties for datasets such as faces or MRI scans [6,21,22]. And as a result, these learned initializations allow the INR to fit unseen (test) images rapidly with far fewer gradient updates while learning more generalizable features [2,21,22]. In this study, we leverage this insight to develop an INR fitting strategy for medical image segmentation. In particular, we present the surprising discovery that an INR carefully optimized to fit many training pairs of images and associated segmentation maps can predict a segmentation map for an unseen test image *when simply fine-tuned to reconstruct the image's pixels.*

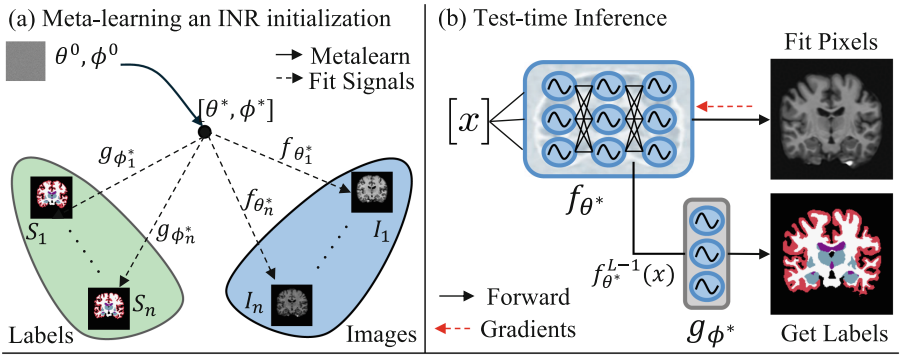


Fig. 1. Overview of MetaSeg. (a) We use a meta-learning framework to learn optimal initial parameters θ^*, ϕ^* for an INR consisting of an L-layer reconstruction network $f_\theta(\cdot)$ and shallow segmentation head $g_\phi(\cdot)$. (b) At test time, optimally initialized INR f_{θ^*} is iteratively fit to the pixels of an unseen test scan. After convergence, the penultimate features $f_{\theta^*}^{L-1}(x)$ are fed as input to the segmentation head $g_\phi^*(\cdot)$ to predict per-pixel class labels.

We propose MetaSeg, a novel meta-learning [7] INR framework that performs this strategy. MetaSeg uses an INR that simultaneously predicts both image pixels and segmentation labels given a coordinate (Fig. 1b), and aims to find an optimal setting of the INR parameters (θ and ϕ), such that the network can be quickly fine-tuned to fit any image-mask pair starting from these parameters. MetaSeg uses a nested, meta-learning procedure (Fig. 1a). In the inner optimization, the INR is fit to reconstruct a randomly selected batch of training image and segmentation mask pairs. The outer (meta) optimization applies a gradient descent step to the current values of θ and ϕ based on the converged inner parameters. These steps iterate to converge to an initialization that works well for both image reconstruction and segmentation objectives from the given image distribution. At test time, we simply fit the INR initialized with the learned optimal parameters to the test image (without needing any

segmentation information), generating viable segmentation masks in just 2 gradient descent updates and near state-of-the-art quality masks after 100 updates.

MetaSeg draws inspiration from Neural Implicit Segmentation Functions (NISFs) [20], which also use an INR to predict pixel and segmentation values. However, instead of forcing the INR to learn image-label correlations, NISF takes an additional latent vector as input to represent an image's semantic content, which must then be optimized at test time to produce a mask. Due to this strategy, NISF must apply several regularization losses on INR parameters and the latent space to succeed. In contrast, MetaSeg offers a much simpler training methodology based on meta-learning, only requiring reconstruction and segmentation losses during training and only a reconstruction loss during testing.

We evaluated MetaSeg on 2D and 3D magnetic resonance imaging (MRI) segmentation tasks using the OASIS-MRI [8,12] dataset. We first find that MetaSeg achieves similar performance for 2D (Dice score: 0.93) and 3D (Dice score: 0.91) MRI segmentation tasks to widely used U-Net baselines [4,14] with 90% fewer parameters, and significantly outperforms NISF [20]. Second, we show that MetaSeg performs well for fine-grained segmentation tasks, accurately segmenting image regions which may be sparse and span only a few pixels. Third, we show that when trained on low-resolution scans, MetaSeg fares better on high-resolution scans than U-Nets, likely due to its continuous representation. Finally, we qualitatively demonstrate through principal component analysis (PCA) visualizations that MetaSeg's learned initial parameters encode semantic information correlated with ground truth segmentations.

2 Methods

We assume a given dataset $\mathcal{D} = \{I_j, S_j\}_{j=1}^{N}$ of N subjects, where I_j and S_j are a d-dimensional image scan and segmentation map, respectively ($d = 2$ and 3 in our experiments). At location $x \in \mathbb{R}^d$, we have $I_j(x) \in \mathbb{R}^D$ and $S_j(x) \in \{0,1\}^C$, where D is the number of image channels, and C is the number of classes.

We propose the MetaSeg framework to estimate $\hat{S}$ for an unseen scan I at test time. MetaSeg consists of an INR (Fig. 1b) with an L-layer MLP $f_\theta : \mathrm{R}^d \rightarrow \mathrm{R}^D$ that predicts $\hat{I}(\mathbf{x})$ given coordinate $\mathbf{x}$, and a segmentation head $g_\phi : \mathrm{R}^h \rightarrow \mathrm{R}^C$ that predicts C class probabilities given the penultimate features, of dimensionality h, computed by f_θ: $f_\theta^{L-1}(\mathbf{x})$. At test time, we aim to fit $f_\theta(\cdot)$ on I for T_f steps, and then simply compute $\hat{S}(\mathbf{x}) = g_\phi(f_\theta^{L-1}(\mathbf{x}))$ in one feed-forward operation.

In Sect. 2.1, we describe MetaSeg's strategy for learning optimal values θ^*, ϕ^* over $\mathcal{D}$, such that $f_{\theta^*}(\cdot)$ and $g_{\phi^*}(\cdot)$ (networks initialized with θ^* and ϕ^*) may be easily fine-tuned to reconstruct and segment a scan from the distribution. In Sect. 2.2, we describe how the INR initialized with these parameters may be used at inference time to estimate a segmentation map for a scan.

2.1 Training: Meta-learning and Segmentation Head Optimization

MetaSeg uses a MAML [7] meta-learning strategy to learn optimal parameters θ^*, ϕ^*, consisting of a nested optimization with *inner* and *outer* (meta) routines (Fig. 1a). Starting with initial random values θ^0 and ϕ^0 at iteration $t = 0$, the outer routine takes T_o gradient descent steps to move the parameters to generalizable values. The gradient at each outer step is based on the results of the inner routine, which fits $f_{\theta^t}(\cdot)$ and $g_{\phi^t}(\cdot)$ for T_i steps on one data pair (I_j, S_j). The outer routine uses the final inner parameter values as a gradient signal to update θ^t, ϕ^t. To avoid a noisy gradient due to overfitting to example j, we set T_i to a small value in practice (e.g., $T_i = 2$).

In the inner routine, at time step t, we optimize both $f_{\theta^t}(\cdot)$ and $g_{\phi^t}(\cdot)$ for a single subject j with a loss function:

$$\mathcal{L}_{inner}(I_j, \hat{I}_j, S_j, \hat{S}_j) = \sum_{\mathbf{x}} \mathcal{L}_{recon}(I_j(\mathbf{x}), \hat{I}_j(\mathbf{x})) + \mathcal{L}_{cls}(S_j(\mathbf{x}), \hat{S}_j(\mathbf{x})), \quad (1)$$

where $\mathcal{L}_{recon}(\mathbf{a}, \mathbf{b}) = || \mathbf{a} - \mathbf{b} ||_2^2$ is a per-pixel reconstruction loss, and

$$\mathcal{L}_{cls}(\mathbf{a}, \mathbf{b}) = \sum_{c=1}^{C} -(1 - \mathbf{b}(c))^{\gamma} \cdot \delta_{c,\mathbf{a}} \log(\mathbf{b}(c)) \quad (2)$$

is a per-pixel multiclass focal classification loss [11], where γ is a hyperparameter. We use a focal loss to account for heavy class imbalance across pixels. In the outer step, we make a gradient descent update:

$$[\theta^{t+1}, \phi^{t+1}] \leftarrow [\theta^t, \phi^t] - \beta \nabla_{[\theta^t, \phi^t]} \big(\mathcal{L}_{inner}(I_j, \hat{I}_j, S_j, \hat{S}_j) \big), \quad (3)$$

where gradient $\nabla_{[\theta^t, \phi^t]}(\cdot)$ computes the difference between the converged parameters $(\theta_j^t,\ \phi_j^t)$ of the inner optimization and current parameters θ^t, ϕ^t. After T_o outer gradient descent steps, we obtain parameters θ^{T_o} and ϕ^{T_o}.

We freeze $\theta^* = \theta^{T_o}$, but further optimize ϕ^{T_o} such that when, at test time, $f_{\theta^*}(\cdot)$ is fit for T_f iterations on any random scan, we can obtain accurate segmentation predictions, in one step. To do so, we first fit $f_{\theta^*}(\cdot)$ separately on each training scan I_j for T_f iterations and populate a dataset of learned scan-specific features and segmentation maps, $\{f_{\theta_j^*}^{L-1}(\mathbf{x}), S_j(\mathbf{x})\}_{j=1}^{N}$. We then globally optimize $g_{\phi}(\cdot)$ by minimizing the loss function $\mathcal{L}_{seg}$:

$$\mathcal{L}_{seg}(\mathcal{D}) = \sum_{\mathbf{x}} \sum_{j} \mathcal{L}_{cls}(S_j(\mathbf{x}), g_{\phi}(f_{\theta_j^*}^{L-1}(\mathbf{x}))). \quad (4)$$

We freeze the converged parameters as ϕ^*.

2.2 Inference: Fit Pixels, Get Labels

At inference time, we assume a given (unseen) scan I. We fit $f_{\theta^*}(\cdot)$ to I for T_f iterations, solely optimizing the pixelwise reconstruction loss $\mathcal{L}_{recon}$. Upon completion, we compute $\hat{S}(\mathbf{x}) = g_{\phi^*}(f_{\theta^*}^{L-1}(\mathbf{x}))$, and apply softmax and argmax operations to $\hat{S}(\mathbf{x})$ to obtain the predicted class per pixel.

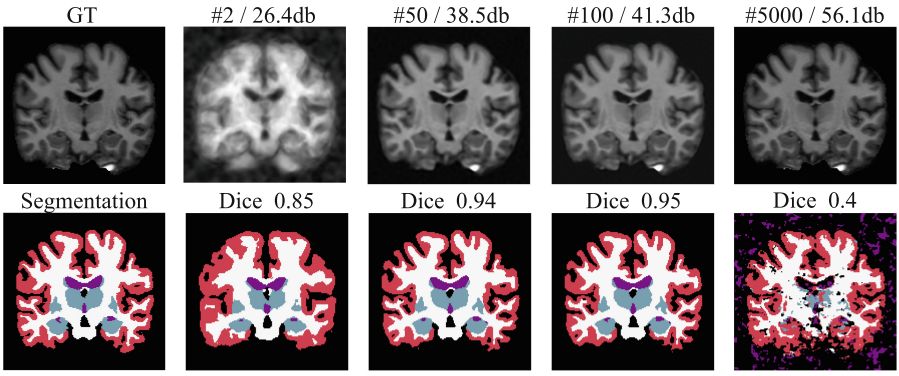


Fig. 2. Example progression of reconstruction and segmentation performance as a MetaSeg INR fits the pixels of a test image. We report PSNR (top) and Dice (bottom) scores. MetaSeg's initialization leads to rapid convergence, with a Dice score of 0.85 emerge as early as $T_f = 2$ iterations. It achieves optimal segmentation (Dice=0.95, PSNR=41.3) at $T_f = 100$ iterations, and eventually declines as it overfits to the image pixels at $T_f = 5000$ iterations.

3 Experiments and Results

We evaluated MetaSeg on the tasks of 2D and 3D MRI image segmentation. We used 414 T1 brain MRI scans from the OASIS-MRI [8,12] dataset, randomly split into 214 training, 100 validation, and 100 test scans. OASIS-MRI provides 5 coarse (background, white matter, gray matter, cortex, and cerebrospinal fluid), 24 fine-grained segmentation labels for 2D slices, and 35 segmentation labels for full 3D MRI volumes. We used aligned coronal sections provided in OASIS-MRI for 2D segmentation and aligned 3D volumes for 3D segmentation. We normalized all images and volumes to have intensity in $[0, 1]$. We padded and resized images to resolution 192×192, and cropped 3D volumes to $160 \times 160 \times 200$ followed by $2\times$ downsampling to yield $80 \times 80 \times 100$ resolution scans.

Implementation. For all experiments, we used a SIREN [19] INR with $L = [6, 5, 5]$ layers, width $h = [128, 512, 256]$, and $w_0 = 30$. We set the segmentation head $g_\phi(\cdot)$ as one fully connected layer with Leaky-ReLU activation, followed by a linear layer with C output values. We used the Adam [10] optimizer with a learning rate of 10^{-4} for inner and outer meta-learning optimizations. When tuning the segmentation head for 5 classes, we reduced the learning rate to $5e-5$. We set $T_i = 2$ and $T_f = 100$, and ran the outer loop for $T_o = 10$ epochs over the training data while validating the performance of optimized parameters θ^t, ϕ^t every 50 steps. For 2D segmentation with 5 classes, we set the focal loss hyperparameter $\gamma = 1.0$, and for 24 output classes we set $\gamma = 2.0$. For 3D MRI volumes which exhibit high background class imbalance, we set $\gamma = 3.0$ and further scaled the reconstruction loss by 0.1 for background pixels.

Table 1. Quantitative performance of MetaSeg and baseline models on 2D and 3D MRI segmentation tasks. MetaSeg achieves comparable Dice scores to popular U-Net baselines with 90% fewer parameters.

Task	Num. classes	Model	Dice Score ↑	Num. parameters ↓
2D MRI Segmentation	5	U-Net [4]	0.96 ± 0.008	7.7M
		MetaSeg	0.93 ± 0.012	83K
	24	U-Net [4]	0.84 ± 0.097	7.7M
		MetaSeg	$0.86 + 0.032$	1.06M
3D MRI Segmentation	5	SegResNet [14]	0.95 ± 0.006	4.7M
		NISF [20]	0.81 ± 0.007	293K
		MetaSeg	0.91 ± 0.011	330K

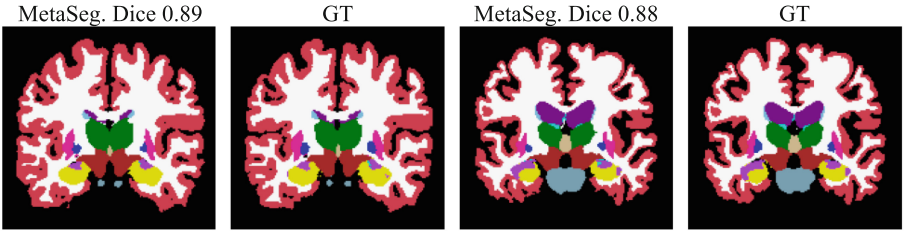


Fig. 3. Coarse and fine-grained 2D brain MRI segmentation results with 24 classes, for two subjects. MetaSeg accurately segments structures and is robust to high variations across subjects for structures such as ventricles (depicted in purple), brain stem (gray), and hippocampus (yellow). MetaSeg also adapts well to structures such as the cerebral cortex (red) which are not compact and localized in space.

Baselines: We adopt a widely used U-Net [15] baseline for 2D MRIs [4] and SegResNet [14] from the Monai [5] package for 3D MRIs. We trained all models using a learning rate of $1e{-}3$ until the validation loss saturated. We also compare against NISF [20], the most closely related INR method in the literature.

3.1 Results

Table 1 presents average Dice [3] scores with standard deviations for MetaSeg and baseline methods for both 2D and 3D cases. For 2D, MetaSeg achieves an excellent Dice score of 0.93 ± 0.012 for 5 classes, comparable to that achieved by the U-Net, but with 90% fewer parameters. Interestingly, as Fig. 2 illustrates, MetaSeg fits a signal remarkably fast, yielding a viable segmentation in just 2 update steps to $f_\theta(\cdot)$ (Dice: 0.85), and reaching peak performance at T_f steps. If fit further, MetaSeg's segmentation accuracy eventually falls off as parameters overfit to the particular scan. MetaSeg also performs well on the finer-grained segmentation task with 24 classes, achieving a Dice score of 0.86 ± 0.032 and outperforming the baseline U-Net. As shown in Fig. 3, MetaSeg provides con-

sistently accurate segmentation results for the ventricles and brain stem, which have drastic variations in size and area across examples.

MetaSeg also performs well for 3D segmentation, achieving an average Dice score of 0.91±0.011 with 90% fewer parameters than SegResNet, and significantly outperforms NISF. Figure 4 depicts rendered coronal, sagittal, and axial planes from the segmented volume showing high agreement with ground truth segmentation planes, also demonstrating that MetaSeg-encoded volumes can be readily queried at any viewing plane. We also explored whether MetaSeg can generate high resolution (2×) 3D segmentation maps when only trained on low resolution scans, due to its underlying continuous representation. MetaSeg achieved a Dice score of 0.78 ± 0.011 on this task, while SegResNet [14] achieved a Dice score of 0.73 ± 0.019. Hence, while there is a significant drop in performance in this scenario, MetaSeg still outperforms U-Nets.

Finally, we conducted an ablation study on parameter initialization strategy for 2D segmentation, with results reported in Table 2. MetaSeg performs significantly better than INRs initialized with random, fixed, and image-only meta-learning strategies. The latter result, in particular, demonstrates the benefit of learning structural priors with joint segmentation supervision.

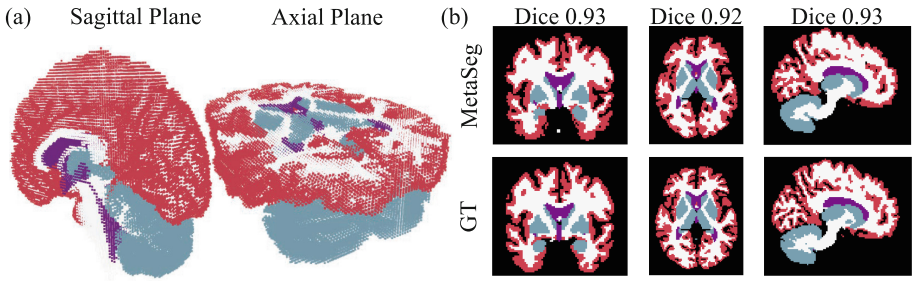


Fig. 4. MetaSeg is capable of accurately segmenting 3D MRI volumes. (a) Various cross sections of one learned volume. (b) Renderings of intermediate coronal, axial, and sagittal planes, with high agreement (Dice≈ 0.93) with ground truth.

Table 2. Ablation studies on INR initialization strategies for 2D segmentation. MetaSeg's approach of jointly learning initial INR and segmentation parameters outperforms random, fixed, and image-only meta-learning initialization strategies.

Init. Strategy	Random	Fixed	Meta-learn, Image Only	MetaSeg
Dice scores	0.30 ± 0.057	0.53 ± 0.1	0.81 ± 0.033	$\mathbf{0.93 \pm 0.012}$

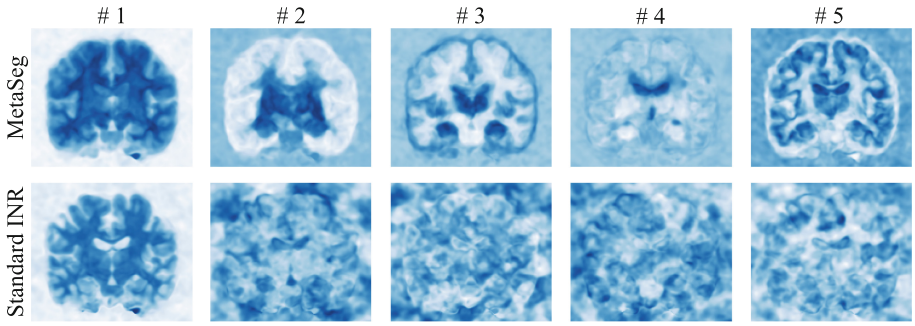


Fig. 5. PCA visualization of penultimate features learned by MetaSeg and a standard INR for a 2D test scan. We performed PCA separately on the features returned by each INR across spatial coordinates. We see a strong correlation for MetaSeg features with anatomical structures. For example, component #2 approximately resembles inner brain regions, #3 resembles regions like hippocampus and basal ganglia, #4 resembles ventricles, and #5 captures the cerebral cortex. On the contrary, a standard INR yields seemingly random features.

4 Discussion and Conclusion

Results demonstrate that MetaSeg achieves strong performance on both 2D and 3D MRI segmentation tasks, with comparable Dice scores to common U-Net baselines [4,14] but with 90% fewer parameters. In addition, MetaSeg significantly outperforms NISF [20], a related INR approach for segmentation. Ablation results demonstrate that our meta-learning strategy to initialize parameters with joint reconstruction and segmentation supervision is crucial for success.

MetaSeg offer a new perspective on the capabilities of INRs for imaging tasks, particularly beyond signal representation, which has been their predominant application in the literature. In particular, the results demonstrate that the meta-learning strategy can be useful for image translation tasks such as segmentation, where there is strong cross-correlation across signals. Indeed, as demonstrated by PCA decomposition in Fig. 5, MetaSeg learns low-dimensional embeddings that seem to jointly model both pixel intensity and semantic structure.

Because an INR explicitly conditions its features on input coordinates, it is likely that an INR will learn highly signal-specific features localized in space. While this is a good property for signal representation, it may lead to poor performance on tasks requiring generalization, such as segmentation. We conducted a pilot test to measure the sensitivity of MetaSeg to the spatial alignment of test scans to the training set distribution by applying random orientation and translation augmentations. We found that Dice score drops by 2%-6% for test images rotated randomly in [5°,15°], and drops by 3%-9% for test images with a random translation in [5–10] pixels. Hence, MetaSeg can be somewhat sensitive to spatial alignment, though further experimentation is needed to understand whether this can be mitigated itself with proper training augmentation.

MetaSeg learns fast, is resource-friendly, and scales well with more dimensions, unlike typical vision and transformer models. It also is easy to train, using only standard reconstruction and classification losses. In an age where AI requires increasing resource and financial investments, MetaSeg provides a fresh new perspective to solving image segmentation tasks.

Acknowledgments. The authors acknowledge support from NIH award R01DE 032051, ARPA-H award D24AC00296, NSF awards IIS-2107313 and EEC-1648449.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arefeen, Y., Levac, B., Stoebner, Z., Tamir, J.: Infusion: Diffusion regularized implicit neural representations for 2d and 3d accelerated mri reconstruction. arXiv preprint arXiv:2406.13895 (2024)
2. Bauer, M., Dupont, E., Brock, A., Rosenbaum, D., Schwarz, J.R., Kim, H.: Spatial functa: scaling functa to imagenet classification and generation (2023). https://arxiv.org/abs/2302.03130
3. Bertels, J.: Optimizing the dice score and jaccard index for medical image segmentation: theory and practice. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 92–100. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_11
4. Buda, M., Saha, A., Mazurowski, M.A.: Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Comput. Biol. Med. **109** (2019). https://doi.org/10.1016/j.compbiomed.2019.05.002
5. Cardoso, M.J., et al.: MONAI: An open-source framework for deep learning in healthcare (Nov 2022). https://doi.org/10.48550/arXiv.2211.02701
6. Dupont, E., Kim, H., Eslami, S.M.A., Rezende, D.J., Rosenbaum, D.: From data to functa: your data point is a function and you can treat it like one. In: 39th International Conference on Machine Learning (ICML) (2022)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, 06–11 Aug, vol. 70, pp. 1126–1135. PMLR (2017). https://proceedings.mlr.press/v70/finn17a.html
8. Hoopes, A., Hoffmann, M., Greve, D.N., Fischl, B., Guttag, J., Dalca, A.V.: Learning the effect of registration hyperparameters with hypermorph. Mach. Learn. Biomed. Imaging **1**(IPMI 2021), 1–30 (Apr 2022). https://doi.org/10.59275/j.melba.2022-74f1
9. Khan, A., et al.: A recent survey of vision transformers for medical image segmentation (2023). https://arxiv.org/abs/2312.00634
10. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 318–327 (2020). https://doi.org/10.1109/TPAMI.2018.2858826

12. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (OASIS): cross-sectional MRI Data in young, middle aged, nondemented, and demented older adults. J. Cognitive Neurosci. **19**(9), 1498–1507 (2007). https://doi.org/10.1162/jocn.2007.19.9.1498
13. Martel, J.N.P., Lindell, D.B., Lin, C.Z., Chan, E.R., Monteiro, M., Wetzstein, G.: Acorn: adaptive coordinate networks for neural scene representation. ACM Trans. Graph. (SIGGRAPH) **40**(4) (2021)
14. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pp. 311–320. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_28
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Saragadam, V., LeJeune, D., Tan, J., Balakrishnan, G., Veeraraghavan, A., Baraniuk, R.G.: Wire: wavelet implicit neural representations. In: Conf. Computer Vision and Pattern Recognition (2023)
17. Saragadam, V., Tan, J., Balakrishnan, G., Baraniuk, R., Veeraraghavan, A.: Miner: multiscale implicit neural representations. In: European Conf. Computer Vision (2022)
18. Shi, J., Zhu, J., Pelt, D.M., Batenburg, K.J., Blaschko, M.B.: Implicit neural representations for robust joint sparse-view ct reconstruction. arXiv preprint arXiv:2405.02509 (2024)
19. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Adv. Neural. Inf. Process. Syst. **33**, 7462–7473 (2020)
20. Stolt-Ansó, N., McGinnis, J., Pan, J., Hammernik, K., Rueckert, D.: Nisf: Neural implicit segmentation functions. In: Greenspan, H., et al. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2023, pp. 734–744. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_70
21. Tancik, M., et al.: Learned initializations for optimizing coordinate-based neural representations. In: CVPR (2021)
22. Vyas, K.K., Humayun, I., Dashpute, A., Baraniuk, R., Veeraraghavan, A., Balakrishnan, G.: Learning transferable features for implicit neural representations. In: Advances in Neural Information Processing Systems, vol. 37 (2024)
23. Zha, R., Zhang, Y., Li, H.: Naf: Neural attenuation fields for sparse-view cbct reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 442–452. Springer (2022). https://doi.org/10.1007/978-3-031-16446-0_42



Forget-MI: Machine Unlearning for Forgetting Multimodal Information in Healthcare Settings

Shahad Hardan[1(✉)], Darya Taratynova[1], Abdelmajid Essofi[1], Karthik Nandakumar[2,3], and Mohammad Yaqub[2]

[1] Department of Machine Learning, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{shahad.hardan,darya.taratynova,abdelmajid.essofi}@mbzuai.ac.ae
[2] Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
mohammad.yaqub@mbzuai.ac.ae
[3] Department of Computer Science, Michigan State University, East Lansing, UAE
nandakum@msu.edu

Abstract. Privacy preservation in AI is crucial, especially in healthcare, where models rely on sensitive patient data. In the emerging field of machine unlearning, existing methodologies struggle to remove patient data from trained multimodal architectures, which are widely used in healthcare. We propose *Forget-MI*, a novel machine unlearning method for multimodal medical data, by establishing loss functions and perturbation techniques. Our approach unlearns unimodal and joint representations of the data requested to be forgotten while preserving knowledge from the remaining data and maintaining comparable performance to the original model. We evaluate our results using performance on the forget dataset ↓, performance on the test dataset ↑, and Membership Inference Attack (MIA) ↓, which measures the attacker's ability to distinguish the forget dataset from the training dataset. Our model outperforms the existing approaches that aim to reduce MIA and the performance on the forget dataset while keeping an equivalent performance on the test set. Specifically, our approach reduces MIA by 0.202 and decreases AUC and F1 scores on the forget set by 0.221 and 0.305, respectively. Additionally, our performance on the test set matches that of the retrained model, while allowing forgetting. Code is available at https://github.com/BioMedIA-MBZUAI/Forget-MI.git.

Keywords: Machine Unlearning · Multimodal Data · Clinical Data · Privacy Preserving · Right to Be Forgotten

1 Introduction

The wide and fast-paced applications of artificial intelligence (AI) may, in many cases, necessitate the use of private data. This usage threatens the privacy and

S. Hardan and D. Taratynova—These authors contributed equally to this work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 204–213, 2026.
https://doi.org/10.1007/978-3-032-04947-6_20

security of individuals or organizations. To address such concerns, the "right to be forgotten" is legislated in many countries around the world [9,19,22]. This requirement aims to guarantee the removal of individual data upon request as well as its effect on model learning, leading to the emergence of the field of machine unlearning. In healthcare settings, the use of medical data in AI models is highly restricted due to its sensitivity and confidentiality. Therefore, machine unlearning can contribute significantly to addressing patient privacy concerns and the robustness of de-identification methods [26]. For instance, machine unlearning can assure the removal of patient data and fulfill withdrawal requests from subjects involved in clinical research [4].

The goal of machine unlearning is to ensure that the unlearned model achieves a precision comparable to that of the original model trained on the full dataset [3], while forgetting the data requested for removal. The process of machine unlearning is illustrated in Fig. 1. Primarily, we have 1) a dataset containing patient data of different modalities that collectively contribute to learning about the patient and 2) an original model trained on the full patient dataset (Step 1, Fig. 1). When a patient requests the removal of their data from the model, we create the forget dataset containing all used modalities. The patient's data is forgotten by unlearning it from the original model, from which we obtain the unlearned model (Step 2, Fig. 1). To evaluate the effectiveness of unlearning, we have to ensure that 1) the unlearned model's performance is close enough to the original model's performance and 2) any attack to retrieve forgotten data from the unlearned model is unsuccessful (Step 3, Fig. 1).

However, it is challenging to confirm the erasure of the forget dataset due to the lack of explainability of most architectures and the randomness in models. One method, known as exact unlearning, involves retraining the model from scratch while excluding the forget data [5,12,27]. However, this comes at a disadvantage because 1) retraining the model is expensive and time-consuming and 2) there is no guarantee to still have access to the training dataset. On the other hand, approximate unlearning efficiently modifies the existing model to avoid the drawbacks of retraining [26]. Furthermore, medical data is collected from multiple sources and in different types, such as images and clinical notes [1,14]. Thus, the fusion of these modalities, a common practice, motivates the need for multimodal unlearning techniques to address the complexities of multimodal approaches. Combining these valuable aspects, our study explores machine unlearning within the context of multimodal medical data.

Among the notable frameworks used for unlearning, teacher-student frameworks have emerged as an effective approach, in which the student model is guided by the teacher through distillation or regularization, ensuring that the student retains overall performance while unlearning the specific data points to be forgotten [8,17,28]. Other machine unlearning methods identify and update only the parameters most affected by the forget set [13,20,24].

In multimodal settings, prior machine unlearning methods focused on modality decoupling in image-text systems [7,10,16,18]. Our work is partly inspired by [7], which proposed breaking associations between modality pairs marked for

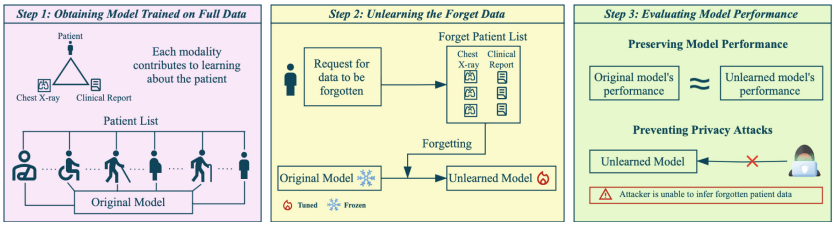


Fig. 1. The original setting contains a model trained on multimodal patient data. When a patient requests their data to be forgotten, we create the forget dataset. This dataset is then forgotten from the original model, obtaining the unlearned model. The unlearned model 1) prevents access to the forgotten data and 2) preserves the original model's performance on the retained data.

deletion while preserving unimodal knowledge. However, this unimodal retention is problematic in healthcare settings. Therefore, a new method must be designed to erase both unimodal and multimodal patient information. Accordingly, we summarize our contributions as follows:

- We introduce, *Forget-MI*, a novel method capable of efficiently unlearning unimodal and joint embeddings from multimodal architectures.
- We are the first to introduce multimodal unlearning in healthcare where we account for the multiple studies per patient in medical datasets.
- We propose a system of loss functions that incorporate noise to help control the distance between the unlearning and original model embeddings to tackle challenges of narrow embedding distributions in medical data.
- We evaluate *Forget-MI* by testing its robustness against attacks, ensuring privacy protection while largely preserving performance on retained data across different unlearning levels.

2 Methodology

Problem Formulation: We introduce, *Forget-MI*, a novel method that unlearns joint and unimodal representations, ensuring patient data is forgotten from the trained model. Given an original model $\mathcal{F}_{\mathbf{og}}$ trained on the full dataset D, we aim to unlearn a subset of data points, resulting in an unlearned model $\mathcal{F}_{\mathbf{ul}}$ that forgets this information while retaining the remaining knowledge from the original model. In order to start unlearning from $\mathcal{F}_{\mathbf{og}}$, we extract the set of points we want to unlearn, forming the forget set $D_f \in D$, then the remaining datapoints form the retain set $D_r = D - D_f$. Starting with $\mathcal{F}_{\mathbf{ul}} = \mathcal{F}_{\mathbf{og}}$, we aim to lead our desired model $\mathcal{F}_{\mathbf{ul}}$ to perform on D_f similar to the way that the original model $\mathcal{F}_{\mathbf{og}}$ would perform on unseen data.

Our unlearning process adjusts the model's loss function in a way that promotes forgetting D_f while preserving the information in D_r, as described in Fig. 2. We achieve this by computing four losses and backpropagating their sum.

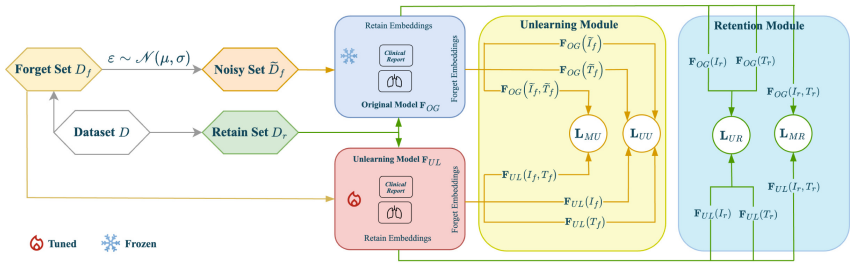


Fig. 2. *Forget-MI*'s Approach: The dataset has images I and text T, and is split into retain D_r and forget D_f, with noise added to the D_f to create $\tilde{D}_f$. We have two models: original $\mathcal{F}_{\mathbf{og}}$ and unlearning $\mathcal{F}_{\mathbf{ul}}$, and we start with $\mathcal{F}_{\mathbf{ul}} = \mathcal{F}_{\mathbf{og}}$. Then, $\mathcal{F}_{\mathbf{ul}}$ unlearns by computing the total loss that consists of: unimodal unlearning $\mathcal{L}_{UU}$, multimodal unlearning $\mathcal{L}_{MU}$, unimodal retention $\mathcal{L}_{UR}$, and multimodal retention $\mathcal{L}_{MR}$. The unlearning losses induce forgetting D_f, while the retention losses preserve the knowledge on D_r.

The losses are defined below, where $[\mathcal{F}(T), \mathcal{F}(I)]$ is a concatenation of the unimodal embeddings produced by a model $\mathcal{F}$, and $\mathcal{F}(T, I)$ is the joint embedding:

Unimodal Unlearning (UU): This loss targets forgetting the unimodal encodings within the forget set D_f. We enable forgetting a neighborhood around the image-text pair $(I_f, T_f) \in D_f$ by adding noise to form $(\tilde{I}_f, \tilde{T}_f)$. So, the UU loss pushes $\mathcal{F}_{\mathbf{ul}}$'s unimodal embeddings on (I_f, T_f) away from $\mathcal{F}_{\mathbf{og}}$'s unimodal embeddings on $(\tilde{I}_f, \tilde{T}_f)$. This distance is computed as follows:

$$\mathcal{L}_{UU} = -Dist\Big([\mathcal{F}_{\mathbf{ul}}(I_f), \mathcal{F}_{\mathbf{ul}}(T_f)], [\mathcal{F}_{\mathbf{og}}(\tilde{I}_f), \mathcal{F}_{\mathbf{og}}(\tilde{T}_f)]\Big). \tag{1}$$

Multimodal Unlearning (MU): This loss targets the edges between modalities in the forget set D_f, by achieving the same functionality of the UU loss but on the joint embeddings to break the correlation between modalities. We set $(I_f, T_f) \in D_f$ as the modality pair to be unlearned from $\mathcal{F}_{\mathbf{og}}$, and $(\tilde{I}_f, \tilde{T}_f)$ as its noisy counterpart. To encourage the joint embedding of $\mathcal{F}_{\mathbf{ul}}$ to dissociate from the joint embedding of $\mathcal{F}_{\mathbf{og}}$, we define the MU loss by the following formula:

$$\mathcal{L}_{MU} = -Dist\Big(\mathcal{F}_{\mathbf{ul}}((I_f, T_f)), \mathcal{F}_{\mathbf{og}}((\tilde{I}_f, \tilde{T}_f))\Big). \tag{2}$$

Unimodal Retention (UR): While unlearning D_f, some information from D_r may also be forgotten, resulting in unintended information loss. To tackle that, UR loss helps the model maintain its performance on the image and text unimodal representation of the retain set. We set $(I_r, T_r) \in D_r$ as the modality pair in the retain set. We want the unimodal encodings in $\mathcal{F}_{\mathbf{og}}$ to be preserved in $\mathcal{F}_{\mathbf{ul}}$, so we define the UR loss by the following formula:

$$\mathcal{L}_{UR} = Dist\Big([\mathcal{F}_{\mathbf{ul}}(I_r), \mathcal{F}_{\mathbf{ul}}(T_r)], [\mathcal{F}_{\mathbf{og}}(I_r), \mathcal{F}_{\mathbf{og}}(T_r)]\Big). \tag{3}$$

Multimodal Retention (MR): Following the UR loss, we retain the joint embeddings in the unlearned model $\mathcal{F}_{\mathbf{ul}}$. We set $(I_r, T_r) \in D_r$ as the modality pair from the retain set. We want $\mathcal{F}_{\mathbf{ul}}$ to maintain the same precision on (I_r, T_r) as $\mathcal{F}_{\mathbf{og}}$, so we define the MR loss by the following formula:

$$\mathcal{L}_{MR} = Dist\Big(\mathcal{F}_{\mathbf{ul}}((I_r, T_r)), \mathcal{F}_{\mathbf{og}}((I_r, T_r))\Big). \tag{4}$$

Total Loss: These losses formulate our objective of unlearning the data in D_f while retaining the desired information in D_r. Unlearning is achieved by backpropagating on the total loss as follows:

$$\mathcal{L} = w_{uu}\mathcal{L}_{UU} + w_{ur}\mathcal{L}_{UR} + w_{mu}\mathcal{L}_{MU} + w_{mr}\mathcal{L}_{MR}, \tag{5}$$

where $w_{uu} + w_{ur} + w_{mu} + w_{mr} = 1$.

3 Experimental Setup

Dataset: We conduct experiments on the MIMIC Chest X-ray (MIMIC-CXR) database [15]. We work on a subset of MIMIC-CXR, adopted from [6], which contains 6,742 images and radiology reports belonging to 1,663 subjects. Our data is labeled according to the four stages of edema: no edema (43%), vascular congestion (25%), interstitial edema (22%), and alveolar edema (10%), leading to a multi-classification problem.

Implementation Details: Since access to the full training dataset is not always feasible, explicitly defining $D_r = D - D_f$ is problematic. Accordingly, in our experiments, the size of the retain set is fixed to match that of the forget set at each epoch, and a subset is constructed by sequentially sampling points from the available retained data. Unlearning begins with an original model that has been trained on the full dataset. We adopt from [6] a multimodal architecture that uses late fusion and encodes CXR images using ResNet and radiology reports (text) using a pre-trained SciBERT model [2]. Furthermore, for the model to forget a neighborhood of the desired data points, we apply noise to the forget set, as denoted in Eqs. (1) and (2). We handle noise on the image and text levels differently. For images, we apply Gaussian noise [10] with hyperparameters μ and σ, representing mean and standard deviation of the noise distribution. We experiment on a set of $\mu = \{0, 0.1, 0.2\}$ and $\sigma = \{0.1, 0.2\}$. For text, we add noise at a character level and a word level randomly [21]. After experimenting with different levels, we observe that the noise has to be small to avoid distorting the model's performance.

In addition, we construct joint embeddings for unlearning using a multimodal adaptation gate that represents the two modalities while maintaining the visual as the dominant one [23]. There are two learning rates used in the unlearning experiments: $1e-4$ and $1e-5$, for 30 epochs. The distance used in Eqs. (1)–(4) is the Euclidean distance.

Forming Forget Datasets: In a healthcare setting, we address cases where patients request the removal of *all* their data, which may encompass multiple studies per patient. Our forget sets include 3%, 6%, and 10% of the original dataset size, aiming to study the effect of the forget set size on the performance of the unlearning model. To ensure fair sample selection, the forget datasets preserve the overall distribution of study counts per patient, including both patients with multiple studies and those with fewer studies.

Evaluation: Following previous works [7,8,10], we employ standard metrics to evaluate the unlearning process. The first metric is **distance** ↓, which is inspired by differential privacy to quantify the difference between the unlearned and original models on the retain and test samples. The second metric is the Membership Inference Attack (**MIA** ↓) in a black-box setting, where we only have access to the model outputs. Following [17], we train an SVM classifier with losses from the retain samples as positive examples and losses from the test samples as negative examples. Finally, we evaluate the unlearned model's performance using **macro-F1** score and **AUC** on two datasets: the forget set (D_f ↓) and the test set (D_t ↑).

4 Results and Disucssions

We compare *Forget-MI* with modality-agnostic approaches, such as NegGrad+ [17], which fine-tunes the original model on D_r while negating the gradient for D_f; SCRUB [17], a teacher-student approach where the student model is trained to selectively disobey the teacher on D_f while aligning on D_r; Catastrophic Forgetting (CF-k) [11], where the first $k = 2$ layers of the original model are frozen and the remaining layers are fine-tuned on D_r; and Exact Unlearning (EU-k) [11], where the unfrozen layers are retrained on D_r. From a multimodal perspective, we evaluate MultiDelete [7], which retains multimodal and unimodal information on D_r while only unlearning D_f on a multimodal level.

Aside from the baselines, we compare *Forget-MI* with a retrained model. We conduct experiments where we explore the effect of different noise and weight settings on the unlearning results. Our experimental settings include: 1) No Noise, adding no distortion to the forget set, 2) Equal weights for all losses, 3) higher weights for Multimodal losses w_{mu} and w_{mr}, 4) higher weights for Unimodal losses w_{uu} and w_{ur}, and 5) higher weights for Retention losses w_{ur} and w_{mr}. This methodology is applied consistently across forget set sizes of 3%, 6%, and 10% of the data, with the Gaussian noise having $\mu = 0$ and $\sigma = 0.1$. We compare our best *Forget-MI* setting for each forgetting percentage with the retrained model and the baselines in Table 1. Additionally, we compare the results of different *Forget-MI* settings with the original and retrained models in Table 2.

Our experiments show that *Forget-MI* consistently outperforms baseline methods, achieving the lowest MIA scores while maintaining comparable performance on the test set relative to the original and retrained models.

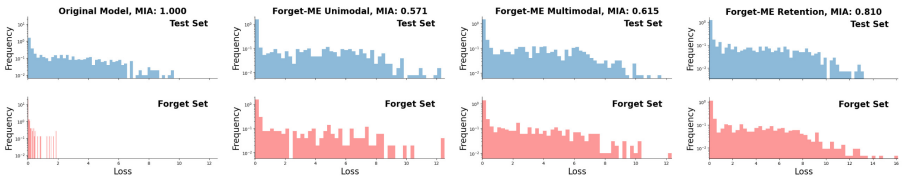
Effect of Forget Percentage: We identify the best *Forget-MI* loss weighting settings for each forgetting percentage. When unlearning 3% of the data, the

Table 1. Comparison between our best *Forget-MI* setting, the retrained model, and the baseline models on forgetting 3%, 6%, and 10% of the data. The Table shows that *Forget-MI* is closer to the retrained model in all cases. The *Forget-MI* experiment used is the best loss weightage setting for each of the percentages, which is then detailed in Table 2. The highlighted results show the best **combination** of metrics.

Method	3%					6%					10%				
	MIA ↓	D_f,AUC ↓	D_f,F1 ↓	D_t,AUC ↑	D_t,F1 ↑	MIA ↓	D_f,AUC ↓	D_f,F1 ↓	D_t,AUC ↑	D_t,F1 ↑	MIA ↓	D_f,AUC ↓	D_f,F1 ↓	D_t,AUC ↑	D_t,F1 ↑
Original	1.000	0.999	0.965	0.677	0.388	1.000	0.999	0.972	0.677	0.388	1.000	0.999	0.970	0.677	0.388
Retrain	0.000	0.566	0.310	0.626	0.362	0.769	0.675	0.395	0.702	0.427	0.190	0.588	0.342	0.629	0.382
NegGrad+	1.000	1.000	0.983	0.678	0.399	1.000	1.000	0.987	0.677	0.388	1.000	1.000	0.986	0.677	0.388
SCRUB	1.000	1.000	0.983	0.679	0.397	1.000	1.000	0.987	0.679	0.394	1.000	1.000	0.985	0.679	0.394
CF-k	1.000	0.999	0.976	0.686	0.416	1.000	0.999	0.977	0.676	0.410	1.000	1.000	0.990	0.668	0.410
EU-k	1.000	0.996	0.952	0.669	0.386	1.000	0.612	0.328	0.631	0.339	1.000	0.629	0.345	0.631	0.339
MultiDelete	0.806	0.858	0.562	0.622	0.300	0.846	0.878	0.604	0.636	0.289	0.952	0.972	0.783	0.656	0.333
Forget-MI	0.571	0.735	0.393	0.625	0.250	0.615	0.654	0.328	0.599	0.270	0.810	0.656	0.313	0.565	0.252

Table 2. *Forget-MI* on forgetting 3%, 6%, and 10% of the data. Here, D_f refers to the forget set and D_t refers to the test set. The *Forget-MI* experiments correspond to higher weighting to the respective losses (e.g., *Forget-MI* Multimodal weights w_{mu} and w_{mr} more), as explained in Sect. 4. The highlighted results show the best **combination** of metrics.

Method	3%					6%					10%				
	MIA ↓	D_f, AUC ↓	D_f, F1 ↓	D_t, AUC ↑	D_t, F1 ↑	MIA ↓	D_f, AUC ↓	D_f, F1 ↓	D_t, AUC ↑	D_t, F1 ↑	MIA ↓	D_f, AUC ↓	D_f, F1 ↓	D_t, AUC ↑	D_t, F1 ↑
Original	1.000	0.999	0.965	0.677	0.388	1.000	0.999	0.972	0.677	0.388	1.000	0.999	0.970	0.677	0.388
Retrain	0.000	0.566	0.310	0.626	0.362	0.769	0.675	0.395	0.702	0.427	0.190	0.588	0.342	0.629	0.382
No Noise	1.000	0.534	0.116	0.508	0.154	1.000	0.511	0.110	0.521	0.090	1.000	0.513	0.104	0.505	0.085
Equal	0.571	0.764	0.385	0.631	0.253	0.846	0.802	0.428	0.637	0.277	0.952	0.796	0.369	0.617	0.240
Multimodal	0.714	0.774	0.384	0.634	0.247	0.615	0.654	0.328	0.599	0.270	0.905	0.740	0.344	0.586	0.231
Unimodal	0.571	0.735	0.393	0.625	0.250	0.615	0.687	0.357	0.610	0.275	0.905	0.764	0.363	0.597	0.236
Retention	0.714	0.646	0.306	0.598	0.245	0.615	0.672	0.339	0.594	0.267	0.810	0.656	0.313	0.565	0.252



(a) Original Model (b) Forgetting 3% (c) Forgetting 6% (d) Forgetting 10%

Fig. 3. The original model vs. best *Forget-MI* at each of forgetting percentages. The subplots show the loss distribution for the forget set relative to the test set. The plots show that, unlike the original model, our model fails at recognizing the test set from the forget set, indicating successful unlearning.

Unimodal setting yields the best results. At 6%, where the MIA score is similar between *Unimodal* and *Multimodal* settings, we consider the *Multimodal* experiment to be outperforming due to its lower performance on the forget set. One reason behind this pattern is that the most distinguishing features in datasets with high inter-patient similarity are not found solely in individual records but

are most uniquely captured in joint modality encodings. Analogously to humans, when our observations are abundant, our ability to remember and learn them increases when we connect our knowledge. On the other hand, when we unlearn 10% of the data, it is best to focus more on the *Retention* losses, as the model tends to compromise more on the performance on the test set due to its objective of forgetting more data.

Evaluation Criterion: Given the trade-off between unlearned model generalizability and the unlearning quality, identifying the optimal combination of metrics is more valuable than the conventional way of evaluation through the test set performance. The evaluation was based on the distance between the retrained model and the unlearning model, the MIA score, the performance on the forget set, and the performance on the test set. One way to demonstrate the efficiency of MIA-based unlearning assessment, according to [25], is to explore the loss distributions in the test and forget datasets of the original and each of the best unlearning models for the forgetting percentages, as shown in Fig. 3. As illustrated, the original model can recognize the difference in the test and forget loss distributions. However, the unlearned models fail to do so, indicating that they no longer retain information from the forget set.

In addition, when observing very similar MIA scores for different configurations under the same forgetting percentage, we compare their performance on the forget and test sets. While we expect the model to struggle at evaluating the forget set, it should not come at the cost of low generalizability on the test set, making the model less useful after unlearning.

Baseline Comparison: We notice a clear difference between modality-agnostic and modality-aware approaches. Other than MultiDelete [7], the baselines showed unsuccessful unlearning. This is explained by the fact that their average distance to the retrained model is 0.323, while the original model's distance is 0.251, indicating that modality-agnostic approaches do not push the model further away from the original. This is opposed to 0.483, 0.410, and 0.561 for our best configurations when forgetting 3%, 6% and 10% of the data, respectively. On the other hand, MultiDelete [7] provides higher F1 than *Forget-MI*. Due to the high data imbalance in this multi-classification problem, *Forget-MI*'s superior ability to forget compromises on the performance on the low represented classes, causing a lower F1.

Effect of Noise: We further demonstrated that introducing small Gaussian noise to the forget set significantly enhances the unlearning performance. Noiseless experiments resulted in the poorest outcomes due to the limited influence of isolated points on the broader data distribution. Thus, the added noise mitigates this by creating a neighborhood around forgotten data points, facilitating broader shifts in the representation of the model. Moreover, increasing the standard deviation consistently raises the MIA score, increases the forget set's performance, and reduces generalization. A higher mean decreases the MIA score and reduces the test set's generalization.

Unlearning Time: The computational cost of unlearning varies substantially across methods. Full retraining is the most time-intensive, requiring up to 14 h. Among the unlearning approaches, NegGrad+, SCRUB, CF-k and EU-k complete within 4 h. Forget-MI and MultiDelete require longer runtimes, taking up to 5 h depending on the forget percentage. This trade-off in runtime is acceptable given the improved unlearning efficacy observed in these methods.

5 Conclusion

The robust performance of *Forget-MI* represents a significant step toward practical machine unlearning in sensitive domains like healthcare, effectively balancing unlearning performance, model utility, and privacy concerns. Generally, we observe that the model is effectively able to unlearn different percentages of data that we aim to forget, and it becomes more challenging as the percentage increases. This suggests inherent limitations in current unlearning algorithms' ability when faced with larger data removal requests. Moreover, future work in the healthcare field can extend to other clinical tasks and experiment with foundation models, which require high generalizability on unseen data, making forgetting more difficult.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. Nat. Med. **28**(9), 1773–1784 (2022)
2. iv: Beltagy, I., Cohan, A., Lo, K.:Scibert: Pretrained contextualized embeddings for scientific text. CoRR abs/ arXiv: 1903.10676 (2019)
3. Bourtoule, L., et al.: Machine unlearning. CoRR, abs/ arXiv: 1912.03817 (2019)
4. Broekstra, R., Aris-Meijer, J., Maeckelberghe, E., Otten, S.: Motives for withdrawal of participation in biobanking and participants' willingness to allow linkages of their data. Euro. J. Human Genet. **30**(1–11), 11 (2021)
5. Cao, Y., Yang, J.: Towards making systems forget with machine unlearning. In: 2015 IEEE Symposium on Security and Privacy, pp. 463–480 (2015)
6. Chauhan, G., et al.: Joint modeling of chest radiographs and radiology reports for pulmonar edema assessment. CoRR, abs/ arXiv: 2008.09884 (2020)
7. Cheng, J., Amiri, H.: Multidelete for multimodal machine unlearning (2024)
8. Chundawat, V.S., Tarun, A.K., Mandal, M., Kankanhalli, M.: Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher (2023)
9. de la Torre, L.: A guide to the california consumer privacy act of 2018. SSRN Electr. J. (2018)
10. Foster, J., Fogarty, K., Schoepf, S., Öztireli, C., Brintrup, A.: Zero-shot machine unlearning at scale via lipschitz regularization. CoRR, abs/ arXiv: 2402.01401 (2024)

11. Goel, S., Prabhu, A., Kumaraguru, P.: Evaluating inexact unlearning requires revisiting forgetting. CoRR, abs/ arXiv: 2201.06640 (2022)
12. Golatkar, A., Achille, A., Polito, M., Soatto, S.: Mixed-privacy forgetting in deep networks, Avinash Ravichandran (2021)
13. Graves, L., Nagisetty, V., Ganesh, V.: Amnesiac machine learning (2020)
14. Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. npj Digital Med. **3**, 12 (2020)
15. Johnson, A., et al.: Mimic-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**, 317 (2019)
16. Kravets, A., Namboodiri, V.: Zero-shot class unlearning in clip with synthetic samples (2024)
17. Kurmanji, M., Triantafillou, P., Hayes, J., Triantafillou, E.: Towards unbounded machine unlearning (2023)
18. Li, J., et al.: Single image unlearning: efficient machine unlearning in multimodal large language models (2024)
19. Mantelero, A.: The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. Comput. Law Sec. Rev. **29**(3), 229–235 (2013)
20. Mehta, R., Pal, S., Singh, V., Ravi, S.N.: Deep unlearning via randomized conditionally independent hessians (2022)
21. Moradi, M., Blagec, K., Samwald, M.: Deep learning models are not robust against noise in clinical text. ArXiv, abs/ arXiv: 2108.12242 (2021)
22. Office of the Privacy Commissioner of Canada. Announcement: Privacy commissioner seeks federal court determination on key issue for canadians' online reputation (Oct 2018)
23. Rahman, W., et al.: Integrating multimodal information in large pretrained transformers. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. Association for Computational Linguistics (July 2020)
24. Shi, J., Ghalyan, N., Gourgoulias, K., Buford, J., Moran, S.: DeepClean: machine unlearning on the cheap by resetting privacy sensitive weights using the fisher diagonal (2024)
25. Triantafillou, E., et al.: Are we making progress in unlearning? findings from the first NeurIPS unlearning competition (2024)
26. Wang, W., Tian, Z., Zhang, C, Yu, S.: A comprehensive survey, Machine unlearning (2024)
27. Wu, Y., Dobriban, E., Davidson, S.: DeltaGrad: rapid retraining of machine learning models. In: Daumé III, H., Singh, A., (eds.) Proceedings of the 37th International Conference on Machine Learning, vol. 119 of Proceedings of Machine Learning Research, 13–18 Jul, pp. 10355–10366. PMLR (2020)
28. Zhou, J.: A unified method to revoke the private data of patients in intelligent healthcare with audit to forget. Nat. Commun. **14**(1) (2023)

Generative Unsupervised Anomaly Detection with Coarse-Fine Ensemble for Workload Reduction in 3D Non-contrast Brain CT of Emergency Room

Jongjun Won[1], Jihwan Kim[1], Joonseo Oh[1], Yereen Yoo[1], Jieun Yum[1], Joonsang Lee[1], Joon Hyung Park[1], Wooyoung Jo[1], Yoojin Nam[1], Hyunki Lee[1], Gil-sun Hong[2(✉)], and Namkug Kim[1(✉)]

[1] Department of Convergence Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea
wjj910@gmail.com, namkugkim@gmail.com

[2] Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, South Korea
hgilsun@gmail.com

Abstract. Neurologic emergencies need to treat unspecified anomalies with various shapes, intensities, and locations in 3D non-contrast brain CT. However, in practice, patients with anomalies take a relatively small portion of total CT volumes. In this situation, excluding unremarkable scans could reduce radiologists' workload. We used a generative unsupervised anomaly detection (GUAD) with 3D Hierarchical Diffusion AutoEncoder (HDAE) model to develop this. In this study, we considered anomalies in two perspectives and made models. One is a Coarse-Morphological anomaly detection Model (CMM), and the other is a Fine-Grained anomaly detection Model (FGM). We ensembled these models' decisions for the exclusion of the unremarkable scans. Models were trained with normal scans of 28,510 from Asan Medical Center (AMC). For evaluation, we mainly used two consecutive test sets of 544 scans from AMC and 1,795 scans from Gangneung Asan Hospital (GNAH). Among clinically significant and unremarkable scans, our study showed [NPV (Negative Predictive Value)/workload reduction] of [98.1%/9.7%] and [96.7%/19.9%] for AMC and GNAH, respectively. Additionally, we used a public dataset (NPV of 98.5%) and five other external hospitals' hemorrhage sets (NPV of 96.0%) to evaluate robustness. Under the reasonable NPV, models showed the potential for workload reduction by omitting unremarkable scans. Compared to individual results of CMM or FGM, the ensembled decision usually shows NPV advantages. Also, with visual results, we observed our model could detect various types of anomalies.

Keywords: 3D Generative Model · Anomaly Detection · Workload Reduction · Coarse-Fine Ensemble

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 214–223, 2026.
https://doi.org/10.1007/978-3-032-04947-6_21

1 Introduction

Generative models (GMs) achieved significant advancements in capturing the underlying data distribution, allowing them to generate realistic data. With various applications of GMs in the medical domain [1, 2], sometimes GMs have been adopted in unsupervised anomaly detection tasks. Diffusion probabilistic models [3, 4] have been used in Generative Unsupervised Anomaly Detection (GUAD) [5–8]. In GUAD, GMs were trained only with normal data. They understood the normal data distribution and could capture the deviation of normal data. When we input normal data, GUAD usually reconstructs scans with little difference, but when we input abnormal data, reconstructed normal-like scans have a significant difference compared to abnormal input. After producing a difference map between input and reconstruction, post-processing yields anomaly scores in areas over the anomaly score threshold.

In the case of neurologic emergencies, various diseases should be treated in different sizes, shapes, locations, intensities, and prevalence rates. Non-contrast brain computed tomography (CT) is usually used as standard screening for fast scanning time. In practice, of the total medical scan readings, the volume of the abnormal group accounts for a smaller portion than that of the normal group. For this reason, excluding normal scans reduces the workload of radiologists. In chest radiographs, some previous studies show that commercial deep-learning software can reduce radiologists' workload by about 8 to 17% [9–12]. Using brain CT modality, we attempted to evaluate the workload reduction in neurological emergencies. We have assumed that GUAD may be an appropriate method for detecting non-specific abnormalities and, in the same sense, able to filter out unremarkable scans with no abnormalities. Also, GUAD has an efficiency that sets variations up as one class of "deviation from the normal group." Therefore, there is no need to build many supervised models for each anomaly. Our goal is to exclude some portion of the normal group by setting the anomaly decision cut-off as low as possible. Of course, achieving high Negative Predictive Values (NPV) is most important.

In this study, we suggest a perspective that anomalies need to be treated as two sides. One is a type of coarse morphological anomaly, and the other is a type of fine-grained anomaly. We build two models of coarse morphological anomaly detection model (CMM) (Fig. 1a) and fine-grained anomaly detection model (FGM) (Fig. 1b). Now, we describe our suggestion with visual results. On one side, we describe the need for CMM. As the Fig. 2a, b, when a model only focuses on reconstructing abnormal "intensity" areas to normal, its reconstruction cannot be considered normal images. Indeed, reconstruction of CMM with modification of the brain's morphology seems like normal images. CMM is necessary to catch morphological anomalies (Fig. 2c. Hydrocephalus, Fig. 2d. Mass effect of hemorrhage). On the other side, we describe the need for FGM. Although the CMM model has a reasonable aspect for anomaly detection, it could change normal brain anatomy (e.g., ventricle or subarachnoid space) as a variation of the normal brain and sometimes yield many false positives (Fig. 2e). Detecting fine-grained anomalies like small infarctions among false positive areas may be challenging. So, when we train FGM, we give hard conditions such as brain anatomy segmentation mask, which consists of subarachnoid space, ventricle, and brain boundary [13]. Hard conditioning makes FGM yield relatively few false positives. As a result, as shown in Fig. 2f, we could set a low anomaly score threshold in the processing

stage and find fine-grained anomalies. In sum, we use complementary models to treat anomalies. CMM is relatively easy to modify input scans' morphology to detect coarse morphological anomalies. FGM detects slight intensity gap anomalies or small-size anomalies. As part of the anomaly decision process (Fig. 1c), we exclude scans predicted as unremarkable from both models simultaneously and consider that they do not need to be referred to radiologists. For the models' architecture (Fig. 1. a, b), we used a 3D expanded hierarchical diffusion autoencoder [14], which consists of a semantic encoder to extract CT's feature and a DDPM [15] U-Net to output the same image as the input CT. HDAE is suitable for our task because it treats images from coarse to fine levels and has high reconstruction performance.

We evaluated our model's ability to distinguish remarkable cases with critical findings in brain CTs from unremarkable cases, which included normal and benign cases. Our model was tested using one internal dataset (AMC: Asan Medical Center) and three external datasets: external dataset 1 (GNAH: Gangneung Asan Hospital), external dataset 2 (an open dataset from the 2019 RSNA Brain Hemorrhage Challenge [16]), and external dataset 3 (a multi-institutional dataset of brain hemorrhages). Notably, the internal dataset and external dataset 1, both collected consecutively from emergency departments, represent real-world clinical data. To evaluate the clinical utility of our model, we focused on negative predictive value (NPV) as the primary evaluation metric and analyzed workload reduction (WLR) across various anomaly decision thresholds. Additionally, we assessed the benefits of using a decision ensemble of FGM and CMM to improve NPVs. Our code is available at https://github.com/Krying/WLR_ANO_3D. In summary, our study makes two key contributions:

- We demonstrate the potential of GUAD for workload reduction by excluding cases without critical findings using real-world datasets
- We propose an anomaly detection approach that considers coarse morphological and fine-grained anomalies through qualitative assessment.

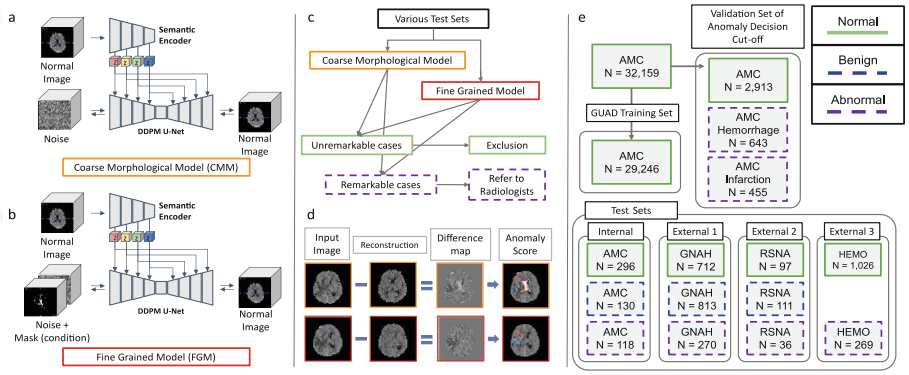


Fig. 1. The overall workflow of our study. (a) and (b) are the model architectures (HDAE) of CMM and FGM. (c) is the process of making anomaly decisions using FGM and CMM. (d) is the process of deriving anomaly scores with the difference map between input and reconstruction images. (e) is the data flow of the train, valid (for anomaly decision cut-off), and various test sets.

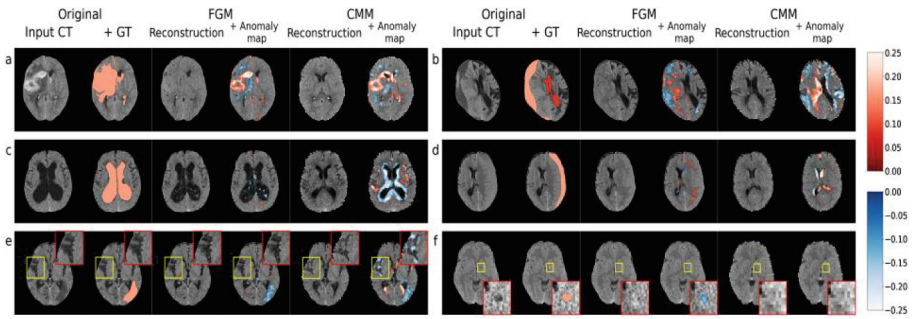


Fig. 2. Various cases of anomalies. (a) intracranial hemorrhage, (b) hemorrhage with hydrocephalus, (c) hydrocephalus, (d) subdural hemorrhage, (e) infarction, (f) hypoxic brain injury. Each case shows the original with lesion mask, reconstruction of FGM and CMM with anomaly map. The red and blue map indicates the hyper and the hypo intensity anomalies, respectively.

2 Material and Methods

2.1 Study Population

In Fig. 1e, we summarized the data used for training, validation, and test. All training and validation data were collected retrospectively (from Jan. 2000 to Aug. 2018) from AMC. Among 32,159 normal scans, 29,246 were used for GUAD model training (age: 44.2 ± 18.6 (Mean ± SD) years; female: 53.5%). The remaining 2,913 normal cases, 643 hemorrhage cases, and 455 infarction cases form a validation set. This validation set was used to determine the cut-off of the anomaly decision. Test datasets consist of one internal test set and 3 external test sets. Test sets of internal (AMC) and external 1 (GNAH) were collected consecutively Feb. 2019 (one month) and from Jan. to May 2019 (3 months), respectively. For external dataset 2, we randomly selected normal cases of 97 from any-hemorrhage group among RSNA dataset [16]. Resembling the class ratio of external dataset 1, we randomly selected benign of 111 and abnormal of 36 cases. External dataset 3 comprised five other external hospitals that had hemorrhage and normal cases. Although this dataset is unsuitable for our theme, which aims to treat various diseases, we used it to confirm our models' robustness for other CT scanners (GE, SIMENS, PHILIPS, and TOSHIBA). The internal set included participants aged 58.7 ± 17.9 years, of whom 51.5% were female. External set 1 had ages of 61.5 ± 17.5 years with 47.7% female participants, and external set 3 had ages of 53.8 ± 18.1 years with 49.7% female participants. Abnormal scans consist of cases with hemorrhage, infarction, mass, hydrocephalus, and others.

2.2 Pre-processing and Post-Processing for Anomaly Scoring

For preprocessing, the Brain Extraction Tool [17] was applied, and we performed depth-wise padding or crop to set all the scans' depth as 32 and then resized to 256 × 256 × 32 (for FGM) and 96 × 96 × 32 (for CMM) respectively. Intensity normalization was performed from Hounsfield units [-10, 90] to [0, 1]. For postprocessing of anomaly scoring, in the difference map, logits under HU 7.5 (FGM) and 10 (CMM) were removed to suppress minor variations, and median smoothing (radius = 1) was applied twice. Small objects with fewer than 75 voxels were removed, and all the absolute values of voxels were summed up.

2.3 Experiment Details

Here, x is the input image, x_t is a noisy image, Z_{enc} are feature vectors, and p is the reverse process [15], HDAE [14] consists of a semantic encoder $Z_{enc} = Enc(x)$ and a conditional Denoising Diffusion Implicit Model (DDIM) [18] $p(x_{t-1} \mid x_t, Z_{enc})$ Instead of DDIM, we used the Denoising Diffusion Probabilistic Model (DDPM) [15]. The semantic encoder extracts semantic vectors from the input images, which are then fed into the corresponding DDPM U-Net layers as condition hierarchically. For the semantic conditioning method, we followed Diffusion AutoEncoder [19] and HDAE as adaptive group normalization.

For training FGM, we gave the condition of the segmentation mask (C_{mask}) to DDPM $p(x_{t-1} \mid x_t, Z_{enc}, C_{mask})$. We used a linear beta noise scheduler $T = 1000$ and only used 550 timesteps. Because our task is just the reconstruction of input image as a normal-like image, we did not need to focus on noise part $T > 550$. Where ϵ is Gaussian noise, and t is the time of a Gaussian diffusion process, we used the loss function defined in Eqs. (2), and (3) for noise prediction, like DDPM. More mathematical formulas can be found in [15, 19]. For inference, we started inference from T = 500 (FGM) and T = 400 (CMM) to preserve patients' identical traits as possible. FGM was trained using normal scan size of 256 × 256 × 32, 10 epochs, learning rate (lr) of 3e-5, and batch size of 3. CMM was trained using normal scan size of 96 × 96 × 32, 22 epochs, lr of 4e-5, and batch size of 8. Experimentally, resolution of 96 × 96 × 32 is enough to find coarse morphological anomalies. Both models used a cosine decay lr scheduler [20] and were trained using an A100 GPU. We used PyTorch framework v2.5.1, CUDA 12.1, and MONAI [21] library v1.4.0.

$$Z_{enc} = Enc(x) \tag{1}$$

$$L_{CMM} = \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{x}_0,\epsilon_t}\left[\left\|\epsilon_\theta\left(x_t, t, z_{enc\ of\ CMM}\right) - \epsilon_t\right\|_2^2\right] \tag{2}$$

$$L_{FGM} = \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{x}_0,\epsilon_t}\left[\left\|\epsilon_\theta\left(x_t, t, z_{enc\ of\ FGM}, C_{mask}\right) - \epsilon_t\right\|_2^2\right] \tag{3}$$

2.4 Evaluation: NPV and Workload Reduction

We used two evaluation metrics NPV and WLR. Metrics' formulation is described in Eqs. (4) and (5). We considered unremarkable cases as negative class.

$$NPV = \frac{(True\ Negative)}{(True\ Negative + False\ Negative)} \tag{4}$$

$$WLR = \frac{(True\ Unremarkable\ scnas)}{(Unremarkable\ scans + Remarkable\ scans)} \tag{5}$$

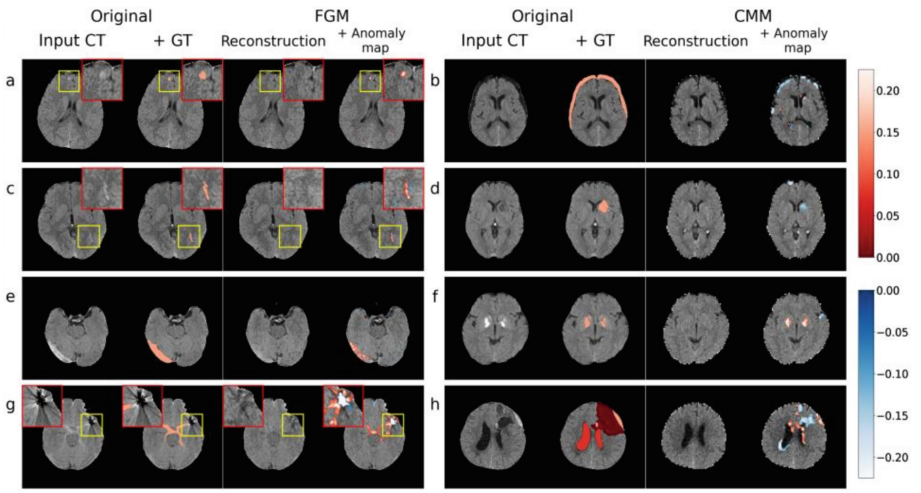


Fig. 3. Various cases of anomalies. (a) cavernoma, (b) subdural hygroma, (c) arteriovenous malformation, (d) infarction, (e) cerebral venous thrombus, (f) metabolic disease, (g) subarachnoid hemorrhage with metal artifact, (h) hemorrhage with hydrocephalus and intraparenchymal mass. Results of (a), (c), (e), and (g) are from FGM, and results of (b), (d), (f), and (h) are from CMM. The red and blue map indicates the hyper and the hypo intensity anomalies, respectively.

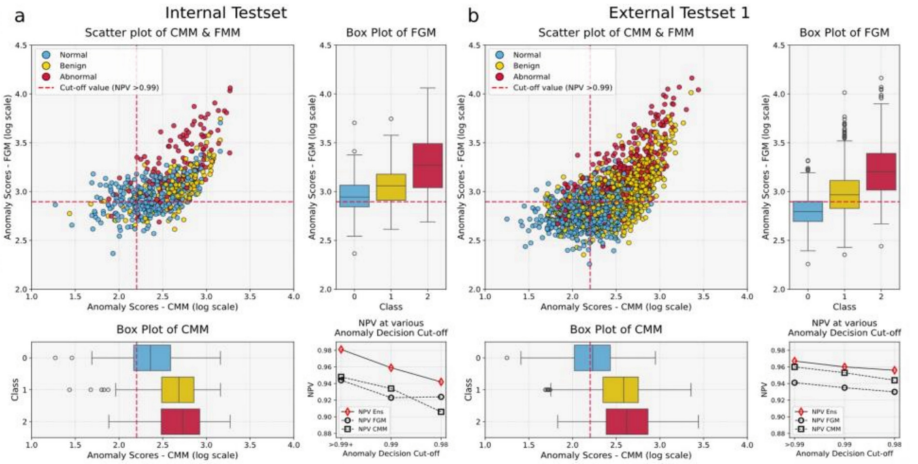


Fig. 4. Scatter plots and box plots of Anomaly Scores and plots of NPVs at various anomaly thresholds. (a) and (b) are the plots of internal and external 1 respectively. Scatter plots indicate the log scaled anomaly scores (x-axis means scores of CMM, and y-axis means scores of FGM). In box plots of anomaly score, class 0, 1, and 2 means 'normal', 'benign', and 'abnormal' respectively. Plots of "NPV at various anomaly decision cut-off values" (lower right of each figure) display the NPVs of FGM, CMM, and ensembled decision. The red markers indicate the NPVs of ensembled decision from CMM and FGM. The black markers indicate each model's NPVs.

Table 1. NPVs and WLRs for test sets at various anomaly decision cut-off (NPVs of 0.98, 0.99, and more than 0.99) from validation set.

Cut-off	Metrics	Internal test	External test 1	External test 2	External test 3
> 99%	NPV	98.1% (53/54)	96.7% (357/369)	98.5% (64/65)	96.0% (360/375)
	WLR	9.7% (53/544)	19.9% (357/1795)	26.2% (64/244)	27.8% (360/1295)
99%	NPV	95.9% (71/74)	96.0% (433/451)	97.1% (68/70)	96.4% (429/445)
	WLR	13.1% (71/544)	24.1% (433/1795)	27.9% (68/244)	33.1% (429/1295)
98%	NPV	94.2% (97/103)	95.6% (503/526)	96.5% (83/86)	96.2% (477/496)
	WLR	17.8% (97/544)	28.0% (503/1795)	34.0% (83/244)	36.8% (477/1295)

Table 2. Comparison of NPVs of FGM, CMM and Ensembled results for test sets at various anomaly decision cut-off (NPVs of 98%, 99%, and more than 99%) from validation set.

Test Sets	Internal test			External test 1			External test 2			External test 3		
Cut-off	>99%	99%	98%	>99%	99%	98%	>99%	99%	98%	>99%	99%	98%
FGM	0.949	0.934	0.906	0.960	0.953	0.944	0.953	0.956	0.927	0.949	0.946	0.939
CMM	0.944	0.923	0.924	0.941	0.935	0.930	**0.989**	**0.971**	**0.971**	0.945	0.944	0.946
Ensemble	**0.981**	**0.959**	**0.942**	**0.967**	**0.960**	**0.956**	0.985	**0.971**	0.965	**0.960**	**0.964**	**0.962**

3 Results and Discussion

3.1 Workload Reduction with Reasonable NPV

We evaluated the NPV and WLR of our models at different anomaly decision cut-offs (98%, 99%, and > 99%) across internal and external test datasets (Table 1). For the highest threshold (>99%), the NPV across the four test sets ranged from 96.0% to 98.5%, while the corresponding WLR varied between 9.7% and 27.8%. As expected, lowering the anomaly decision cut-off to 99% led to a slight decrease in NPV (ranging from 95.9% to 97.1%), but with an increase in WLR (13.1%–33.1%). Further reducing the cut-off to 98% resulted in a more pronounced trade-off, with NPVs between 94.2% and 96.5%, accompanied by WLRs of 17.8%–36.8%. These results indicate that higher cut-offs ensure higher NPVs but limit the extent of workload reduction. Conversely, lower thresholds yield greater workload reduction but at the cost of reduced NPV.

3.2 The Effect of Model Ensemble

To assess the impact of our ensemble approach, we compared NPVs of the individual models (FGM and CMM) against the ensembled model at different cut-offs (Table 2). For critical findings, excluding external test 2, the ensembled model outperformed the individual models across all thresholds. For the real-world dataset (internal, external dataset 1), at NPV of > 99%, the ensemble achieved an NPV of 98.1% for internal test, compared to 94.9% for FGM and 94.4% for CMM. 96.7% and, for external test 1, the ensemble achieved an NPV of 96.7%, compared to 96.0% for FGM and 94.1% for CMM. The coarse-fine ensemble improved NPV performance over FGM and CMM alone (Fig. 4). These results suggest that combining FGM and CMM enhances model reliability, in identifying critical cases with high NPV. This approach would be particularly valuable in settings where reducing false negatives is paramount, such as emergency departments and acute care settings.

3.3 Balancing Sensitivity and Workload Reduction

An important consideration in deploying AI models for clinical decision support is the balance between sensitivity (ensuring high NPV) and WLR. Our study shows that setting a higher anomaly decision cut-off (e.g. > 99%) maintains NPVs close to 98% while still achieving a reasonable WLR (~10–35%). However, a more relaxed threshold (e.g., 98%) increases WLR (~20–44%) at the cost of a slight NPV decline. Institutions must determine the optimal threshold based on their specific patient population and clinical workflow constraints. These findings highlight the importance of carefully selecting decision thresholds to minimize missed critical cases while maintaining workload efficiency.

3.4 Various Anomaly Cases and Missed Cases

Figure 3 shows the cases of anomaly detection exist in real-world conditions. In real-world conditions, at NPV of > 99%, our model classified 54 cases as unremarkable in the internal test dataset, among which one case was later identified as critical. In the

external dataset, 12 out of 369 cases classified as unremarkable were later found to be critical. This underscores the need for human oversight and potential hybrid models where AI serves as an assistive tool rather than a sole decision-maker.

3.5 Limitation

The primary limitation of our study is the lack of comparative analysis. To the best of our knowledge, no prior studies have specifically addressed workload reduction (WLR) on brain CT. Our study may serve as one of the initial steps in this research direction. Given the novelty of the topic, our primary goal was to demonstrate the clinical relevance and potential impact of applying deep learning in this context. Additionally, the model's inference time per scan was approximately 88 s for CMM and 120 s for FGM on an RTX 3090 GPU (VRAM $<$ 6 GB). While these times are reasonable, there is potential for further optimization.

4 Conclusion

Our model has the potential to significantly alleviate the workload in emergency radiology by prioritizing cases requiring urgent review. The ability to maintain high NPVs across diverse datasets suggests that the model could be deployed across different institutions with minimal performance degradation. Additionally, we suggest a new ensemble approach to enhance the model's performance. Future work should focus on real-time deployment simulations and evaluating the impact on radiologist efficiency. In conclusion, our study provides strong evidence that deep learning-based anomaly detection can serve as a reliable tool for optimizing radiology workflows, improving efficiency, and maintaining diagnostic accuracy in emergency and acute care settings.

Acknowledgments. This research was supported by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI22C0471, RS-2022-KH130266); the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (RS-2024–00355370, RS-2025–00553009); and the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea (2024IP0058).

Disclosure of Interests. The authors have no competing interests.

References

1. Kazerouni, A., et al.: Diffusion models in medical imaging: A comprehensive survey. Med. Image Anal. **88**, 102846 (2023)
2. Yang, L., et al.: Diffusion models: A comprehensive survey of methods and applications. ACM Comput. Surv. **56**(4), 1–39 (2023)
3. Sohl-Dickstein, J., et al. Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. PMLR (2015)
4. Song, Y., et al.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, (2020)

5. Wyatt, J., et al. Anoddpm: anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
6. Bercea, C.I., et al., Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. arXiv preprint arXiv:2305.19643, (2023)
7. Fontanella, A., et al.: Diffusion models for counterfactual generation and anomaly detection in brain images. IEEE Trans. Medical Imaging (2024)
8. Behrendt, F., et al.: Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris. Comput. Biol. Med. **186**, 109660 (2025)
9. Schalekamp, S., et al.: Performance of AI to exclude normal chest radiographs to reduce radiologists' workload. European Radiol., 1–9 (2024)
10. Plesner, L.L., et al.: Using AI to identify unremarkable chest radiographs for automatic reporting. Radiology **312**(2), e240272 (2024)
11. Dyer, T., et al.: Diagnosis of normal chest radiographs using an autonomous deep-learning algorithm. Clin. Radiol. **76**(6), 473. e9–473. e15 (2021)
12. Plesner, L.L., et al.: Autonomous chest radiograph reporting using AI: estimation of clinical impact. Radiology **307**(3), e222268 (2023)
13. Cai, J.C., et al., Fully automated segmentation of head CT neuroanatomy using deep learning. Radiol. Artifi. Intell. **2**(5), e190183 (2020)
14. Lu, Z., et al.: Hierarchical diffusion autoencoders and disentangled image manipulation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2024)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
16. Flanders, A.E., et al.: Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiol. Artifi. Intell. **2**(3), e190211 (2020)
17. Akkus, Z., et al.: Robust brain extraction tool for CT head images. Neurocomputing **392**, 189–195 (2020)
18. Song, J., Meng, C., Ermon, S:. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, (2020)
19. Preechakul, K., et al. Diffusion autoencoders: toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
20. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, (2016)
21. Pinaya, W.H., et al.: Generative AI for medical imaging: extending the MONAI framework. arXiv. arXiv preprint arXiv:2307.15208, (2023)

GLM-SFNet: Global-Local Vision-Mamba with Semantic Fusion for Medical Image Segmentation

Jiahui Chen[1], Fei Qi[1(✉)], Chengyuan Chang[1], Qinjie Hu[1], Kaiwen Fu[1], Xiaotian Wang[1], and Kun Liu[2]

[1] Xidian University, Xi'an, Shaanxi 710126, China
fred.qi@ieee.org
[2] Hebei University of Technology, Tianjin 300401, China

Abstract. Mamba-based architectures have shown promising performance in medical image segmentation. Accurate segmentation demands effective capture and integration of both global context and local details. However, existing methods often lack a balanced approach to extracting and fusing global and local information within the encoder and decoder. To address this issue, we introduce Global-Local Vision-Mamba with Semantic Fusion Network (GLM-SFNet), which is designed for balanced global-local feature processing in medical image segmentation. In the encoder, GLM-SFNet employs a Local-Global Vision State Space block (LGVSS). LGVSS strategically integrates four-directional scanning Mamba to capture comprehensive global context while incorporating Learnable Descriptive Convolution (LDC) to ensure detailed local feature extraction. For the decoder, we propose a Semantic Fusion Decoder (SFD), which achieves enhanced information integration and boundary precision by strategically combining global and local semantic fusion modules. Extensive experiments on three benchmark datasets demonstrate that GLM-SFNet achieves state-of-the-art segmentation performance while maintaining a lightweight architecture.

Keywords: Medical Image Segmentation · Mamba · Attention · Semantic Fusion · Global Context · Local Features

1 Introduction

Segmentation is a fundamental task in medical image analysis [26], providing essential visual references for clinical diagnosis and improving both efficiency and accuracy. It has been widely applied in skin lesion analysis [2,5] and organ segmentation in abdominal CT scans [12].

J. Chen and F. Qi—Equal contributions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 224–234, 2026.
https://doi.org/10.1007/978-3-032-04947-6_22

U-Net [20] represents a breakthrough in medical image segmentation, with its encoder-decoder structure and skip connections inspiring numerous improvements [1,16]. While U-Net and its variants [1] effectively capture local features, convolutional kernels inherently struggle to model long-range dependencies. SwinUNet [3] incorporates Swin Transformer [15,24] to enhance global context modeling but at the cost of detail loss and increased computational overhead. TransUNet [4] and MISSFormer [10] leverage both convolutional and transformer-based approaches.The high computational complexity of models employing Transformer [3,4,10] limits their the clinical applications.

Recently, Mamba [7] has reduced the computational complexity to linear time, while preserving global modeling capabilities for long sequences, as compared to the quadratic complexity of the Transformer [24]. Mamba has been applied to computer vision [8,14,31] for its efficiency. Specifically, Slice-Mamba [6], VM-UNet [21], VM-UNet v2 [30], and UltraLight VM-UNet [28] have adopted Mamba architecture to efficiently learn visual representations in medical image segmentation.

However, focusing mainly on global context modeling, these Mamba-based segmentation models [6,21,28,30] often have limitations in preserving spatial local details, hindering for accurate boundary segmentation. Firstly, in the encoder, employing Mamba for long-range dependency modeling, these methods may not adequately represent local spatial relationships and geometric properties. These local details are crucial for accurately identifying organ boundaries and lesion regions [29]. Although VMamba [14] applies a cross-scan module to mitigate this issue, it still fails to maintain spatial consistency between adjacent pixels. Secondly, concerning the decoder, these methods suffer from poor semantic alignment in feature fusion due to the reconstruction-oriented upsampling. Poor semantic alignment, caused by independent upsampling of high-level features and potential spatial distortion of low-level features, reduces the semantic integrity of the fused high-resolution outputs. Moreover, existing decoders do not effectively integrate global context and local details.

To overcome above issues, we introduce a Global-Local Vision-Mamba with Semantic Fusion Network (GLM-SFNet), which has a Local-Global Vision State Space Block (LGVSS) for enhanced local-global representation and a Semantic Fusion Decoder (SFD) for improved semantic integration. LGVSS combines a four-directional scanning Mamba pathway for long-range dependency modeling with a Learnable Descriptive Convolution (LDC) [9] branch to capture local details. SFD, a plug-and-play decoder, employs Global Cross-scale Fusion (GCF) and Local Cross-scale Fusion (LCF) modules. GCF enhances semantic alignment in deep stages through global cross-attention, and LCF aligns high-level semantic features with low-level spatial features in shallow stages using local cross-attention. The collaborative combination of GCF and LCF within SFD ensures semantic coherence and spatial continuity throughout decoding.

The main contributions of this paper are as follows: (1) We propose LGVSS, which effectively captures long-range dependencies and local details in images. (2) We propose SFD, a plug-and-play medical image segmentation decoder. It

performs synchronized upsampling and feature fusion to enhance the hierarchical flow of semantic information from high-level to low-level features. (3) GLM-SFNet achieves state-of-the-art performance on three public datasets. On the Synapse dataset, it reaches an average DSC of 84.82% and an HD95 of 11.87mm.

2 Method

In this section, we introduce the overall architecture of GLM-SFNet and present details of each component.

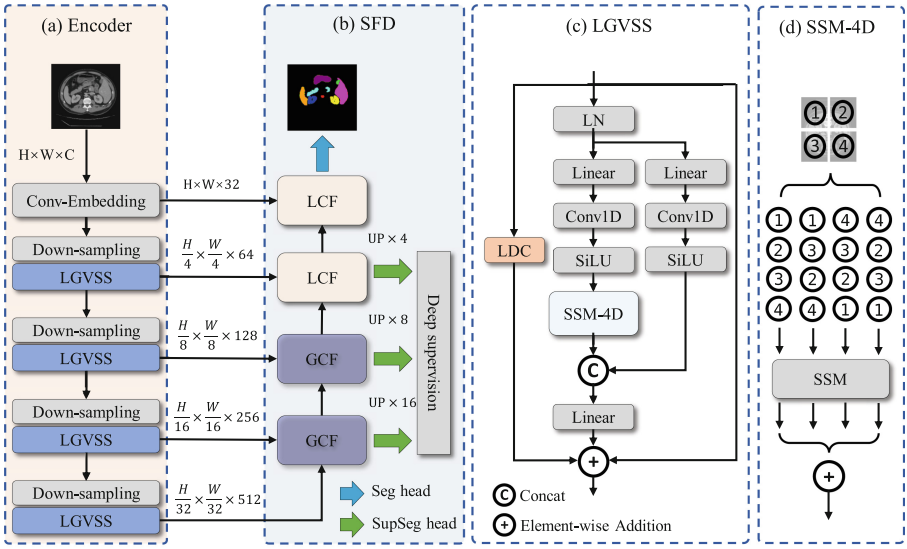


Fig. 1. The overall architecture of GLM-SFNet. (a) A multi-scale encoder with four hierarchical stages. (b) Semantic Fusion Decoder (SFD). (c) Local-Global Vision State Space Block (LGVSS). (d) Four-directional scan State Space Model (SSM-4D).

2.1 GLM-SFNet

As shown in Fig. 1, following the mainstream encoder-decoder design, the proposed GLM-SFNet consists of a Mamba-based encoder and a Semantic Fusion Decoder (SFD). According to Fig. 1(a), the encoder takes an input image $F \in \mathbb{R}^{H\times W\times C}$ and processes it through a Conv-Embedding layer, adjusting the channel dimension to produce a feature map $F' \in \mathbb{R}^{H\times W\times 32}$. This feature map F' is then passed through four consecutive stages. Each stage has a Down-Sampling layer and an LGVSS block, for multi-scale feature extraction. As illustrated in Fig. 1(b), the SFD has four stages symmetric to the encoder. In

its two deeper stages, the SFD applies Global Cross-scale Fusion (GCF) to fuse high-level features with fine semantic alignment. In contrast, Local Cross-scale Fusion (LCF) integrates high-level semantics with low-level spatial features in the two shallower layers. In addition, SFD includes a Seg head to generate the final segmentation masks and three auxiliary SupSeg heads for deep supervision over intermediate outputs.

For reproducibility, we detail the implementation of the aforementioned simple layers. The Conv-Embedding comprises two 3×3 convolutional layers. Each Down-Sampling layer contains a max pooling layer followed by two 3×3 convolutional layers. The SupSeg head comprises a bilinear interpolation layer and a 1×1 convolutional layer, while the Seg head includes a 3×3 convolutional layer, a bilinear interpolation layer, and a 1×1 convolutional layer. The details of LGVSS, GCF, and LCF are provided in subsequent subsections.

2.2 Local-Global Vision State Space Block (LGVSS)

The structure of an LGVSS block is depicted in Fig. 1(c). Compared against the MambaVision Mixer [8], an improved State Space Model (SSM), the LGVSS block has a distinct main pathway, a left-side branch, and a residual connection on the right. The main pathway comprises two structural similar parallel branches. One branch utilizes the four-directional scanning SSM (SSM-4D), as illustrated in Fig. 1(d), to enhance multi-directional long-range dependency modeling. The other symmetric branch compensates for potential information loss and local forgetting caused by sequential scanning by omitting SSM-4D. The outputs of both branches are concatenated and projected through a linear layer. At the left side, the Learnable Descriptive Convolution (LDC) [9] branch captures local spatial structures in the 2D feature map by introducing learnable local descriptors to extract fine-grained texture features. The residual connection stabilizes gradient propagation. Finally, the outputs from all three branches are summed to generate the final output of LGVSS.

2.3 Global Cross-Scale Fusion (GCF)

GCF integrates high-level semantic information between encoder and decoder features within deeper stages of the network. It facilitates effective feature fusion with the Multi-Head Attention (MHA) mechanism [24], as depicted in Fig. 2(a). The inputs are derived from consecutive stages of the encoder and decoder, denoted as $F_{\text{enc}} \in \mathbb{R}^{H \times W \times C/2}$ and $F_{\text{dec}} \in \mathbb{R}^{H/2 \times W/2 \times C}$, respectively.

Initially, Batch Normalization (BN) is applied to normalize features as F'_{enc} and F'_{dec}. These normalized features are subsequently fed into MHA, where F'_{enc} serves as queries (Q), while F'_{dec} provides keys (K) and values (V), each after a global linear embedding. MHA is then applied to obtain fused features, effectively integrating semantic content across scales. Since keys and values span the entire spatial space of feature maps, MHA operates globally. Finally, the fused features are further refined through channel attention and spatial attention [27] to enhance feature representation capability.

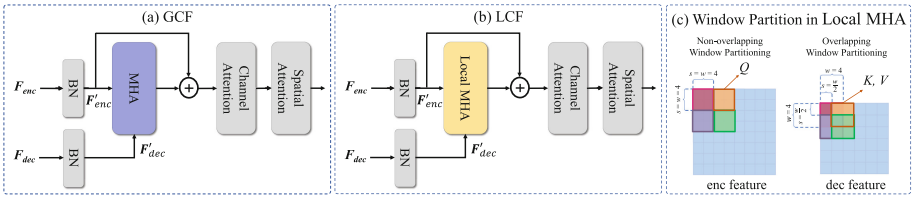


Fig. 2. (a) Global Cross-scale Fusion (GCF). (b) Local Cross-scale Fusion (LCF). (c) Non-overlapping (left) and overlapping (right) window partitioning in Local MHA.

2.4 Local Cross-Scale Fusion (LCF)

LCF aims to enhance semantic fusion in features of shallow stages while achieving fine-grained details. To address challenges of locality and sparsity of attention in shallow-stage features [11], Local MHA is adopted. Local MHA can reduce redundancy and improve fusion efficiency by computing cross-attention only between corresponding local windows of encoder and decoder features.

As shown in Fig. 2(b), LCF adopts a computational process similar to GCF, but replaces MHA with Local MHA as its core component. To apply Local MHA, encoder features F'_{enc} are partitioned into non-overlapping windows of size $w \times w$ with a stride of $s = w$. The window size w is set to 4 when the spatial sizes ratio between F'_{enc} and F'_{dec} is two, whereas if the ratio is four, w is adjusted to 8. These windows are then linearly embedded to produce queries (Q). In parallel, decoder features F'_{dec} are partitioned into overlapping windows of a fixed size $w = 4$ with a stride $s = w/2$, ensuring alignment between encoder and decoder windows. These decoder windows are then projected through a Linear layer to generate keys (K) and values (V). The processes of non-overlapping and overlapping window partitioning are depicted in Fig. 2(c). This strategic alignment of window partitioning enables low-level spatial features to effectively query and integrate high-level semantic information, thereby capturing the fine-grained details essential for boundary segmentation.

In addition, LCF applies dynamic embedding on normalized decoder features F''_{dec} to derive F'_{dec} through two stacked dynamic convolutions. This process enhances local feature representation and represents another difference from GCF.

3 Experiments and Results

3.1 Datasets and Implementation

Datasets and Metrics. To validate the effectiveness of the proposed network, we conducted experiments on three publicly available datasets: Synapse [12], ISIC2017 [2], and ISIC2018 [5]. For Synapse, we followed the configurations of TransUNet [4] and evaluated performance with the Dice Similarity Coefficient (DSC) and the 95% Hausdorff Distance (HD95). For ISIC2017 and ISIC2018, we used the UltraLight VM-UNet [28] protocol, evaluating performance with DSC,

Sensitivity (SE), Specificity (SP), and Accuracy (ACC). HD95 is measured in millimeters (mm), and other metrics are in percentages (%).

Implementation Details. GLM-SFNet was implemented using Python 3.11, PyTorch 2.3.1, and CUDA 12.1. All experiments were conducted on an NVIDIA RTX 4090 GPU. During training, we applied random rotation and flipping for data augmentation and used the AdamW optimizer. For Synapse, we set the input resolution to 224×224, the batch size to 24, and the maximum number of training epochs to 600. The initial learning rate was 0.001, and the loss function was a weighted combination of Dice loss and cross-entropy loss, with both weights set to 1. For ISIC2017 and ISIC2018, the input resolution was set to 256×256, the batch size to 8, and the maximum number of training epochs to 250. The initial learning rate was 0.01, and the loss function was a weighted combination of BCE loss and Dice loss, with weights of 1 for both.

3.2 Quantitative and Qualitative Segmentation Results

As shown in Table 1, our GLM-SFNet achieved superior performance on the Synapse dataset, with an average DSC of 84.82% and HD95 of 11.87 mm, outperforming all baseline models. Compared with PVT-EMCAD-B2 [19], it achieved a 1.19% DSC improvement and a 3.81 mm HD95 reduction. Compared with VM-UNet [21], it achieved a 3.74% DSC improvement and a 7.34 mm HD95 reduction. For small organs, GLM-SFNet achieved a 4.19% DSC improvement for the gallbladder (GB) and a 2.73% DSC improvement for the pancreas (PC). For organs with complex boundaries, it achieved a 1.39% DSC improvement for the stomach (SM). As shown in Table 2, GLM-SFNet also performed well on the ISIC2017 and ISIC2018 datasets, achieving DSC of 92.18% and 90.64%, which are 1.27% and 1.24% higher than UltraLight VM-UNet [28]. These results confirm that GLM-SFNet can effectively model long-range dependencies while capturing fine-grained local textures, enabling precise segmentation of organ boundaries and lesion areas.

As shown in Fig. 3, we present visual comparisons between GLM-SFNet and other methods on the Synapse dataset. Our method achieves precise boundary delineation for complex-shaped organs, as demonstrated in stomach (SM) segmentation (first row) and liver segmentation (second row). For small organs like pancreas (PC) and kidney right (KR) (third row), GLM-SFNet shows superior performance while effectively addressing over-segmentation and under-segmentation issues. This stems from our encoder's effective modeling of long-range dependencies and accurate capture of local textural details, which provide comprehensive multi-scale features. The decoder further improves performance through cross-scale semantic fusion of hierarchical encoder features, achieving refined boundary accuracy.

Table 1. Comparative experimental results on the Synapse dataset. **Bold** indicates the best performance, and **underline denotes the second-best.**

Model	Average DSC↑	Average HD95↓	Aorta	GB	KL	KR	Liver	PC	Spleen	SM
U-Net [20]	70.11	44.69	84.00	56.70	72.41	62.64	86.98	48.73	81.48	67.96
Att-UNet [17]	71.70	34.47	82.61	61.94	76.07	70.42	87.54	46.70	80.67	67.66
TransUNet [4]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUNet [3]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
MT-UNet [25]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
MissFormers [10]	81.96	18.20	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
PVT-CASCADE [18]	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.69
TransCASCADE [18]	82.68	17.34	86.63	68.48	87.66	**84.56**	94.43	65.33	90.79	83.52
VM-UNet [21]	81.08	19.21	86.40	69.41	86.16	82.76	94.17	58.80	89.51	81.40
SliceMamba [6]	81.95	16.04	87.78	68.77	**88.30**	84.26	95.25	64.49	86.91	79.82
PVT-EMCAD-B2 [19]	83.63	15.68	88.14	68.87	88.08	84.10	**95.26**	68.51	**92.17**	83.92
GLM-SFNet	**84.82**	**11.87**	**88.32**	**74.78**	87.49	84.35	95.14	**71.24**	91.98	**85.31**

Table 2. Comparative experimental results on the ISIC2017 and ISIC2018 datasets. **Bold** indicates the best, and underline denotes the second-best.

Methods	ISIC2017				ISIC2018			
	DSC ↑	SE ↑	SP ↑	ACC ↑	DSC ↑	SE ↑	SP ↑	ACC ↑
U-Net [20]	89.89	87.93	98.12	96.13	88.51	87.35	97.44	95.39
VM-UNet [21]	90.70	88.37	98.42	96.45	88.91	88.09	97.43	95.54
VM-UNet v2 [30]	90.45	87.68	98.49	96.37	89.02	89.59	97.02	95.51
MALUNet [22]	88.96	88.24	97.62	95.83	89.31	88.90	97.25	95.48
LightM-UNet [13]	90.80	88.39	98.46	96.49	88.98	88.29	97.65	95.55
EGE-UNet [23]	90.73	89.31	98.16	96.42	88.19	90.09	96.38	95.10
UltraLight VM-UNet [28]	90.91	90.53	97.90	96.46	89.40	86.80	**97.81**	95.58
GLM-SFNet	**92.18**	**90.87**	**98.53**	**97.08**	**90.64**	**90.59**	97.50	**96.04**

3.3 Ablation Study

Table 3 presents the ablation experiments conducted on the Synapse dataset to validate the effectiveness of SFD. We replaced the decoders of PVTv2-EMCAD-B0, PVTv2-EMCAD-B2 [19], and SwinUNet [3] with SFD and compared their performances. The results demonstrated that SFD improved DSC of the three models by 1.2%, 0.79%, and 3.47%, and reduced the HD95 by 0.29 mm, 2.1 mm, and 7.03 mm, respectively. These results confirmed the advantages of SFD in cross-scale semantic fusion and delineate complex boundary shapes.

Table 4 evaluates the impact of the LDC branch and the number of Scanning Directions (#SD) in LGVSS. Firstly, without LDC, unidirectional, bidirectional, and four directional (4D) scanning were applied to analyze the effect of the number of SDs. The results showed that increasing the number of SDs effectively enhanced the performance of LGVSS, while the number of parameters remained unchanged due to shared scanning components, and GFLOPs slightly increased.

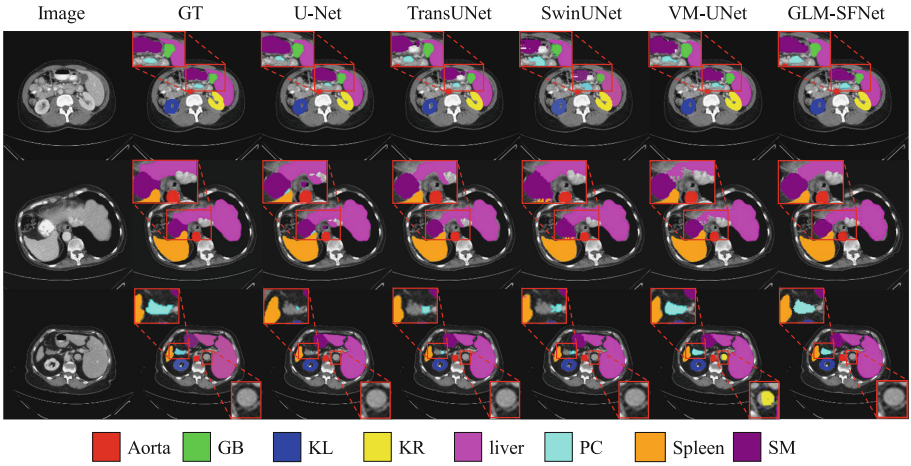


Fig. 3. Qualitative results of different methods.

Table 3. Effectiveness of SFD (ours) on the Synapse dataset.

Encoders	PVTv2-B0		PVTv2-B2		SwinUNet Encoder	
Decoders	EMCAD	SFD	EMCAD	SFD	SwinUNet Decoder	SFD
DSC	81.97	**83.14**	83.63	**84.42**	79.13	**82.60**
HD95	17.39	**17.10**	15.68	**13.58**	21.55	**14.52**

Table 4. Effect of different components in LGVSS.

LGVSS		DSC	Params(M)	GFLOPs
LDC	#SD			
No	1	83.18	7.74	4.16
No	2	83.53	7.74	4.17
No	4	84.05	7.74	4.18
Yes	4	**84.82**	14.35	4.64

Table 5. Impact of the global and local component combination in SFD.

SFD stages				DSC	Params(M)	GFLOPs
L	L	L	L	75.77	17.64	4.64
L	L	L	G	77.67	15.04	4.62
L	L	G	G	**84.82**	14.35	4.64
L	G	G	G	83.27	14.16	4.90
G	G	G	G	80.35	14.10	14.81

Adopting LDC with the 4D scanning scheme, DSC was further improved by 0.77%, with moderate increase in parameters and minor rise in GFLOPs.

Table 5 further analyzes the different combinations of LCF (L) and GCF (G) in SFD to determine the optimal decoding configuration. The table lists the four stages of the decoder from shallow to deep layers from left to right. The experimental results revealed that deep-layer features contain rich semantic information, while shallow-layer features exhibit abundant local details, both are crucial for segmentation. Ultimately, we found that the LLGG configuration achieved the best balance between performance and computational efficiency.

4 Conclusions

In this study, we introduced the Global-Local Vision-Mamba with Semantic Fusion Network (GLM-SFNet), designed to achieve a balanced integration of global contexts and local details in medical image segmentation. Our innovation lies in both encoder and decoder. In encoder, the Local-Global Vision State Space (LGVSS) block enables comprehensive global context capture and detailed local feature extraction. The Semantic Fusion Decoder (SFD) enhances information integration and boundary precision. Extensive experiments on three public datasets demonstrate that this balanced design improves segmentation performance over existing state-of-the-art methods. Our work demonstrates that a balanced encoder-decoder architecture is critical in achieving better segmentation results.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (grant numbers. 62176201 and 62173124), and the Beijing-Tianjin-Hebei Collaborative Basic Research Special Project (grant number F2024202102).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azad, R., et al.: Medical Image Segmentation Review: The success of U-Net (Nov 2022)
2. Berseth, M.: ISIC 2017 - Skin Lesion Analysis Towards Melanoma Detection (Mar 2017)
3. Cao, H., et al.: Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. In: ECCV 2022 Medical Computer Vision Workshop. LNCS, vol. 13803, pp. 205–218. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25066-8_9
4. Chen, J., et al.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation (Feb 2021)
5. Codella, N., al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
6. Fan, C., Yu, H., Huang, Y., Wang, L., Yang, Z., Jia, X.: Slicemamba with neural architecture search for medical image segmentation. arXiv preprint arXiv:2407.08481 (2024)
7. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
8. Hatamizadeh, A., Kautz, J.: Mambavision: a hybrid mamba-transformer vision backbone. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 25261–25270 (2025)
9. Huang, P.K., Ni, H.Y., Ni, Y.Q., Hsu, C.T.: Learnable descriptive convolutional network for face anti-spoofing. In: 33rd British Machine Vision Conference (BMVC). BMVA Press, London, UK (2022-11-21/2022-11-24)
10. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: MISSFormer: an effective transformer for 2D medical image segmentation. IEEE Trans. Med. Imaging **42**(5), 1484–1494 (2023)

11. Jiao, J., et al.: Dilateformer: multi-scale dilated transformer for visual recognition. IEEE Trans. Multimedia **25**, 8906–8919 (2023)
12. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-atlas Labeling Beyond Cranial Vault—Workshop Challenge, Munich, Germany, vol. 5, p. 12 (2015)
13. Liao, W., Zhu, Y., Wang, X., Pan, C., Wang, Y., Ma, L.: Lightm-unet: mamba assists in lightweight unet for medical image segmentation. arXiv preprint arXiv:2403.05246 (2024)
14. Liu, Y., et al.: VMamba: visual state space model. Adv. Neural. Inf. Process. Syst. **37**, 103031–103063 (2025)
15. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (Mar 2021)
16. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3523–3542 (2022)
17. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. In: Medical Imaging with Deep Learning, 4-6 July, Amsterdam (2018)
18. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 6222–6231 (2023)
19. Rahman, M.M., Munir, M., Marculescu, R.: EMCAD: efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11769–11779 (2024)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Ruan, J., Li, J., Xiang, S.: VM-UNet: Vision mamba unet for medical image segmentation. arXiv preprint arXiv:2402.02491 (2024)
22. Ruan, J., Xiang, S., Xie, M., Liu, T., Fu, Y.: MALUNet: a multi-attention and light-weight unet for skin lesion segmentation. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1150–1156. IEEE (2022)
23. Ruan, J., Xie, M., Gao, J., Liu, T., Fu, Y.: Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 481–490. Springer (2023). https://doi.org/10.1007/978-3-031-43901-8_46
24. Vaswani, Aet al.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
25. Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R.: Mixed transformer u-net for medical image segmentation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2390–2394. IEEE (2022)
26. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: a survey. IET Image Proc. **16**(5), 1243–1267 (2022)
27. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

28. Wu, R., Liu, Y., Liang, P., Chang, Q.: Ultralight VM-UNet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. arXiv preprint arXiv:2403.20035 (2024)
29. Yang, Z., Farsiu, S.: Directional connectivity-based segmentation of medical images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11525–11535 (2023)
30. Zhang, M., Yu, Y., Jin, S., Gu, L., Ling, T., Tao, X.: VM-UNET-V2: rethinking vision mamba unet for medical image segmentation. In: International Symposium on Bioinformatics Research and Applications. pp. 335–346. Springer (2024). https://doi.org/10.1007/978-981-97-5128-0_27
31. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: efficient visual representation learning with bidirectional state space model. In: Proceedings of the 41st International Conference on Machine Learning, pp. 62429–62442. PMLR, Vienna, Austria (Jul 2024)

Harnessing EHRs for Diffusion-Based Anomaly Detection on Chest X-Rays

Harim Kim[1], Yuhan Wang[2], Minkyu Ahn[1], Heeyoul Choi[1], Yuyin Zhou[2], and Charmgil Hong[1(✉)]

[1] Handong Global University, Pohang, South Korea
{hrkim,minkyuahn,hchoi,charmgil}@handong.ac.kr
[2] University of California, Santa Cruz, CA, USA
{ywan1332,yzhou284}@ucsc.edu

Abstract. Unsupervised anomaly detection (UAD) in medical imaging is crucial for identifying pathological abnormalities without requiring extensive labeled data. However, existing diffusion-based UAD models rely solely on imaging features, limiting their ability to distinguish between normal anatomical variations and pathological anomalies. To address this, we propose Diff3M, a multi-modal diffusion-based framework that integrates chest X-rays and structured Electronic Health Records (EHRs) for enhanced anomaly detection. Specifically, we introduce a novel Image-EHR Cross-Attention module to incorporate structured clinical context into the image generation process, improving the model's ability to differentiate normal from abnormal features. Additionally, we develop a static masking strategy to enhance the reconstruction of normal-like images from anomalies. Extensive evaluations on CheXpert and MIMIC-CXR/IV demonstrate that Diff3M achieves state-of-the-art performance, outperforming existing UAD methods in medical imaging. Our implementation is available at https://github.com/nth221/Diff3M

Keywords: Diffusion Models · Multi-modality · Anomaly Detection

1 Introduction

Recent AI advancements have remarkably improved radiographic image analysis, providing crucial diagnostic support to radiologists [7,9,19]. However, current clinical practices rely on AI tools designed for predefined disease predictions [21]. This reliance raises the risk of overlooking clinically significant but previously uncharacterized pathological features. To overcome this limitation, unsupervised anomaly detection (UAD) offers a promising direction, with the capability to identify undefined pathological features without prior annotation [2].

Generative model-based anomaly detection has gained interest in recent years [1,25], distinguishing anomalies by regenerating input data into anatomically plausible representations [25,26]. Among these approaches, diffusion-based

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 235–245, 2026.
https://doi.org/10.1007/978-3-032-04947-6_23

models excel in industrial domains, due to their powerful generation capabilities [11,20]. These models extract semantic features characterizing normal data and use them as a reference to reconstruct the input into a normal-like image. However, the variability in individual anatomical structures within radiographic images complicates the extraction of features that manifest normal data, making it difficult to enhance detection performance [3].

To address this challenge, we propose a Diffusion-based Multi-modal Medical Anomaly Detection (Diff3M) framework that leverages both chest X-rays and structured Electronic Health Records (EHRs) for enhanced anomaly detection. While previous studies show that integrating EHR data with medical images enhances clinical prediction [30,32], its potential in anomaly detection remains largely unexplored. To bridge this gap, we introduce a novel Image-EHR Cross-Attention (IECA) mechanism that allows EHR data to provide additional semantic context, enabling the model to differentiate between normal anatomical variations and pathological anomalies. This mitigates the impact of anatomical variability in the generation process, resulting in more reliable anomaly detection. Additionally, we propose a novel masking strategy for reconstructing anatomically normal-appearing images. Existing diffusion-based anomaly detection methods primarily employ random masking with predefined textures [13,31], which lacks generalizability. In contrast, we introduce a Pixel-level Checkerboard Masking (PCM) module, designed to capture a broader range of anomalies by incorporating static masking patterns that generalize beyond specific textures. Throughout the reverse diffusion process, Diff3M reconstructs masked regions into normal-like images by leveraging both the noised input and EHR-guided semantic embeddings, enhancing its ability to detect and characterize anomalies effectively.

We conduct comprehensive experiments using two major chest radiography datasets, CheXpert [14] and MIMIC-CXR/IV [15,16], both offering rich collections of chest X-rays paired with clinical information. Our evaluation, combining quantitative metrics and qualitative analysis, shows that the proposed method outperforms existing UAD approaches for medical images. The contributions of this paper can be summarized as follows:

1. The first diffusion-based medical anomaly detection framework that effectively integrates EHR data as conditioning.
2. A novel Image-EHR Cross-Attention module to generate EHR embeddings that incorporate structured clinical information into the detection pipeline.
3. A new static masking strategy to enhance the reconstruction of normal-like images from anomalies.
4. State-of-the-art performance in medical UAD, surpassing existing methods.

2 Background

Diffusion-Based UAD. Diffusion-based models reconstruct anomalies into normal-like images. AnoDDPM [29] improves normal-like image generation using

Simplex noise [22] but underperforms compared to standard denoising autoencoders [6]. mDDPM [13] applies random masking but targets specific anomalies rather than undefined abnormalities. DDAD [20] and the more recent DiAD [11] condition the reverse process on extracted image features, which may mix normal and anomalous traits, leading to generalization issues.

Feature-Based UAD. Feature-based UAD methods detect anomalies by analyzing features extracted from pre-trained foundation models [2]. PaDiM, CFA and PatchCore [4,18,24] use memory bank to compare features from normal training and testing data, while RD4AD [5] and MambaAD [10] employ knowledge distillation to detect anomalies by measuring discrepancies between teacher and student networks. However, the general-purpose foundation models may struggle to extract meaningful medical features.

Diffusion Models. The Denoising Diffusion Probabilistic Model (DDPM) [12] consists of two main processes: a forward process $q(\cdot)$ and a reverse process $p_\theta(\cdot)$. The forward process gradually adds Gaussian noise to input data $\mathbf{x}$ over T steps, mapping it to a standard normal distribution. Each step t $(t = 1, ..., T)$ introduces noise $\epsilon^{(t)} \sim N(0, 1)$, scaled by a variance parameter $\beta_t \in (0, 1)$ like $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$. The reverse process is a learnable procedure that reconstructs data from $\mathbf{x}_T$ to generate sample $\mathbf{x}_0$ resembling the training data. It is trained by minimizing the KL divergence of $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ and $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. It can be interpreted as approximating each step of the reverse process as the inverse operation of the corresponding forward process step. Consequently, the model learns to predict the noise $\epsilon^{(t)}$ using $\mathcal{L} := \|\epsilon_\theta^{(t)}(\mathbf{x}_t) - \epsilon^{(t)}\|_2^2$.

The Denoising Diffusion Implicit Model (DDIM) [27] enhances DDPM efficiency by reformulating the forward process with a reversed ordinary differential equation (ODE), making it deterministic and non-Markovian [28]:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \sqrt{\bar{\alpha}_{t+1}} \left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}} \right) \mathbf{x}_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} + 1} \right) \epsilon_\theta^{(t)} \right]. \tag{1}$$

where $\bar{\alpha}_t := \prod_{s=1}^{t}(1 - \beta_s)$. Since the processes are no longer Markovian, DDIM enables sample generation with fewer steps than T. The method proposed in this paper is built upon DDIM.

3 Proposed Framework

We present **Diff3M**, a diffusion-based UAD framework trained exclusively on normal data (Fig. 1). It comprises two key steps: (1) a *Noise Prediction* (NP) Network for estimating noise $\epsilon_\theta^{(t)}$ in the reverse diffusion process and (2) a *Masked Pixel Generation* (MPG) Network for reconstructing masked inputs into anatomically plausible images $\tilde{\mathbf{x}}_t$. To effectively integrate EHRs for enhanced anomaly detection, we propose two key designs. First, we introduce a novel *Image-EHR Cross Attention* (IECA) module, which extracts EHR embeddings from paired X-ray images $\mathbf{x}$ and EHR data $\mathbf{r}$, conditioning the MPG Network and the NP

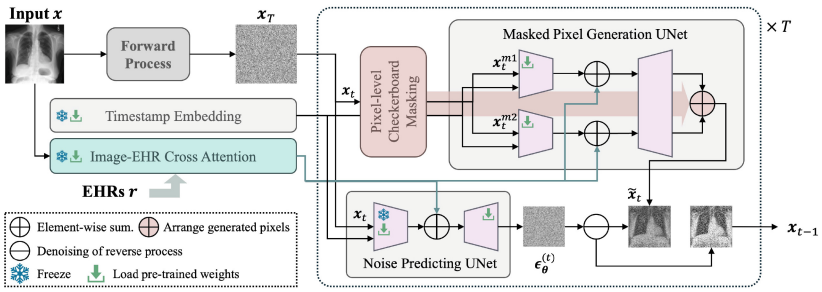


Fig. 1. Training process of Diff3M. Image-EHR Cross Attention (IECA) incorporates EHR embeddings as conditional inputs for the noise prediction (NP) network at each reverse process step. A novel Pixel-level Checkerboard Masking (PCM) strategy is employed to improve generating masked inputs into normal-like images from the Masked Pixel Generation (MPG) Network.

Network (Sect. 3.1). This enables a more semantically informed diffusion process. Additionally, following the original DDPM framework [12], we incorporate timestamp embeddings to capture temporal information. Second, we propose a *Pixel-level Checkerboard Masking* (PCM) strategy to enhance image reconstruction by generating masked inputs from $\mathbf{x}_t$, which the MPG Network then reconstructs into normal-appearing images (Sect. 3.2).

3.1 Image-EHR Cross Attention (IECA)

Existing diffusion-based models [11,20] extract normal semantic features solely from input images, potentially incorporating anatomical anomalies. To mitigate this, we propose an Image-EHR Cross Attention (IECA) module to derive an embedding that represents the X-ray image through EHR data, reducing the influence of anomalous features. As shown in Fig. 2-(a), we first use a feature tokenizer [8] to generate EHR embeddings, producing f tokens based on tabular features. The generated tokens $F \in \mathbb{R}^{f \times d}$ are then used to compute token-wise weights $\mathbf{w}_r$ based on their similarity to the input image embedding $\mathbf{e}$. The final conditional embedding $\mathbf{c}_r$ is then obtained as a weighted sum of F:

$$\begin{aligned} &F = \text{FeatureTokenizer}(\mathbf{r}), \quad \mathbf{r} \in \mathbb{R}^{1\times(\text{col})}, \quad F \in \mathbb{R}^{f\times d}, \\ &\mathbf{c}_r = \mathbf{w}_r F, \quad \mathbf{w}_r = \sigma_{softmax}((F\mathbf{e}^T)/\sqrt{d}), \quad \mathbf{e} = \text{Encoder}(\mathbf{x}), \quad \mathbf{e} \in \mathbb{R}^{1\times d} \end{aligned} \tag{2}$$

Since the generated conditional embedding $\mathbf{c}_r$ is based on the similarity with $\mathbf{e}$, it implicitly encodes the EHRs information to align with the semantic features of the input image. $\mathbf{c}_r$ is then used as a conditional input for MPG and NP to provide normal semantic features for the input image.

3.2 Pixel-Level Checkerboard Masking (PCM)

Existing masking-based methods apply random masking with specific textures to simulate predefined target anomalies, training models to reconstruct the

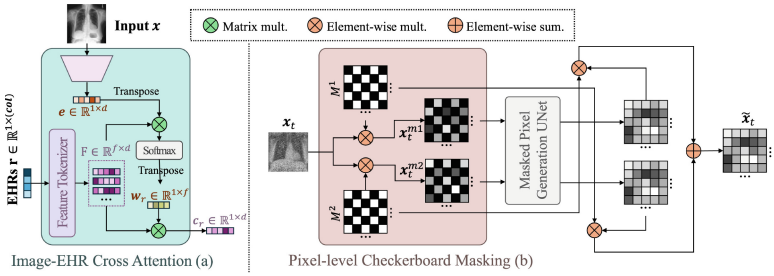


Fig. 2. Details of the proposed modules. (a) The Image-EHR Cross Attention module incorporates structured clinical context into the image generation process. (b) The Pixel-level Checkerboard Masking module facilitates the regeneration of anomalies into normal-like images using two types of static masks.

simulated patterns into anatomically acceptable shapes [13,31]. However, such approaches often struggle to detect real pathological abnormalities that deviate from predefined patterns. To avoid reliance on specific anomaly patterns and encourage the model to depend on auxiliary information—specifically, the EHR conditional embedding from IECA, we employ static masking that deliberately obscures image regions. This approach enables the detection of a broader range of anomalies beyond those seen in the training data.

As illustrated in Fig. 2-(b), PCM consists of two complementary checkerboard masks (M^1, M^2) of input image size. A scale parameter s adjusts masking intensity by time step t to prevent later losses affecting earlier steps. At time step t, applying M^1 and M^2 to noised input $\mathbf{x}_t$ yields $\mathbf{x}_t^{m1}$ and $\mathbf{x}_t^{m2}$. To obtain the final prediction, we apply the opposite mask to each reconstructed image. The formulation is as follows:

$$\begin{aligned} M^1_{i,j} = (i+j) \bmod 2, \quad M^2_{i,j} = 1 - (i+j) \bmod 2, \\ M^{1,2} \leftarrow (M^{1,2} \times (1-s) + s), \quad s = t/T, \\ \tilde{\mathbf{x}}_t = \mathrm{MPG}(\mathbf{x}_t^{m1}) \odot M^2 + \mathrm{MPG}(\mathbf{x}_t^{m2}) \odot M^1, \quad \mathbf{x}_t^{m1}, \mathbf{x}_t^{m2} = \mathbf{x}_t \odot M^{1,2} \end{aligned} \tag{3}$$

The regenerated $\tilde{\mathbf{x}}_t$ is used for loss function during training and for generating the previous step image $\mathbf{x}_{t-1}$ during detection.

Training and Detection. To train the noise prediction capability of the NP Network and the masked image reconstruction capability of the MPG Network, we compute the mean squared error (MSE) between output of each network $(\epsilon_\theta^{(t)}, \tilde{\mathbf{x}}_t)$ and its corresponding target $(\epsilon^{(t)}, \mathbf{x}_t)$ as follows:

$$\mathcal{L}_{Diff3M} = \mathbb{E}_{t,\mathbf{x},\mathbf{c}_r,\epsilon}\left[\lambda \|\epsilon_\theta^{(t)} - \epsilon^{(t)}\|_2^2 + (1-\lambda)\|\tilde{\mathbf{x}}_\mathbf{t} - \mathbf{x}_t\|_2^2\right] \tag{4}$$

where $\lambda \in (0,1)$ is a weighting factor. For more efficient training, the base DDPM and IECA module can be pre-trained on X-ray images and EHR data (Fig. 1).

Algorithm 1 outlines the Diff3M detection process. Using the DDIM sampling strategy, the forward process introduces noise into the image up to T' ($T' < T$) steps. During the reverse process, the model generates $\mathbf{x}_{t-1}$ using produced noise $\epsilon_\theta^{(t)}$ and image $\tilde{\mathbf{x}}_t$ following [28]. The final anomaly score measures the difference between the reconstructed image $\mathbf{x}_0$ and the original input $\mathbf{x}$. Samples with high anomaly scores are flagged as potential anomalies.

Algorithm 1. Anomaly detection using Diff3M

Input: input image $\mathbf{x}$, EHRs $\mathbf{r}$, timestamp t, noise level T'
Output: image-level anomaly score a
for all t from 0 to $T' - 1$ **do**
    $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \sqrt{\bar{\alpha}_{t+1}} \left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}} \right) \mathbf{x}_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \right) \epsilon_\theta^{(t)} \right]$
end for
$\mathbf{c}_r \leftarrow IECA(\mathbf{r}, \mathbf{x})$
for all t from T' to 1 **do**
    $t_{emb} \leftarrow TimestampEmbedding(t)$
    $\tilde{\mathbf{x}}_t \leftarrow MPG(\mathbf{x}_t^{m1}, \mathbf{x}_t^{m2}, \mathbf{c}_r, t_{emb}), \quad (\mathbf{x}_t^{m1}, \mathbf{x}_t^{m2}) \leftarrow PCM(\mathbf{x}_t)$
    $\mathbf{x}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{\tilde{\mathbf{x}}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta^{(t)}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta^{(t)}, \quad \epsilon_\theta^{(t)} \leftarrow NP(\mathbf{x}_t, \mathbf{c}_r, t_{emb})$
end for
$a \leftarrow diff(\mathbf{x}, \mathbf{x}_0), \quad diff(\mathbf{x}, \mathbf{x}_0) \in \{\|\mathbf{x} - \mathbf{x}_0\|_2^2, \max_{i,j} |\mathbf{x}^{i,j} - \mathbf{x}_0^{i,j}|\}$
return a

4 Experiments and Results

Datasets and Evaluation Metrics. CheXpert [14] and MIMIC-CXR [15] both contain X-ray images, the same types of demographic information, and 12 disease classes. We define normal samples as cases with *No Finding*, while all other cases are considered anomalies. For demographic features, we use sex, age, and AP/PA view. In MIMIC-CXR, we further incorporate BMI, blood pressure, height, and weight from MIMIC-IV [17] as additional EHR features, selecting records within three months of the X-ray imaging date. CheXpert has 16,969 normal samples for training, and (26/176) (normal/anomaly) samples for testing. MIMIC has 29,310 normal samples for training, and (487/1646) (normal/anomaly) samples for testing. We evaluate image-level anomaly detection using AUROC and AUPRC to compare performance across various detection thresholds.

Implementation Details. Diff3M is trained using 8 Nvidia RTX A5000 GPUs. All images are resized to 256×256. IECA, MPG, and NP adopt the UNet-based architecture from [28]. We use Adam optimizer with a maximum time step T of 1,000, a batch size of 3, a weighting factor λ of 0.5 and a learning rate of 10^{-4}. For Chexpert, the model is trained from scratch for 60,000 iterations. For MIMIC, the model is pre-trained on CheXpert, and fine-tuned with EHRs for 70,000 iterations. During inference, DDIM sampling is applied with 600/400 steps for CheXpert/MIMIC.

Table 1. Performance comparison on CheXpert and MIMIC datasets. *demo*: demographics information for IECA. *ehr*: additional EHRs from MIMIC-IV for IECA.

Dataset	Metric	Feature-based			Diffusion-based		
		PatchCore	RD4AD	MambaAD	DiAD	Ours$_{demo}$	Ours$_{demo,ehr}$
CheXpert	AUROC	0.584	0.529	0.594	**0.664**	**0.664**	–
	AUPRC	0.888	0.719	0.889	0.917	**0.931**	–
MIMIC	AUROC	0.573	0.555	0.570	0.597	0.610	**0.617**
	AUPRC	0.796	0.800	0.779	0.803	0.818	**0.821**

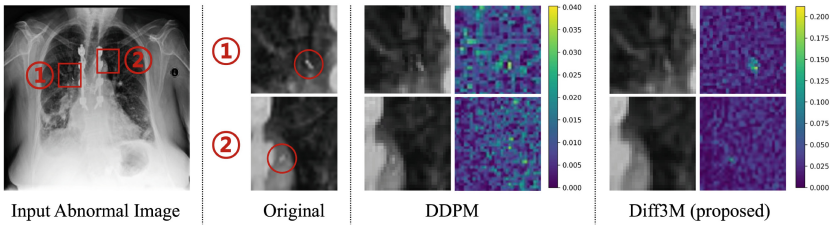


Fig. 3. Qualitative comparison. Red circles indicate anomalous artifacts. (Color figure online)

4.1 Performance Comparison

We compare Diff3M with state-of-the-art (SOTA) feature-based models, *e.g.*, PatchCore [24], RD4AD [5], and MambaAD [10], along with the diffusion-based model DiAD [11]. For qualitative evaluation, we compare the anomaly maps, computed as the absolute pixel-level difference between the input and output.

Quantitative and Qualitative Results. As shown in Table 1, Diff3M achieves SOTA performance in image-level anomaly detection for both CheXpert and MIMIC. For MIMIC, Diff3M improves AUROC by approximately 2.0–5.5% over previous approaches. This highlights the importance of effectively leveraging EHR data to enhance anomaly detection. For CheXpert, the performance of Diff3M using only demographics-based EHRs is comparable to DiAD. This suggests that demographic information alone may not fully capture anatomical variations. Overall, diffusion-based approaches outperform feature-based methods, likely because foundation models struggle to extract effective features in the medical domain, as discussed in Sect. 2.

Figure 3 compares the anomaly regeneration capability of DDPM and Diff3M. We examine an abnormal image containing statistically anomalous artifacts. Unlike DDPM, Diff3M successfully transforms anomalous artifacts into normal-like structures for each region, assigning significantly higher anomaly scores compared to surrounding areas. These results demonstrate that Diff3M effectively generates anomaly maps, which can serve as a valuable medical assistive tool.

Table 2. Ablation study on the modules of Diff3M (MSE/max_{abs}).

Dataset	Metric	DDPM	DDPM+PCM	Ours$_{demo}$	Ours$_{demo,ehr}$
CheXpert	AUROC	0.500/0.500	0.553/0.640	0.547/**0.664**	-
	AUPRC	0.838/0.867	0.902/0.922	0.899/**0.931**	-
MIMIC	AUROC	0.562/0.551	0.592/0.562	0.610/0.567	**0.617**/0.568
	AUPRC	0.800/0.791	0.807/0.793	0.818/0.799	**0.821**/0.798

Table 3. Analysis of attention weights for each input feature (mean±std). *entire*: analysis across entire samples. $top_{10\%}$: analysis for the top 10% of feature values.

input feature f		BMI	BP$_{max}$	BP$_{min}$	Height	Weight
mean $\mathbf{w}_r^{(f)}$	*entire*	**0.262**±0.117	0.183±0.043	0.116±0.037	0.210±0.040	0.230±0.038
	$top_{10\%}$	**0.475**±0.033	0.124±0.022	0.081±0.024	0.145±0.021	0.174±0.013

4.2 Ablation Study

Table 2 shows that adding PCM and IECA sequentially improves performance on both datasets. This demonstrates that each module contributes to anomaly detection and that their combination further enhances performance. We also compare the performance of MSE and pixel-level maximum absolute difference (max_{abs}) as anomaly scoring methods. The results show that MSE is more effective for MIMIC, whereas max_{abs} performs better for CheXpert. To understand this discrepancy, we analyze anomaly detection performance for individual disease classes. In MIMIC, switching to max_{abs} leads to a 12.3–19.9% performance drop for *Fracture* and *Pleural Other* cases. Conversely, in CheXpert, the number of positive samples for these classes is extremely low (0–1), which explains why using max_{abs} does not negatively impact performance. These findings suggest that different scoring functions may be more suitable for different diseases. Therefore, when using the anomaly detection models as a medical assistive tool, it may be necessary to review anomalies based on different scoring functions.

Effects of EHRs. Table 3 presents the mean of attention weights ($\mathbf{w}_r$) assigned to each IECA input feature f in the MIMIC test dataset. Among the EHR features, BMI has the highest average weight. In particular, for the top 10% of samples, BMI accounts for 47.5%, indicating a significantly dominant contribution. This suggests that BMI is the most representative feature for capturing the anatomical characteristics of input images. Unlike height and weight, which independently provide limited information about body shape, BMI (calculated as weight/height^2) offers a more meaningful approximation of body shape. These findings demonstrate that IECA effectively combines EHR features, extracting conditional embeddings that reflect anatomical information. In the case of demographic features, most of the weights converged to zero. These results suggest that utilizing richer EHR features is crucial for robust anomaly detection.

5 Conclusion

This paper proposes Diff3M, a novel framework that leverages EHR data for enhancing medical anomaly detection. By integrating EHR data through Image-EHR Cross Attention (IECA) and enforcing structured feature learning with Pixel-level Checkerboard Masking (PCM), Diff3M consistently outperforms existing diffusion-based and feature-based methods on CheXpert and MIMIC datasets. In the future work, we will apply lightweight models such as latent diffusion [23] to analyze the trade-off between detection performance and efficiency resulting from structural shifts in the model. In addition, we will explore incorporating richer clinical data to further enhance its diagnostic support capabilities.

Acknowledgments. This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program (RS-2024-00431394) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: GANomaly: semi-supervised anomaly detection via adversarial training. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11363, pp. 622–637. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20893-6_39
2. Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., Li, X.: Bmad: benchmarks for medical anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4042–4053 (2024)
3. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Diffusion models with implicit guidance for medical anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 211–220. Springer (2024). https://doi.org/10.1007/978-3-031-72120-5_20
4. Defard, T., Setkov, A., Loesch, A., Audigier, R.: PaDiM: a patch distribution modeling framework for anomaly detection and localization. In: Del Bimbo, A., et al. (eds.) ICPR 2021. LNCS, vol. 12664, pp. 475–489. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68799-1_35
5. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9737–9746 (2022)
6. Fan, Y., Liao, H., Huang, S., Luo, Y., Fu, H., Qi, H.: A survey of emerging applications of diffusion probabilistic models in mri. Meta-Radiol., 100082 (2024)
7. Feng, X., Huang, Q., Li, X.: Ultrasound image de-speckling by a hybrid deep network with transferred filtering and structural prior. Neurocomputing **414**, 346–355 (2020)
8. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Adv. Neural. Inf. Process. Syst. **34**, 18932–18943 (2021)

9. Han, Z., et al.: Contrastive diffusion model with auxiliary guidance for coarse-to-fine pet reconstruction. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 239–249. Springer (2023). https://doi.org/10.1007/978-3-031-43999-5_23
10. He, H., et al.: Mambaad: exploring state space models for multi-class unsupervised anomaly detection. arXiv preprint arXiv:2404.06564 (2024)
11. He, H., et al.: A diffusion-based framework for multi-class anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 8472–8480 (2024)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
13. Iqbal, H., Khalid, U., Chen, C., Hua, J.: Unsupervised anomaly detection in medical images using masked diffusion model. In: International Workshop on Machine Learning in Medical Imaging, pp. 372–381. Springer (2023). https://doi.org/10.1007/978-3-031-45673-2_37
14. Irvin, J et al.: Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)
15. Johnson, A., et al.: Mimic-iv (version 3.1). physionet (2024)
16. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr database (version 2.0. 0). Physionet **2**, 5 (2019)
17. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: Mimic-iv. PhysioNet., pp. 49–55 (2020). https://physionet.org/content/mimiciv/1.0/ Accessed 23 August 2021
18. Lee, S., Lee, S., Song, B.C.: Cfa: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. IEEE Access **10**, 78446–78454 (2022)
19. Mao, J., et al.: Medsegfactory: Text-guided generation of medical image-mask pairs. arXiv preprint arXiv:2504.06897 (2025)
20. Mousakhan, A., Brox, T., Tayyub, J.: Anomaly detection with conditioned denoising diffusion models. arXiv preprint arXiv:2305.15956 (2023)
21. Obuchowicz, R., Lasek, J., Wodziński, M., Piórkowski, A., Strzelecki, M., Nurzynska, K.: Artificial intelligence-empowered radiology–current status and critical review. Diagnostics **15**(3), 282 (2025)
22. Perlin, K.: Improving noise. In: Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, pp. 681–682 (2002)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
24. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328 (2022)
25. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-anogan: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. **54**, 30–44 (2019)
26. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 146–157. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_12
27. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

28. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 35–45. Springer (2022). https://doi.org/10.1007/978-3-031-16452-1_4
29. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: Anoddpm: anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 650–656 (2022)
30. Yao, W., Yin, K., Cheung, W.K., Liu, J., Qin, J.: Drfuse: learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In: Proceedings of the AAAI conference on artificial intelligence, vol. 38, pp. 16416–16424 (2024)
31. Zhang, X., Li, N., Li, J., Dai, T., Jiang, Y., Xia, S.T.: Unsupervised surface anomaly detection with diffusion probabilistic model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6782–6791 (2023)
32. Zhou, Y., et al.: Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. arXiv preprint arXiv:2111.11665 (2021)



Hierarchical Anatomy-Aware Guidance for Brain Tissue Microstructure Reconstruction from T1-Weighted MRI

Yuxing Li and Chuyang Ye(✉)

School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China
chuyang.ye@bit.edu.cn

Abstract. Tissue microstructure information reconstructed from diffusion *magnetic resonance imaging* (MRI) provides crucial brain tissue information for brain disease analysis. However, clinical imaging time constraints often limit the availability of diffusion MRI, thus prompting research into tissue microstructure reconstruction from clinically feasible MRI modalities, such as T1-weighted MRI. Recent Transformer-based generative adversarial networks demonstrate potential by capturing long-range dependencies via self-attention in general MRI synthesis tasks, yet the significant gap between diffusion and T1-weighted MRI limits their ability to achieve optimal performance, leading to anatomical inconsistency in the reconstructed tissue microstructure maps. To address the problem, we propose a *hierarchical anatomy-aware guidance* (HAAG) framework for brain tissue microstructure reconstruction from T1-weighted MRI. First, we consider a two-level strategy to introduce the anatomical priors for the Transformer. At the input level of the Transformer, we propose an adaptive semantic embedding module that seamlessly integrates anatomical structure category information, providing semantic-level guidance for tissue microstructure reconstruction. At the feature modeling level of the Transformer, we propose a distance-guided self-attention mechanism to achieve effective information fusion of anatomical structures that balances both global and local contexts. Then, we consider a more general approach to verify the anatomical consistency at the output level of the whole synthesis network. We develop an anatomy-aware discriminative loss that encourages anatomical consistency between the input and output modalities. HAAG was validated on a public brain MRI dataset for reconstruction of tissue microstructure from T1-weighted MRI. The results demonstrate that our method significantly improves the quality of tissue microstructure reconstruction.

Keywords: Tissue Microstructure Reconstruction · Anatomical Prior · Consistency Generation

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 246–256, 2026.
https://doi.org/10.1007/978-3-032-04947-6_24

1 Introduction

Tissue microstructure information reconstructed from diffusion *magnetic resonance imaging* (MRI) has demonstrated great value into human brain researches [10]. However, the demand for long scanning times makes the acquisition of diffusion MRI impractical in clinical settings, thereby impeding the use of tissue microstructure information for clinical use. To address the problem, it is possible to apply generic MRI synthesis approaches so that tissue microstructure reconstruction is achieved via synthesis from clinically feasible MRI modalities, such as T1-weighted MRI. For example, earlier MRI synthesis works use registration-based strategies [15], compressed sensing [18], and random forest regression [9]. More Recent methods use deep learning techniques, mostly with generators and discriminators trained by adversarial learning [6,7,27]. These methods are based on *convolutional neural networks* (CNNs) [1,5,12–14,23] or Transformer-based generators [4,22]. Additionally, diffusion models [8,16,17] have been developed for better MRI synthesis performance, but they usually require considerable computational resources.

Among the DL models, the Transformer-based *generative adversarial network* (GAN) is more practical in resource-limited clinical scenarios with moderate computational efficiency. Thus, we focus on the use of Transformer-based MRI synthesis networks for brain tissue microstructure reconstruction from clinically feasible MRI. However, direct application of existing Transformer-based models for tissue microstructure reconstruction only leads to suboptimal performance due to the insensitivity of clinically feasible MRI to tissue microstructure, which greatly increases the synthesis difficulty and leads to anatomical inconsistency in the synthesized result. Thus, improvement of Transformer-based methods is still required for brain tissue microstructure reconstruction.

In this study, we introduce a *Hierarchical Anatomy-Aware Guidance* (HAAG) framework into Transformer-based brain MRI synthesis models for tissue microstructure reconstruction from clinically feasible MRI. HAAG uses anatomical prior knowledge, which has been neglected by previous synthesis methods, to guide the synthesis model at different levels. First, to improve the Transformer framework, a two-level approach is suggested for integrating the anatomical prior knowledge. At the input level, an adaptive semantic embedding module is proposed to integrate anatomical structure category information for semantic guidance. At the feature modeling level, a distance-guided self-attention mechanism is proposed to balance global and local contexts and blend information across anatomical structures. Second, to obtain a more general way to maintain anatomical consistency during network training, at the output level of the whole generator, an anatomy-aware discriminative loss is used to enforce anatomical agreement between input and output images, which encourages structural precision of the reconstruction results. For validation we applied the proposed methods to synthesize the NODDI [25] tissue microstructure maps, including the *Intra-Cellular Volume Fraction* (ICVF) and *Orientation Dispersion* (OD) maps, from T1-weighted MRI with the public *Human Connectome Project* (HCP) [20]

dataset. The results show that HAAG improves the quality of tissue microstructure reconstruction compared with other approaches.

2 Methods

2.1 Problem Formulation

Given a clinically feasible brain MRI modality, we aim to synthesize brain tissue microstructure maps from it. We assume that a training dataset of source MRI scans $\mathbf{X}^{\mathrm{s}}$ (i.e., clinically feasible MRI) and the corresponding target tissue microstructure maps $\mathbf{X}^{\mathrm{t}}$ from the same subjects are available. A synthesis model is trained on this paired data to learn the mapping from source to target images.

2.2 HAAG: Hierarchical Anatomy-Aware Guidance

We assume that the incorporation of prior knowledge about brain anatomy can aid tissue microstructure synthesis from the source MRI modality. To this end, we propose the HAAG framework. It combines two methods to integrate anatomical prior knowledge. First, we introduce the design of learnable semantic embedding with anatomical category information for the input of self-attention. Then, we explain how relative distance prior knowledge benefits self-attention. Second, we introduce a broader strategy for consistency-aligned feature generation and then present the overall brain MRI synthesis model with HAAG.

Adaptive Semantic Embedding, ASE. In the original patch embedding, the input image feature $\mathbf{x} \in \mathbb{R}^{H \times W \times D_{\mathrm{token}}}$ was reshaped into a flattened 2D sequence of patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where H, W, C, N and D_{token} are respectively the height, the width, the feature channel of input, the number of patches and the embedding dimension of the Transformer tokens, P^2 is the patch size, and $N = HW/P^2$.

In the proposed *Adaptive Semantic Embedding* (ASE) module, the anatomical sequence for each patch which is denoted by $\mathbf{C}_p \in \{1, 2, \ldots, K\}^{P^2 \cdot C}$ with K as the number of anatomical categories, is firstly mapped to an embedding space using a learnable embedding matrix $\mathbf{E} \in \mathbb{R}^{K \times D_{\mathrm{emb}}}$ with embedding dimension D_{emb}:

$$\mathbf{Em}_p = \mathbf{E}[\mathbf{C}_p] \in \mathbb{R}^{N \times (P^2 \cdot D_{\mathrm{emb}})}. \tag{1}$$

Then, the embedding matrix $\mathbf{Em}_p$ is projected to the same feature space of the image feature tokens using a learnable linear transformation matrix $\mathbf{W}_{\mathrm{s}} \in \mathbb{R}^{(P^2 \cdot D_{\mathrm{emb}}) \times D_{\mathrm{token}}}$ to gather patch's semantic information from spatial perspective:

$$\mathbf{ASE}_p = \mathbf{Em}_p \cdot \mathbf{W}_{\mathrm{s}} \in \mathbb{R}^{N \times D_{\mathrm{token}}}. \tag{2}$$

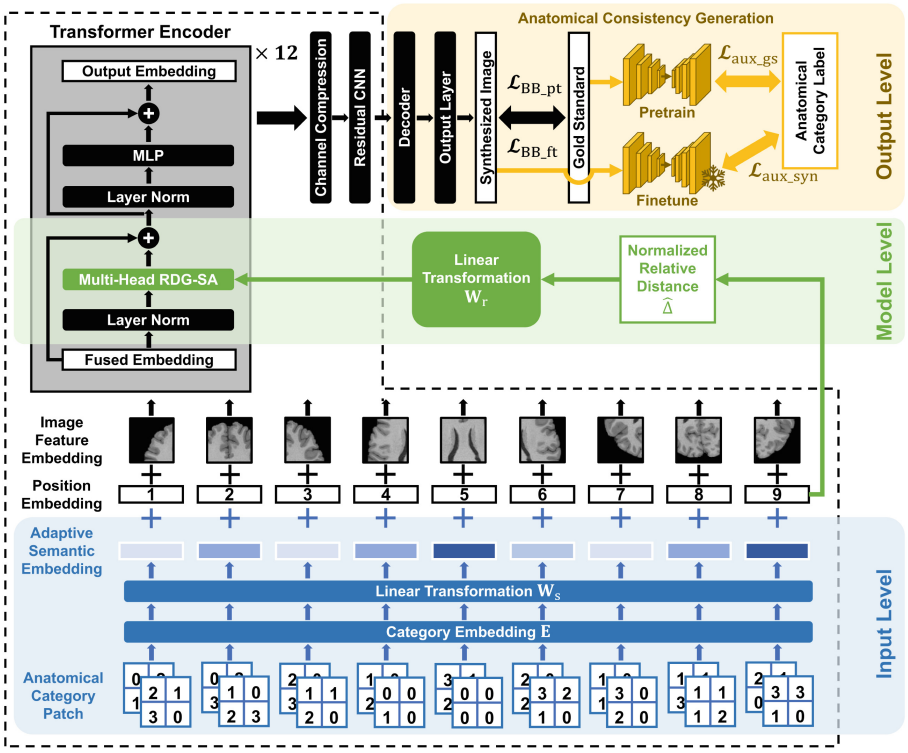


Fig. 1. A schematic of the generator of the brain MRI synthesis model for tissue microstructure reconstruction with Hierarchical Anatomy-Aware Guidance framework.

Finally, the adaptive semantic embedding $\mathbf{ASE}_p$, image feature embedding $\mathbf{X}_p$, and positional embedding $\mathbf{PE}_p$ are fused as follows:

$$\mathbf{X}_p^{'} = \mathbf{X}_p + \mathbf{ASE}_p + \mathbf{PE}_p. \tag{3}$$

Note that to prevent overfitting, dropout is applied to the fused embedding $\mathbf{X}_p^{'}$ with a dropout probability $\gamma = 0.1$.

Relative Distance-Guided Self-attention, RDG-SA. Suppose N image tokens are sent into the self-attention module, and the i-th token is denoted by a row vector $\boldsymbol{X}_i^{'} \in \mathbb{R}^{D_{\text{token}}}$. For each token $\boldsymbol{X}_i^{'}$, the spatial position (x_i, y_i) can be obtained by:

$$x_i = \left\lfloor \frac{i-1}{N_W} \right\rfloor, \quad y_i = (i-1) \bmod N_W, \tag{4}$$

where $i \in \{1, 2, \ldots, N\}$ and $N_W = \frac{W}{P}$.

Given the spatial position, we have the horizontal and vertical relative distances for each pair of tokens $\boldsymbol{X}_i^{'}$ and $\boldsymbol{X}_j^{'}$ as:

$$\Delta x_{ij} = |x_i - x_j|, \quad \Delta y_{ij} = |y_i - y_j|. \tag{5}$$

Thus, the total relative distance Δ_{ij} is defined as the sum of the horizontal and vertical distances:

$$\Delta_{ij} = \Delta x_{ij} + \Delta y_{ij}. \tag{6}$$

Then, the relative distance is obtained by being normalized with the maximum possible distance $\Delta_{\max} = N_H + N_W$, where $N_H = \frac{H}{P}$:

$$\hat{\Delta}_{ij} = \frac{\Delta_{ij}}{\Delta_{\max}}. \tag{7}$$

With the spatial position prior knowledge about brain anatomy, it is possible to construct a relative distance-guided relation matrix $\mathbf{R} \in \mathbb{R}^{N \times N}$ for the input tokens, where the (i, j)-th entry $R_{i,j}$ of $\mathbf{R}$ roots from the complement of normalized relative distance $\hat{\Delta}_{ij}$, forecasting that higher $R_{i,j}$ indicates that i-th token and j-th token are more likely to inform each better:

$$\mathbf{R} = \mathbf{W}_{\mathrm{r}}(\mathbf{1} - \hat{\mathbf{\Delta}}), \tag{8}$$

where $\mathbf{W}_{\mathrm{r}} \in \mathbb{R}^{N \times N}$ is a learnable linear transformation matrix.

We propose to supplement the conventional self-attention with relation matrix $\mathbf{R}$ and design the *Relative Distance-Guided Self-Attention* (RDG-SA) as:

$$\text{RDG-SA}(\mathbf{X}^{'}) = \text{softmax}\left(\frac{\mathbf{X}^{'}\mathbf{W}_{\mathrm{Q}}(\mathbf{X}^{'}\mathbf{W}_{\mathrm{K}})^{\top}}{\sqrt{d_k}} + \mathbf{R}\right)\mathbf{X}^{'}\mathbf{W}_{\mathrm{V}}, \tag{9}$$

where $\mathbf{W}_{\mathrm{Q}}$, $\mathbf{W}_{\mathrm{K}}$, and $\mathbf{W}_{\mathrm{V}}$ are three learnable projection matrices in the Transformer, and d_k is the projected dimension. In this way, this enhanced attention mechanism prioritizes interactions between spatially proximate tokens, effectively modeling the local dependencies.

Multi-head self-attention plays a crucial role in the success of the Transformer [21]. RDG-SA can be extended to multi-head RDG-SA as well, which allows diverse attention for richer representations.

Anatomical Consistency Generation, ACG. In order to explicitly ensure anatomical consistency between the synthesized and target images, we propose an *Anatomical Consistency Generation* (ACG) method where a perceptual loss is introduced.

First, a simple UNet-like perceptual network is constructed with 4 downsampling and 4 upsampling blocks to classify tissue types. Pretrained on gold standard images, it uses the loss function:

$$\mathcal{L}_{\mathrm{pt}} = \mathcal{L}_{\mathrm{BB_pt}} + \mathcal{L}_{\mathrm{aux_gs}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{BB_pt}}$ is the backbone (BB)'s pretraining loss and $\mathcal{L}_{\mathrm{aux_gs}}$ is the cross-entropy loss computed between the gold standard images' classification result and tissue-category labels.

During the fine-tuning phase, the perceptual network is frozen, and the synthesized images are used as input to compute the tissue-type classification error as the loss to back-propagate:

$$\mathcal{L}_{\mathrm{ft}} = \mathcal{L}_{\mathrm{BB_ft}} + \mathcal{L}_{\mathrm{aux_syn}}, \tag{11}$$

where $\mathcal{L}_{\mathrm{BB_ft}}$ is BB's fine-tuning loss and $\mathcal{L}_{\mathrm{aux_syn}}$ is the cross-entropy loss computed between the synthesized images' classification result and the tissue-category labels.

Overall Synthesis Model. Integrating ASE and RDG-SA forms the first component of the HAAG framework, concentrating on enhancing the Transformer by integrating anatomical priors. The second component of HAAG framework is the ACG module, verifying the anatomical consistency of generator's intermediate feature. The architectures of these three mechanisms are respectively shown in Fig. 1.

2.3 Implementation Details

Our selected Transformer-based model ResViT [4] has achieved superior performance in brain MRI synthesis. It has a generator and a discriminator, trained adversarially. Unlike CNN-based models that capture only local contexts, ResViT improves the generator's bottleneck with a Transformer-based module using self-attention for long-range dependencies.

The implementation of our method is based on the open-source code of ResViT. The data preprocessing process includes: (1) computing tissue microstructure and brain parcellation maps respectively from diffusion MRI and structural MRI, and (2) co-registering diffusion MRI and microstructure maps with structural MRI. Like ResViT [4], the proposed method takes 2D MRI slices as input, and the output slices are concatenated to form the 3D synthesis result. The training procedure also follows ResViT, which includes a pretraining phase and a fine-tuning phase. For more details that are not related to the HAAG framework, please refer to ResViT [4]. The codes for the proposed HAAG framework are available at https://github.com/Winterborner/HAAG.

3 Experiments

Data Description and Experimental Settings. To evaluate our proposed methods, we focused on synthesizing ICVF and OD maps from T1-weighted images. This is challenging because ICVF and OD, usually reconstructed from diffusion MRI [25], capture tissue microstructure while T1-weighted images aren't sensitive to it. We chose T1-weighted images as the source modality as they're commonly available in MRI acquisitions.

Experiments were performed on the publicly available HCP dataset [20]. We selected 61 subjects: 36 for training with 1 subject set aside for validation and 25 for testing. Each subject has a T1-weighted scan (0.7 mm isotropic voxel size) and a diffusion MRI scan (270 diffusion gradients on three b-shells: $b = 1000, 2000, 3000$ s/mm^2 and 1.25 mm isotropic voxel size).

ICVF and OD maps estimated from diffusion MRI scans using AMICO [3] were used for training the model with training subjects and as the gold standard for evaluation with test subjects. These maps were rigidly registered with T1-weighted images using the registration tool in ANTs [2]. For anatomical prior knowledge, we used FSL-FAST [26] to perform brain parcellation on T1-weighted images. For clarity in the results, the first and second components of the HAAG framework are denoted as 'Ours1' and 'Ours2' respectively.

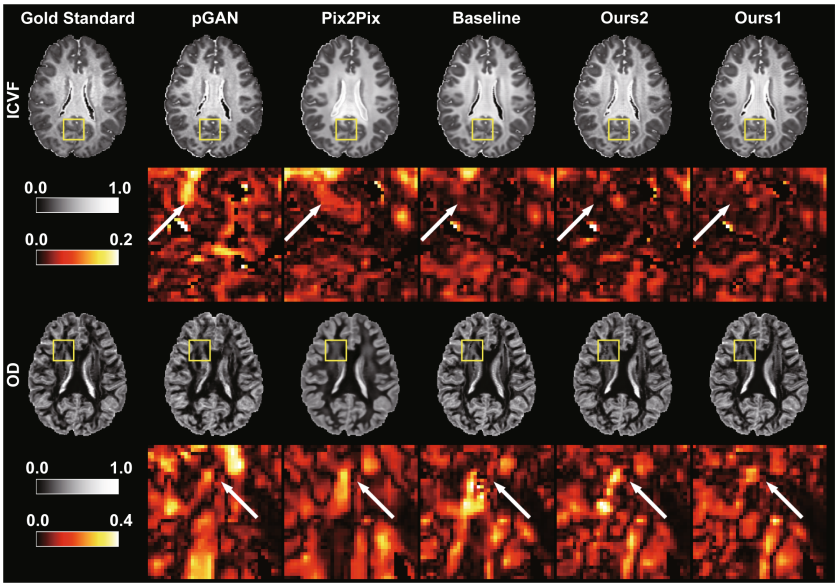


Fig. 2. Representative results of ICVF and OD synthesis and their zoomed error maps with gold standard shown for reference.

Evaluation of Synthesis Results. We compared our approaches with three advanced image synthesis methods: pGAN [5], the 3D version of Pix2Pix (denoted by 'Pix2Pix') [19], and ResViT (denoted by 'Baseline') [4]. An example of the synthesized result of the proposed methods is shown in Fig. 2, where the gold standard is shown as well. Also, the zoomed error maps for the square regions highlighted in the synthesis results are displayed. The zoomed error maps of both proposed methods appear darker than those of competing methods, which indicates better agreement of the proposed methods with the

Table 1. Means and *standard deviations* (stds) of the MAE, RMSE, and PSNR of the synthesis results for the proposed and competing methods. The best results are highlighted in bold. The proposed method with the best performance was compared with all competing methods with paired Student's t-tests ($^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$). The effect sizes (Cohen's d) for the comparison are also shown.

		MAE	RMSE	PSNR
		Mean±Std$_d^p$	Mean±Std$_d^p$	Mean±Std$_d^p$
ICVF	pGAN	$(5.83 \pm 0.29)_{\times 10^{-2}}{}^{***}_{4.06}$	$(8.53 \pm 0.38)_{\times 10^{-2}}{}^{***}_{4.50}$	$21.39 \pm 0.39^{***}_{4.24}$
	Pix2Pix	$(5.03 \pm 0.34)_{\times 10^{-2}}{}^{***}_{1.75}$	$(7.25 \pm 0.40)_{\times 10^{-2}}{}^{***}_{1.67}$	$22.81 \pm 0.47^{***}_{1.64}$
	Baseline	$(5.02 \pm 0.65)_{\times 10^{-2}}{}^{***}_{1.18}$	$(7.12 \pm 0.60)_{\times 10^{-2}}{}^{***}_{1.13}$	$22.98 \pm 0.73^{***}_{1.13}$
	Ours2	$(4.48 \pm 0.31)_{\times 10^{-2}}{}^{**}_{0.32}$	$(6.62 \pm 0.40)_{\times 10^{-2}}{}^{**}_{0.29}$	$23.60 \pm 0.54^{**}_{0.30}$
	Ours1	$\mathbf{(4.36 \pm 0.40) \times 10^{-2}}$	$\mathbf{(6.48 \pm 0.50) \times 10^{-2}}$	$\mathbf{23.79 \pm 0.68}$
OD	pGAN	$(9.86 \pm 0.15)_{\times 10^{-2}}{}^{***}_{16.91}$	$(12.70 \pm 0.18)_{\times 10^{-2}}{}^{***}_{16.28}$	$17.93 \pm 0.13^{***}_{15.04}$
	Pix2Pix	$(7.55 \pm 0.16)_{\times 10^{-2}}{}^{***}_{3.81}$	$(9.66 \pm 0.21)_{\times 10^{-2}}{}^{***}_{2.78}$	$20.31 \pm 0.19^{***}_{2.78}$
	Baseline	$(7.19 \pm 0.17)_{\times 10^{-2}}{}^{***}_{1.79}$	$(9.45 \pm 0.22)_{\times 10^{-2}}{}^{***}_{1.87}$	$20.50 \pm 0.20^{***}_{1.87}$
	Ours2	$(7.07 \pm 0.15)_{\times 10^{-2}}{}^{***}_{1.22}$	$(9.24 \pm 0.20)_{\times 10^{-2}}{}^{***}_{1.08}$	$20.69 \pm 0.19^{***}_{1.09}$
	Ours1	$\mathbf{(6.85 \pm 0.19) \times 10^{-2}}$	$\mathbf{(8.98 \pm 0.26) \times 10^{-2}}$	$\mathbf{20.93 \pm 0.25}$

Table 2. Means and stds of the MAE, RMSE, and PSNR from ablation experiments. The best results are highlighted in bold.

		MAE	RMSE	PSNR
		Mean±Std	Mean±Std	Mean±Std
ICVF	Ours1	$\mathbf{(4.36 \pm 0.40) \times 10^{-2}}$	$\mathbf{(6.48 \pm 0.50) \times 10^{-2}}$	$\mathbf{23.79 \pm 0.68}$
	w/o RDG-SA	$(4.49 \pm 0.34) \times 10^{-2}$	$(6.62 \pm 0.41) \times 10^{-2}$	23.60 ± 0.55
	w/o ASE	$(4.55 \pm 0.37) \times 10^{-2}$	$(6.77 \pm 0.43) \times 10^{-2}$	23.41 ± 0.55
OD	Ours1	$\mathbf{(6.85 \pm 0.19) \times 10^{-2}}$	$\mathbf{(8.98 \pm 0.26) \times 10^{-2}}$	$\mathbf{20.93 \pm 0.25}$
	w/o RDG-SA	$(6.91 \pm 0.14) \times 10^{-2}$	$(9.02 \pm 0.18) \times 10^{-2}$	20.89 ± 0.14
	w/o ASE	$(7.07 \pm 0.17) \times 10^{-2}$	$(9.25 \pm 0.22) \times 10^{-2}$	20.68 ± 0.21

gold standard. Moreover, the synthesized results of both proposed methods contain more accurate details than those of competing methods.

Quantitatively, we measured the *mean absolute error* (MAE), *root mean squared error* (RMSE), and *peak signal-to-noise ratio* (PSNR) of the synthesis results of the proposed and baseline methods, and presented results in Table 1. Ours1 had the lowest MAE and RMSE, along with the highest PSNR, while Ours2 was the second-best. These results show the superior synthesis quality of both proposed methods. Additionally, for all cases, the difference between the optimal method Ours1 and competing methods is statistically significant ($p < 0.001$) with a large effect size ($d > 0.8$).

Ablation Experiments. The ablation study delved deeper into how ASE and RDG-SA affect the synthesis quality of Ours1, with results presented in Table 2. Note that the baseline backbone for this ablation study is that of 'Baseline' in Table 1. ASE and RDG-SA were crucial in improving the precision of the synthesized images, where the complete setup resulted in optimal performance of Ours1.

4 Discussion and Conclusion

We have proposed the HAAG framework, two hierarchical anatomy-aware guidance methods, for brain tissue microstructure reconstruction from clinically feasible MRI. The HAAG framework, with its ASE-RDG-SA combination module and ACG module, incorporates brain anatomy prior knowledge at the Transformer's input and feature-modelling level, and the whole generator-output alignment level. This promotes anatomical consistency between source and target images. Experiments performed on the public HCP dataset for ICVF and OD synthesis from T1-weighted images show that the HAAG framework improves the synthesis quality.

While recent related works have made significant contributions to structural preservation in image synthesis, notably mutual information loss for modality alignment [11] and hierarchical granularity approaches [24], our work represents the first framework leveraging structural MRI and anatomical priors specifically for brain tissue microstructure reconstruction. A limitation of our work is the restricted evaluation on pathological cases, which may affect model generalizability. In future work, we will validate the clinical utility of our method on diverse patient cohorts.

Acknowledgments. This work was supported by the Beijing Municipal Natural Science Foundation (7242273) and the Key Program of National Natural Science Foundation of China (82330057).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Armanious, K., et al.: MedGAN: medical image translation using GANs. Comput. Med. Imaging Graph. **79**, 101684 (2020)
2. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage **54**(3), 2033–2044 (2011)
3. Daducci, A., Canales-Rodríguez, E.J., Zhang, H., Dyrby, T.B., Alexander, D.C., Thiran, J.P.: Accelerated microstructure imaging via convex optimization (AMICO) from diffusion MRI data. Neuroimage **105**, 32–44 (2015)
4. Dalmaz, O., Yurt, M., Çukur, T.: ResViT: residual vision transformers for multimodal medical image synthesis. IEEE Trans. Med. Imaging **41**(10), 2598–2614 (2022)

5. Dar, S.U., Yurt, M., Karacan, L., Erdem, A., Erdem, E., Cukur, T.: Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE Trans. Med. Imaging **38**(10), 2375–2388 (2019)
6. Goodfellow, I., et al.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
8. Jiang, L., Mao, Y., Wang, X., Chen, X., Li, C.: Cola-diff: conditional latent diffusion model for multi-modal mri synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 398–408. Springer (2023). https://doi.org/10.1007/978-3-031-43999-5_38
9. Jog, A., Carass, A., Roy, S., Pham, D.L., Prince, J.L.: Random forest regression for magnetic resonance image synthesis. Med. Image Anal. **35**, 475–488 (2017)
10. Kamiya, K., Hori, M., Aoki, S.: Noddi in clinical research. J. Neurosci. Methods **346**, 108908 (2020)
11. Kang, M., Chikontwe, P., Won, D., Luna, M., Park, S.H.: Structure-preserving image translation for multi-source medical image domain adaptation. Pattern Recogn. **144**, 109840 (2023)
12. Lee, D., Kim, J., Moon, W.J., Ye, J.C.: CollaGAN: collaborative GAN for missing image data imputation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2487–2496 (2019)
13. Li, H., et al.: DiamondGAN: unified multi-modal generative adversarial networks for mri sequences synthesis. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 795–803. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_87
14. Liu, J., Pasumarthi, S., Duffy, B., Gong, E., Datta, K., Zaharchuk, G.: One model to synthesize them all: multi-contrast multi-scale transformer for missing data imputation. IEEE Trans. Med. Imaging **42**(9), 2577–2591 (2023)
15. Miller, M.I., Christensen, G.E., Amit, Y., Grenander, U.: Mathematical textbook of deformable neuroanatomies. Proc. Natl. Acad. Sci. **90**(24), 11944–11948 (1993)
16. Özbey, M., et al.: Unsupervised medical image translation with adversarial diffusion models. IEEE Trans. Med. Imaging (2023)
17. Pan, S., et al.: Cycle-guided denoising diffusion probability model for 3d cross-modality mri synthesis. arXiv preprint arXiv:2305.00042 (2023)
18. Roy, S., Carass, A., Prince, J.: A compressed sensing approach for mr tissue contrast synthesis. In: International Conference on Information Processing in Medical Imaging, pp. 371–383. Springer (2011)
19. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20730–20740 (2022)
20. Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K.: The WU-Minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
21. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
22. Yan, S., Wang, C., Chen, W., Lyu, J.: Swin transformer-based gan for multi-modal medical image translation. Front. Oncol. **12**, 942511 (2022)
23. Yang, H., Sun, J., Yang, L., Xu, Z.: A unified hyper-GAN model for unpaired multi-contrast mr image translation. In: de Bruijne, M., et al. (eds.) MICCAI

2021. LNCS, vol. 12903, pp. 127–137. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_12
24. Yu, Z., et al.: Hifi-syn: hierarchical granularity discrimination for high-fidelity synthesis of mr images with structure preservation. Med. Image Anal. **100**, 103390 (2025)
25. Zhang, H., Schneider, T., Wheeler-Kingshott, C.A., Alexander, D.C.: NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage **61**(4), 1000–1016 (2012)
26. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging **20**(1), 45–57 (2002)
27. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

Hierarchical Part-Based Generative Model for Realistic 3D Blood Vessel

Siqi Chen[1], Guoqing Zhang[1,2], Jiahao Lai[1], Bingzhi Shen[3], Sihong Zhang[1], Caixia Dong[4], Xuejin Chen[5], and Yang Li[1(✉)]

[1] Shenzhen Key Laboratory of Ubiquitous Data Enabling, Tsinghua Shenzhen International Graduate School, Tsinghua University, Beijing, China
csq23@mails.tsinghua.edu.cn, yangli@sz.tsinghua.edu.cn
[2] Pengcheng Laboratory, Shenzhen, China
[3] School of Life Science and the Key Laboratory of Convergence Medical Engineering System and Healthcare Technology, Ministry of Industry and Information Technology, Beijing Institute of Technology, Beijing, China
[4] Institute of Medical Artificial Intelligence, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
[5] MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei, China

Abstract. Advancements in 3D vision have increased the impact of blood vessel modeling on medical applications. However, accurately representing the complex geometry and topology of blood vessels remains a challenge due to their intricate branching patterns, curvatures, and irregular shapes. In this study, we propose a hierarchical part-based framework for 3D vessel generation that separates the global binary tree-like topology from local geometric details. Our approach proceeds in three stages: (1) *key graph generation* to model the overall hierarchical structure, (2) *vessel segment generation* conditioned on geometric properties, and (3) *hierarchical vessel assembly* by integrating the local segments according to the global key graph. We validate our framework on real-world datasets, demonstrating superior performance over existing methods in modeling complex vascular networks. This work marks the first successful application of a part-based generative approach for 3D vessel modeling, setting a new benchmark for vascular data generation. The code is available at: https://github.com/CybercatChen/PartVessel.git.

Keywords: Vasculature · Vessel Generation · 3D Shape Modeling · Hierarchical Structure · Part-based Method

1 Introduction

With the rapid progress in computer graphics, 3D visual technology has significantly enhanced the ability to accurately model medical data [21,22,24]. In particular, 3D vessel generation is crucial for precisely simulating the complex structure of the vasculature, enabling a range of medical applications, from diagnostic assessments [8] to treatment planning [12]. It also supports critical tasks

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 257–267, 2026.
https://doi.org/10.1007/978-3-032-04947-6_25

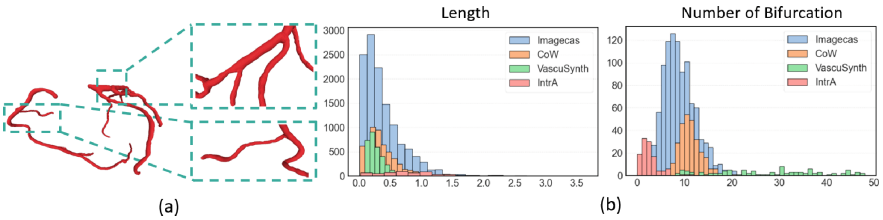


Fig. 1. **(a)** Visualization of a Real-world Coronary Artery Dataset. The vascular network displays a hierarchical, tree-like structure, with localized curvatures and complex branching patterns. **(b)** The histograms of vessel length and number of bifurcations for four different datasets.

such as preoperative simulations [14] and medical image analysis [31], facilitating more accurate and effective decision-making. Moreover, this technology generates detailed datasets that can be applied to downstream tasks like vessel segmentation [5,29] and labeling [27], further improving automated analysis.

Unlike common 3D modeling approaches for objects with regular shapes [3]– such as chairs, tables, or airplanes, which typically feature fixed and predictable structures – vascular modeling poses unique challenges. Those conventional methods [11,20,25], often designed for rigid and uniform objects, are not well suited for modeling blood vessels. As illustrated in Fig. 1, the vasculature is characterized by a high degree of complexity, with numerous bifurcations that vary in both number and location. Additionally, the blood vessels exhibit intricate curvature and irregular, non-uniform characteristics. The morphological diversity of vascular structures in real datasets further compounds these challenges. Such complexity demands a model that can effectively capture fine details and accurately represent the diversity of vascular geometries.

To model vascular structures in detail, several methods have been proposed. Point cloud-based approaches designed for common objects [11], while effective in representing 3D objects, struggle with capturing the geometry of tubular and elongated structures due to their discrete nature. TreeDiffusion [17] utilizes implicit neural fields for modeling anatomical trees, but its flexibility and accuracy are limited in capturing complex vascular geometries. VesselVAE [6] employs skeletal graphs to effectively capture vascular structure. However, it generates the entire vascular network without explicitly addressing the unique characteristics of individual branches. As a result, it performs well for simpler vessels with fewer bifurcations, but its fidelity declines for vessels with many branches.

Although these methods have considered the geometric properties of blood vessels, an effective model must capture both the global and local characteristics. Globally, most vessels follow a tree-like hierarchical structure, with endpoints and bifurcations largely defining their organization. Locally, blood vessels exhibit similar geometric shapes, with variations in radius and length, but they can be viewed as segments of tubular curves. Based on these observations,

we propose a Hierarchical Part-based Vessel Generative Framework, where the global vascular structure is represented by a tree-shaped key graph based on the vessel skeleton, and local segments are modeled as sequential curves. This decomposition naturally captures both the branching topology and local geometric details of the vasculature. Our contributions are as follows: (1) To the best of our knowledge, this is the first work to employ a part-based approach for 3D vessel generation. (2) We explicitly represent the global structure of vessels as a key graph and local segments as sequential curves, significantly enhancing vascular modeling detail. (3) We validate our approach on real-world datasets, demonstrating superior performance over existing methods in modeling complex vascular networks.

2 Related Work

Vessel Generation. Although deep learning has been applied to vascular generation, research in this domain remains limited. Wolterink et al. [23] first employed GANs to generate single-vessel representations sequentially. Feldman et al. [6] later introduced VesselVAE to model branching vascular structures, but it can only handle a small number of bifurcations. Sinha et al. [17] explored implicit neural representations (INRs) and diffusion models for vessel generation, yet fidelity remains a challenge. While these methods offer valuable insights, the complexity of the task still limits practice, highlighting the need for further work to enhance both fidelity and structural accuracy.

3D Part-Based Shape Modeling. Part-based methods in deep learning decompose complex 3D shapes into semantically meaningful components, learned separately and then assembled into a complete structure [4,9]. GRASS [10] and StructureNet [13] first underscored the importance of hierarchical structure, while CompoNet [16] increased shape diversity via transformations of parts and their combinations. Drawing on these approaches, we note a parallel to blood vessels, which exhibit a global tree-like structure and locally tubular geometry.

3 Methodology

Overview. We adopt a skeleton representation and construct a binary-tree *key graph* based on the skeleton's bifurcation and terminal points. In this key graph, each edge corresponds to a distinct vessel segment. Our part-based generative model comprises three stages, as illustrated in Fig. 2. The first stage learns the global binary tree structure. The second stage models each vessel segment. Finally, the third stage assembles the synthesized segments according to the key graph to reconstruct the complete vessel. In what follows, we provide a detailed explanation of the methods employed in each stage.

Stage 1. Key Graph Generation. Recursive Autoencoders (RAE) were initially introduced by [18,19] and later applied to object modeling by [10], as

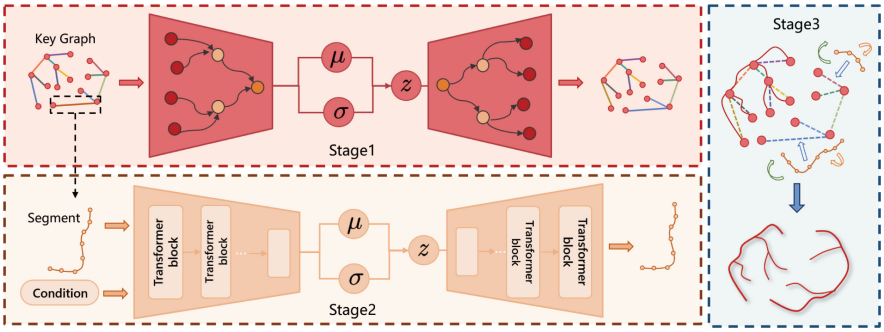


Fig. 2. Overall pipeline of our method. **Stage 1.** Key Graph Generation: learn a global hierarchical tree. **Stage 2.** Vessel Segment Generation: model local 3D curve based on geometric conditions. **Stage 3.** Hierarchical Vessel Assembly: reconstruct the vessel skeleton by assembling segments based on the global layout.

well as vessel modeling by [6]. To extend RAE into a generative framework, we employ a *Recursive Variational Autoencoder (RVAE)* to model and synthesize a key graph representation of the vascular network. Each node's attributes consist of three parts: (1) 3D spatial coordinates $[x, y, z]$, (2) direction $[n_x, n_y, n_z]$ of the local segment, and (3) a geometric descriptor $C = [\ell, \delta, \kappa, \rho]$ that characterizes the local segment's properties, which will be explained in the following stage.

Encoding Phase. Starting from the leaf nodes and moving upward, we aggregate child node features into their parent node. Let v_{parent} be a parent node's attribute, and h_{left} and h_{right} be the hidden states of its left and right children, respectively. The parent's hidden state h_{parent} is computed as follows:

$$h_{parent} = \text{MLP}(\text{concat}[v_{parent}, h_{left}, h_{right}]). \tag{1}$$

By repeatedly performing this operation up the tree, we eventually reach the root node $\mathbf{z}_{root}$, which serves as a global latent embedding for the entire graph.

Decoding Phase. We reverse the procedure to reconstruct node attributes from the root down to the leaves. For each parent node with hidden state h_{parent}, we first use a classifier $\mathbf{y} = \text{NodeCLS}(\hat{h}_{parent})$ to determine whether left and/or right children exist. If a left child is predicted, its attribute is computed by:

$$\hat{v}_{left} = \text{MLP}_{\text{left}}(\hat{h}_{parent}). \tag{2}$$

We then obtain the left child's hidden representation by:

$$\hat{h}_{left} = \text{MLP}(\text{concat}[\hat{h}_{parent}, \hat{v}_{left}]). \tag{3}$$

Similarly, if a right child is predicted, we compute $\hat{v}_{right}$ and $\hat{h}_{right}$ in the same manner. This recursive process continues until all nodes are reconstructed.

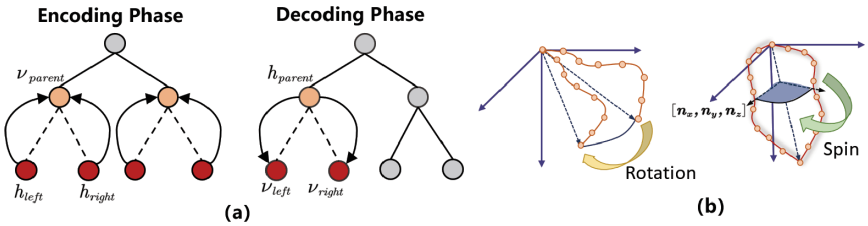


Fig. 3. The encoding and decoding process of the model in *Stage 1*. The two types of rotation processes in *Stage 3*.

Loss Function. The total loss includes three terms: an $\text{MSE}(v, \hat{v})$ for node attribute reconstruction, a $\text{CrossEntropy}(\hat{\mathbf{y}}, \mathbf{y})$ for node-level classification, and a KL divergence to regularize the latent space. These terms together define the final objective:

$$\text{Loss} = \text{MSE}(\hat{v}, v) + \text{CrossEntropy}(\hat{\mathbf{y}}, \mathbf{y}) + D_{\text{KL}}(q(\mathbf{z}_{root}) \| p(\mathbf{z}_{root})). \tag{4}$$

Stage 2. Vessel Segment Generation. For each vessel segment identified in the key graph, we represent the skeleton as an ordered sequence in 3D space. Each point along the sequence is described by $\mathbf{x} = [x, y, z, r]$, where r is the vessel radius. To capture the segment's shape, we introduce the geometric descriptor $C = [\ell, \delta, \kappa, \rho]$ as conditional variables, corresponding to key graph attributes. Specifically, ℓ represents the segment's length, δ is the straight-line distance between the endpoints, κ quantifies the segment's curvature, and ρ indicates the segment's depth within the binary tree, accounting for variations at different branching levels. These features collectively capture both local geometry and broader structural attributes, enabling accurate generative modeling of vessel segments.

Encoding and decoding. We adopt a Transformer-based variational framework in which each point is treated as a token. By conditioning on the geometric descriptor C, the Transformer can synthesize more realistic vessel segments. The encoder maps the input sequence to a latent representation, and the decoder then generates new sequences from this latent space.

Loss Function. The total loss includes three terms: an $\text{MSE}(\mathbf{x}, \hat{\mathbf{x}})$ that measures the reconstruction error, a $\text{CrossEntropy}(\ell, \hat{\ell})$ to ensure the sequence's length, and a KL divergence for generation. Hence, the objective is given by:

$$\text{Loss} = \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + \text{CrossEntropy}(\ell, \hat{\ell}) + D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}, C) \| p(\mathbf{z}|C)). \tag{5}$$

Stage 3. Hierarchical Vessel Assembly. In this final stage, we sample a latent vector from the latent space in *Stage 1*, and decode it into a key graph, and assemble this graph with the individually synthesized vessel segments from *Stage 2* to build the complete skeleton in the same frame. Specifically, we employ

a depth-first search traversal starting from the root node of the key graph. At each step of the traversal, we first attach the corresponding vessel segment by applying scaling and translation, ensuring spatial alignment and orientation.

We then rotate the segment so its local direction aligns with the orientation $[n_x, n_y, n_z]$ derived from the key graph. Figure 3(b) shows the two kinds of rotation processes. These rotations ensure directional consistency with the overall vessel geometry. By progressively attaching each segment via the above transformations, we obtain a complete vessel skeleton. Finally, following the mesh reconstruction method in [30] , we leverage the predicted radius of each segment to reconstruct the final 3D vessel surface from the skeleton.

4 Experiments

Dataset and Data Preparation. We conducted experiments on two real-world datasets and one synthetic dataset, all of which are publicly available. (a) ImageCAS [28]: 1,000 real 3D CCTA scans of coronary arteries, presenting significant anatomical variability. We are the first study to address this challenging dataset. (b) VascuSynth [7]: A synthetic dataset comprising 120 generated 3D vascular trees with varying numbers of bifurcations. (c) Processed CoW [15,26]: 300 processed 3D vascular meshes of intracranial arteries. All datasets used 90% for training and 10% for testing.

We perform a series of preprocessing steps on 3D volumes, to extract skeletons with radius information and derive key graphs. First, we apply morphological operations to the binary label volumes to obtain both the skeleton and its corresponding surface. Next, we adopt an adaptive mapping [30] method to construct a key graph and build a maximum spanning tree, where the root, leaf, and branch nodes are identified as vertices of the key graph.

Baselines and Metrics. To evaluate our proposed method, we selected three baseline models for comparison. First, we included a state-of-the-art point cloud generation model [11]. Meanwhile, we incorporated two models specifically designed for vessel generation: TreeDiffusion [17] (D = 128, L = 5) and VesselVAE [6]. To provide a comprehensive evaluation of vessel reconstruction and generation, we utilize metrics based on both point clouds and graph representations. For point clouds sampled from meshes, we report Jensen-Shannon Divergence (JSD) and Chamfer distance (CD) [1,25] to evaluate generation and reconstruction quality respectively. For assembled skeleton graphs, we focus on examining the geometric and topological properties of skeleton graphs, including the maximum mean discrepancy for degree distributions (Deg.) and the Laplacian spectrums (Spec.). To accurately assess the reconstruction performance of skeleton graphs, which are essentially 3D geometric graphs, we adopt the Graph Wasserstein distance (GWD) from [2].

Implementation Details. All experiments were conducted using PyTorch on an NVIDIA A800 GPU with the Adam optimizer. The training process is conducted in two sequential stages. In *stage 1*, we set the latent space to 512 dimensions. We started with a learning rate of 0.001, which was reduced by 0.2 every

100 epochs. The model was trained for 20k epochs, with a batch size of 128, converging in approximately 12 h. In *stage 2*, we set the latent space to 64 dimensions and the batch size to 512. The learning rate was set to 0.0002, training for 2k epochs, which required about 3 h to train. We set transformer blocks to 4 layers with 4 multi-head self-attention in both the encoder and decoder, and the maximum sequence length was set to 200.

Experimental Result. Table 1 compares our method with three state-of-the-art approaches under point cloud-based and graph-based metrics. Our model achieves competitive performance across most tasks, particularly in graph-based evaluation of structural fidelity and topological consistency. In contrast, VesselVAE struggles with complex branching involving numerous sequential points, resulting in consistently sub-optimal performance. PointDiffusion demonstrates the strongest reconstruction metrics across all datasets. However, it falls short on generation metrics and sample quality, indicating an inability to adequately capture vascular geometry and topology. We also observed that TreeDiffusion reports a high JSD score on the ImageCAS dataset but shows poor qualitative generation results. Because JSD is primarily designed to evaluate the spatial distribution of point clouds, however, it fails to account for the accuracy of their topological structures.

Although PointDiffusion and TreeDiffusion outperform our method on certain metrics, a visual comparison of their reconstruction and generation samples reveals some fundamental problems in their modeling techniques. Some reconstruction and generation samples are visualized in Figs. 4 and 5. Particularly,

Table 1. Comparison of vessel reconstruction and generation on point- and graph-based evaluation metrics. CD and JSD are multiplied by 10^3. The best results are denoted in **Bold**. The second-best result is Underlined.

Dataset	Method	Point-based		Graph-based		
		JSD	CD	Deg.	Spec.	GWD
ImageCAS	VesselVAE	92.0	89.8	1.259	1.165	0.141
	PointDiffusion	82.2	**1.1**	–	–	–
	TreeDiffusion	**31.9**	36.2	1.092	0.188	0.087
	Our	50.1	24.4	**0.601**	**0.079**	**0.029**
VascuSynth	VesselVAE	87.4	59.1	1.911	1.207	0.115
	PointDiffusion	87.1	**7.0**	–	–	–
	TreeDiffusion	42.2	35.1	0.445	**0.099**	0.083
	Our	**40.1**	34.8	**0.190**	0.159	**0.049**
CoW	VesselVAE	88.9	35.6	1.740	1.037	0.025
	PointDiffusion	91.0	**1.0**	–	–	–
	TreeDiffusion	47.7	20.9	**0.726**	0.277	0.038
	Our	**44.1**	15.1	1.475	**0.182**	**0.017**

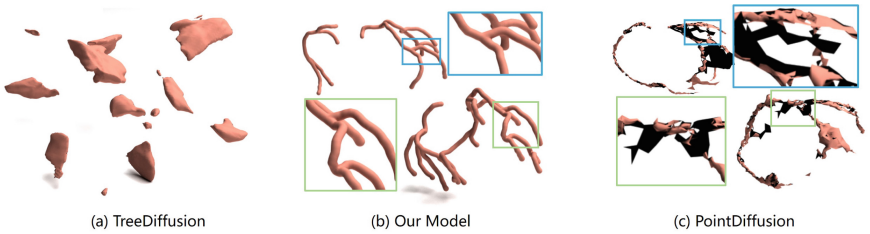


Fig. 4. Reconstruction result from three different methods on ImageCAS dataset. Our approach produces more robust and anatomically consistent results compared to point cloud-based and INR-based methods.

due to the complex data distribution in the ImageCAS dataset, all comparison models failed to successfully generate the basic morphology of the vessels. As shown in Fig. 4, our model not only replicates complex branching structures with high accuracy but also captures subtle morphological variations in the vessels. Point cloud-based methods, by contrast, have difficulty distinguishing between the interior and exterior surfaces of the vessels, leading to multiple holes in the reconstructed meshes and negatively impacting downstream tasks such as vascular analysis. Our skeleton graph-based approach effectively handles complex branching structures and accurately captures morphological variations, underscoring its robustness under real-world conditions. By directly leveraging skeleton and radius information, our method naturally avoids the pitfalls of point cloud-based reconstructions and produces more robust mesh reconstructions for tubular structures.

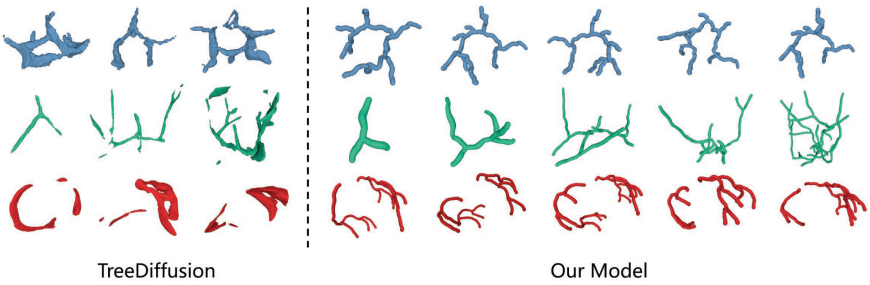


Fig. 5. Examples of generation results from TreeDiffusion and our model on CoW, VascuSynth, and ImageCAS datasets (from top to bottom).

Figure 5 compares the generative performance of our approach against the highly competitive TreeDiffusion, using TreeDiffusion's best-performing samples. As shown, TreeDiffusion often produces irregular, block-like shapes and disconnected components across all datasets, indicating structural anomalies. In

contrast, our model maintains vascular continuity and produces more realistic, anatomically consistent vessel networks.

5 Conclusion

In this paper, we propose a hierarchical part-based framework for 3D vessel generation that separates the global tree-like structure from local geometries. Our approach proceeds in three stages: first, we employ a recursive variational autoencoder to construct a key graph to capture the vascular hierarchy. Subsequently, we introduce a transformer-based VAE to synthesize detailed vessel segments, which are then assembled into a complete vessel in the final stage. Experimental results on three public datasets demonstrate that our model consistently preserves vascular continuity and authentic local curve characteristics.

Acknowledgments. This work is supported in part by the Natural Science Foundation of China (Grant 62371270).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.J.: Learning representations and generative models for 3D point clouds. In: Proceedings of the 35th International Conference on Machine Learning (ICML), pp. 40–49. PMLR (2018)
2. Belli, D., Kipf, T.: Image-conditioned graph generation for road network extraction. arXiv preprint arXiv:1910.14388 (2019)
3. Chang, A.X., et al.: ShapeNet: an information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
4. Chaudhuri, S., Ritchie, D., Wu, J., Xu, K., Zhang, H.: Learning generative models of 3d structures. In: Computer Graphics Forum, vol. 39, pp. 643–666. Wiley Online Library (2020)
5. Dong, C., Xu, S., Dai, D., Zhang, Y., Zhang, C., Li, Z.: A novel multi-attention, multi-scale 3D deep network for coronary artery segmentation. Med. Image Anal. **85**, 102745 (2023)
6. Feldman, P., Fainstein, M., Siless, V., Delrieux, C., Iarussi, E.: VesselVAE: recursive variational autoencoders for 3D blood vessel synthesis. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14220, pp. 67–76. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43907-0_7
7. Hamarneh, G., et al.: Vascusynth: simulating vascular trees for generating volumetric image data with ground-truth segmentation and tree analysis. Comput. Med. Imaging Graph. **34**(8), 605–616 (2010)
8. Hochmuth, A., Spetzger, U., Schumacher, M.: Comparison of three-dimensional rotational angiography with digital subtraction angiography in the assessment of ruptured cerebral aneurysms. Am. J. Neuroradiol. **23**(7), 1199–1205 (2002)
9. Li, J., Niu, C., Xu, K.: Learning part generation and assembly for structure-aware shape synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11362–11369 (2020)

10. Li, J., Xu, K., Chaudhuri, S., Yumer, E., Zhang, H., Guibas, L.: Grass: generative recursive autoencoders for shape structures. ACM Trans. Graph. (TOG) **36**(4), 1–14 (2017)
11. Luo, S., Hu, W.: Diffusion probabilistic models for 3D point cloud generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2837–2845 (2021)
12. Lyu, X., Cheng, L., Zhang, S.: The RETA benchmark for retinal vascular tree analysis. Sci. Data **9**(1), 397 (2022)
13. Mo, K., et al.: StructureNet: hierarchical graph networks for 3D shape generation. **38**(6) (2019)
14. Paetzold, J.C., et al.: Whole brain vessel graphs: a dataset and benchmark for graph learning and neuroscience. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2021)
15. Prabhakar, C., et al.: 3D vessel graph generation using denoising diffusion. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15011, pp. 3–13. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72120-5_1
16. Schor, N., Katzir, O., Zhang, H., Cohen-Or, D.: CompoNet: learning to generate the unseen by part synthesis and composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8759–8768 (2019)
17. Sinha, A., Hamarneh, G.: TrIND: representing anatomical trees by denoising diffusion of implicit neural fields. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15012. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72390-2_33
18. Socher, R., Huval, B., Bath, B., Manning, C.D., Ng, A.: Convolutional-recursive deep learning for 3D object classification. Adv. Neural Inf. Process. Syst. **25** (2012)
19. Socher, R., Lin, C.C., Manning, C., Ng, A.Y.: Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 129–136 (2011)
20. Vahdat, A., et al.: LION: latent point diffusion models for 3D shape generation. Adv. Neural. Inf. Process. Syst. **35**, 10021–10039 (2022)
21. Wang, Z., et al.: Cardiovascular medical image and analysis based on 3D vision: a comprehensive survey. Meta-Radiology **2**(4), 100102 (2024)
22. Wang, Z., Yi, R., Wen, X., Zhu, C., Xu, K.: Cardiovascular medical image and analysis based on 3D vision: a comprehensive survey. Meta-Radiol. 100102 (2024)
23. Wolterink, J.M., Leiner, T., Isgum, I.: Blood vessel geometry synthesis using generative adversarial networks. arXiv preprint arXiv:1804.04381 (2018)
24. Xu, Q.C., Mu, T.J., Yang, Y.L.: A survey of deep learning-based 3D shape generation. Comput. Vis. Media **9**(3), 407–442 (2023)
25. Yang, G., Huang, X., Hao, Z., Liu, M.Y., Belongie, S., Hariharan, B.: PointFlow: 3D point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4541–4550 (2019)
26. Yang, K., et al.: Benchmarking the cow with the TopCow challenge: topology-aware anatomical segmentation of the circle of Willis for CTA and MRA. ArXiv, pp. arXiv–2312 (2024)
27. Yao, L., et al.: Tag-net: topology-aware graph network for centerline-based vessel labeling. IEEE Trans. Med. Imaging **42**(11), 3155–3166 (2023)
28. Zeng, A., et al.: ImageCAS: a large-scale dataset and benchmark for coronary artery segmentation based on computed tomography angiography images. Comput. Med. Imaging Graph. **109**, 102287 (2023)

29. Zhang, G., Dong, C., Li, Y.: Topology-preserving hard pixel mining for tubular structure segmentation. In: 34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, 20–24 November 2023. BMVA (2023)
30. Zhang, G., Li, Y.: A geometric algorithm for blood vessel reconstruction from skeletal representation. In: Peng, W., Cai, Z., Skums, P. (eds.) ISBRA 2024. LNCS, vol. 14954, pp. 114–126. Springer (2024). https://doi.org/10.1007/978-981-97-5128-0_10
31. Zhao, J., Chen, X., Xiong, Z., Zha, Z.J., Wu, F.: Graph representation learning for large-scale neuronal morphological analysis. IEEE Trans. Neural Netw. Learn. Syst. **35**(4), 5461–5472 (2022)



Hierarchical Spatio-Temporal Segmentation Network for Ejection Fraction Estimation in Echocardiography Videos

Dongfang Wang, Jian Yang(✉), Yizhe Zhang, and Tao Zhou(✉)

PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{dongfangwang,csjyang}@njust.edu.cn, taozhou.ai@gmail.com

Abstract. Automated segmentation of the left ventricular endocardium in echocardiography videos is a key research area in cardiology. It aims to provide accurate assessment of cardiac structure and function through Ejection Fraction (EF) estimation. Although existing studies have achieved good segmentation performance, their results do not perform well in EF estimation. In this paper, we propose a Hierarchical Spatio-temporal Segmentation Network (HSS-Net) for echocardiography video, aiming to improve EF estimation accuracy by synergizing local detail modeling with global dynamic perception. The network employs a hierarchical design, with low-level stages using convolutional networks to process single-frame images and preserve details, while high-level stages utilize the Mamba architecture to capture spatio-temporal relationships. The hierarchical design balances single-frame and multi-frame processing, avoiding issues such as local error accumulation when relying solely on single frames or neglecting details when using only multi-frame data. To overcome local spatio-temporal limitations, we propose the Spatio-temporal Cross Scan (STCS) module, which integrates long-range context through skip scanning across frames and positions. This approach helps mitigate EF calculation biases caused by ultrasound image noise and other factors. We achieved state-of-the-art results on three datasets. Our code is available at https://github.com/DF-W/HSS-Net.

Keywords: Echocardiography · Ejection fraction · Segmentation

1 Introduction

The primary of echocardiography analysis is to assess cardiac function, where accurate segmentation of the left ventricular endocardium is essential to measure the heart's Ejection Fraction (EF) [1,3]. However, achieving automatic echocardiography segmentation presents significant challenges. Firstly, the quality of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 268–278, 2026.
https://doi.org/10.1007/978-3-032-04947-6_26

ultrasound images is often compromised by noise, artifacts, and boundary blurring [18,19]. Secondly, the heart undergoes complex motion and deformation during each heartbeat cycle.

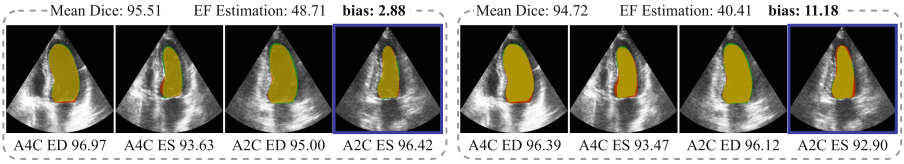


Fig. 1. Left ventricular segmentation maps for ED and ES frames from A2C and A4C views, with Dice segmentation metrics displayed at the bottom. **It is evident that smaller segmentation errors in the key regions (base and apex) within the blue box lead to greater EF calculation deviations**. The green, red, and yellow represent the ground truth, prediction, and overlapping regions, respectively.

Moreover, physicians typically only annotate the End-Diastolic (ED) and End-Systolic (ES) key frames, resulting in limited and sparse annotation data. Therefore, the automated segmentation method is trained on limited annotated data, providing a reliable foundation for the accurate evaluation of cardiac function clinical metrics.

Recently, deep learning methods for echocardiography video segmentation have emerged rapidly. Although effectiveness achieved, their results are not satisfactory when applied to ejection fraction estimation [16,21,22,24]. Many studies focus on 2D segmentation of ED and ES frames, with numerous unlabeled frames left unused, resulting in an inability to capture the continuity of cardiac motion [8,12,27]. To address this, Painchaud *et al.* [13] proposed a post-processing framework that leverages cardiac anatomical priors to enhance inter-frame consistency. However, its performance is highly dependent on the quality of the initial segmentation. Similarly, Wei *et al.* [20] used generated pseudo-labels for collaborative learning of segmentation and tracking, while the method is constrained by the quality of the pseudo-labels. The Transformer architecture, with multi-head self-attention [17], has been widely adopted in video object segmentation [2]. Temporal continuity across frames offers valuable segmentation cues. To exploit this, some methods add specialized modules atop self-attention to capture temporal information. However, this may cause over-reliance on multi-frame relations while overlooking fine details. Moreover, modeling long sequences with many labeled frames introduces heavy computational costs.

In this paper, we propose a Hierarchical Spatio-temporal Segmentation Network (HSS-Net) for echocardiography video segmentation, enhancing the precision of ejection fraction estimation. The low-level stages utilize convolutional networks to process single-frame images, preserving fine details, while the high-level stages employ the Mamba architecture to capture spatio-temporal relationships across multiple frames. By better integrating local and global features, the

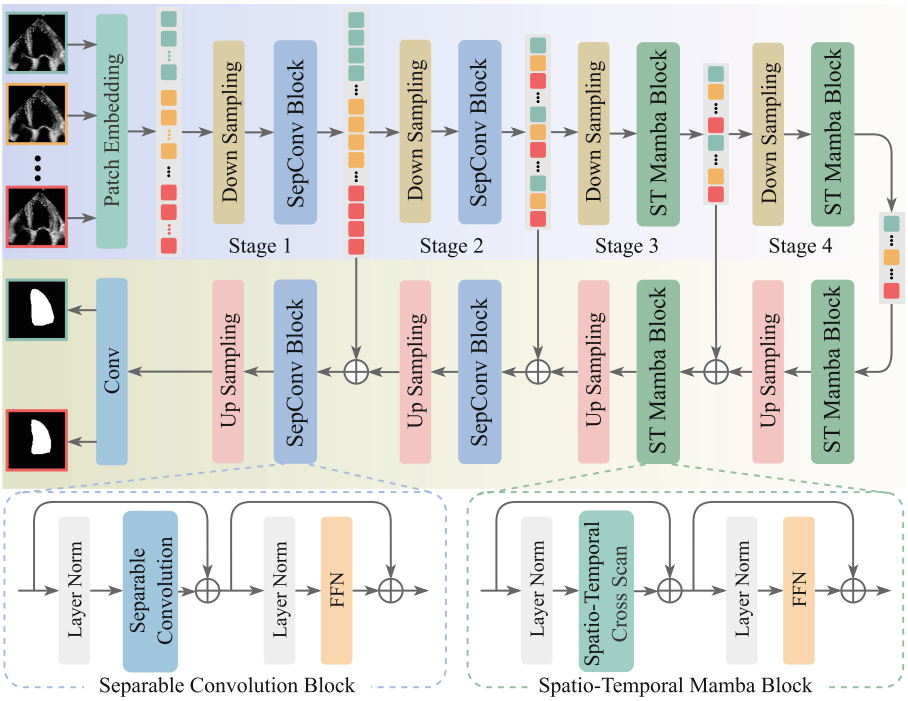


Fig. 2. Illustration of the proposed HSS-Net framework, which is a symmetric Encoder-Decoder architecture.

model reduces segmentation errors in key regions (*e.g.*, the base and apex) and mitigates volume calculation biases caused by local errors, as shown in Fig. 1. The network captures inter-frame motion patterns, enhancing the temporal consistency of the segmentation results and improving EF estimation stability. To fully utilize unlabeled data and strengthen the capture of dynamic cues, we propose a spatio-temporal cross scan module, which captures dynamic cues from different spatio-temporal perspectives. This module integrates long-range dependencies, such as apex motion changes and the correlation of lateral wall contraction, through a skip-connections-based spatio-temporal scan. Additionally, global dynamic modeling reduces sensitivity to interference factors, preventing abnormal jitter in segmentation boundaries and ensuring the reliability of EF estimation.

Contributions: i) We propose a hierarchical spatio-temporal segmentation framework, which combines convolutional and Mamba architectures for detailed local and temporal processing. **ii)** A STCS Module is proposed to capture dynamic cues from multiple perspectives, enhancing robustness and accuracy. **iii)** We present a novel jump scanning mechanism, which breaks local correlations to integrate global information, improving generalization across diverse samples. **iv)** Experimental results show the effectiveness of the proposed model.

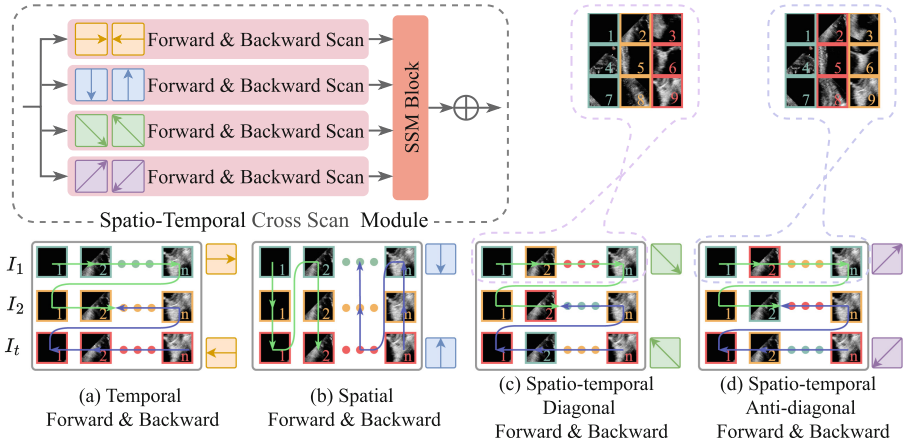


Fig. 3. Spatio-temporal cross scan module and its scanning sequence (I_t denotes the t-th frame in the video clip, and n denotes the number of patch sequences).

2 Methodology

Overview. The architecture of our HSS-Net, illustrated in Fig. 2, consists of two modules: Encoder and Decoder. The first and second stages of the Encoder are primarily composed of stacked separable convolution blocks for low-level feature capture within single-frame images. The third and fourth stages are mainly composed of stacked spatio-temporal Mamba blocks for high-level feature capture across multiple frames. The Decoder architecture is symmetric to the Encoder, with a different number of stacked blocks. It is used to fuse multi-scale features and predict the segmentation mask. Specifically, given a video clip of T frames, denoted as $\mathbf{V} = \{I_1, I_2, ..., I_T\}$, we first apply patch embedding to divide these frames into different patches. The patch sequence is then fed into the encoder, resulting in the i-th stage feature $\mathbf{F_i}$, with a size of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$, where H and W represent the height and width of the original frames, respectively, and $i \in \{1, 2, 3, 4\}$. Finally, these multi-scale features are passed to the decoder, where operations such as inter-frame perception, intra-frame perception, and up-sampling are performed to generate the predicted segmentation results.

2.1 Separable Convolution Block

Fig. 2 illustrates the structure of the separable convolution block. For the i-stage feature embedding of low-level $\mathbf{F_i} \in \mathbb{R}^{T \times C_i \times H_i \times W_i}, i \in \{1, 2\}$ of the given video clip, layer normalization ($\mathcal{LN}(\cdot)$) is performed before performing separable convolution operations ($\mathcal{SC}(\cdot)$). We follow the inverted separable convolution module from MobileNetV2 [15]. Subsequently, feeding the result into the FeedForward Network ($\mathcal{FFN}(\cdot)$) layer to capture low-level features such as edges and textures

in single-frame images. The output features are either passed to the next stacked block or forwarded to the next stage. This process can be expressed as:

$$\mathbf{F_i} = \mathcal{SC}(\mathcal{LN}(\mathbf{F_i})) + \mathbf{F_i}, \quad \mathbf{F_i} = \mathcal{FFN}(\mathcal{LN}(\mathbf{F_i})) + \mathbf{F_i}. \tag{1}$$

2.2 Spatio-Temporal Mamba Block

The structure of the spatial-temporal Mamba block is illustrated in Fig. 2. For the i-stage feature embedding of high-level $\mathbf{F_i} \in \mathbb{R}^{T \times C_i \times H_i \times W_i}, i \in \{3, 4\}$ of the given video clip, we transpose the channel and time dimensions and flatten the spatio-temporal feature embedding into a one-dimensional long sequence $\mathbf{s_i} \in \mathbb{R}^{C_i \times T H_i W_i}$. This sequence $\mathbf{s_i}$ is then fed into the spatio-temporal cross scan ($\mathcal{STCS}(\cdot)$) module and FFN layers. The STCS module establishes long-range dependencies both inter-frame and intra-frame from different spatio-temporal perspectives. This process can be defined as follows:

$$\mathbf{s_i} = \mathcal{STCS}(\mathcal{LN}(\mathbf{s_i})) + \mathbf{s_i}, \quad \mathbf{s_i} = \mathcal{FFN}(\mathcal{LN}(\mathbf{s_i})) + \mathbf{s_i}. \tag{2}$$

Finally, the output feature sequences are reshaped back to their original shape, and after down-sampling, the feature embeddings are passed to the next stage.

Spatio-temporal Cross Scan Module: As shown in Fig. 3, the STCS module with state space model (S6) [6] is designed for the spatio-temporal sequence modeling of video frames. It selectively scans the input sequence from various spatio-temporal perspectives, capturing intricate spatio-temporal relationships and providing a comprehensive understanding of the context.

To better understand and explore the spatio-temporal relationships among frames, we first unfold each frame's patches into sequences along rows and columns. As illustrated in Fig. 3, the patches of each frame are unfolded along rows to form temporal sequences, while the patches at the same position in different frames are unfolded along columns to form spatial sequences. The STCS module offers four different scanning modes: temporal, spatial, spatio-temporal diagonal, and spatio-temporal anti-diagonal. As depicted in Fig. 3(a), the module scans simultaneously along the temporal sequence in both forward and backward directions to explore bidirectional temporal dependencies. In Fig. 3(b), the module scans simultaneously along the spatial sequence in both upward and downward directions to explore bidirectional spatial dependencies. The selective spatio-temporal scanning explicitly considers both intra-frame and inter-frame coherencies and leverages the SSM to establish long-range dependencies of intra-frame and inter-frame.

The heart motion during a heartbeat is not fully synchronous, leading to similarities and differences in information among video frames. To disrupt local data correlations, integrate global information, capture diverse features, and enhance the model's generalization and understanding of cardiac structure and motion, we propose a new spatio-temporal diagonal and anti-diagonal scanning method (see Fig. 3(c) and (d)). We rearrange spatial sequence positions into diagonal

and anti-diagonal patterns and scan temporally in forward and backward directions. This cross-frame and cross-position scanning improves the model's ability to integrate global information and understand cardiac motion and structure.

Loss Function: During training, our loss function includes the Dice loss $\mathcal{L}_{dice}$ [11] and binary cross-entropy loss $\mathcal{L}_{bce}$. Thus, the total loss function $\mathcal{L}_{total} = \alpha\mathcal{L}_{dice}(P, G) + (1-\alpha)\mathcal{L}_{bce}(P, G)$, where G denotes the ground-truth, P denotes the predicted masks, and the balance weight $\alpha = 0.8$ in our experiments.

Table 1. The quantitative results on the CAMUS. FLOPs represent the average computational complexity per frame at a size of 256×256.

Methods	Params	FLOPs	corr	bias ± std	Dice	HD95
UNet++ [28]	26.9M	37.7G	81.68	6.05 ± 6.81	91.87	16.16
TransUNet [4]	105.3M	38.6G	86.22	**1.72 ± 6.07**	92.73	13.71
SegFormer [23]	47.4M	20.9G	79.10	7.84 ± 7.21	91.17	18.46
H2Former [7]	33.7M	33.1G	82.35	5.79 ± 6.87	91.78	16.06
SSCF [22]	53.7M	15.1G	84.48	4.11 ± 6.37	92.59	14.18
PKEchoNet [21]	25.7M	7.2G	76.20	4.13 ± 8.45	93.02	12.93
VideoMamba [10]	75.6M	22.0G	75.21	8.02 ± 8.00	91.53	16.43
Vivim [25]	59.6M	20.6G	78.09	5.75 ± 7.35	92.79	12.74
HSS-Net (Ours)	31.2M	5.6G	**90.47**	2.43 ± 5.02	**93.89**	**11.29**

Table 2. The quantitative results on the EchoNet-Pediatric and EchoNet-Dynamic datasets. The HD95 metrics are reported in pixels.

Methods	EchoNet-Pediatric				EchoNet-Dynamic			
	corr	bias ± std	Dice	HD95	corr	bias ± std	Dice	HD95
UNet++ [28]	69.33	7.62 ± 10.16	90.73	3.65	73.91	9.43 ± 9.02	91.50	2.99
TransUNet [4]	73.09	6.54 ± 9.81	91.11	3.52	74.17	4.67 ± 9.51	91.92	2.96
SegFormer [23]	66.72	6.25 ± 10.77	91.10	3.52	73.12	7.07 ± 9.37	92.07	2.90
H2Former [7]	69.77	6.24 ± 10.06	90.89	3.58	74.78	6.10 ± 9.23	91.68	3.12
SSCF [22]	63.29	5.28 ± 11.82	91.07	3.50	74.87	6.26 ± 9.16	92.35	2.80
PKEchoNet [21]	65.04	6.23 ± 11.21	91.00	3.57	75.43	4.34 ± 9.50	92.45	2.71
VideoMamba [10]	67.34	6.39 ± 11.39	91.06	3.50	78.62	4.50 ± 8.29	92.48	2.75
Vivim [25]	69.92	5.59 ± 10.31	91.12	3.46	81.12	7.02 ± 7.47	92.46	2.73
HSS-Net (Ours)	**76.91**	**1.29 ± 8.68**	**91.90**	**3.23**	**84.50**	**0.95 ± 6.75**	**92.67**	**2.66**

3 Experiments

Datasets: In this study, three publicly available echocardiography video datasets are used, namely CAMUS [9], EchoNet-Pediatric [14], and EchoNet-Dynamic [24]. • **CAMUS** comprises 500 cases acquired at the University Hospital of St Etienne (France), each including 2D apical 2-chamber and 4-chamber view videos, with annotations provided for all frames. • **EchoNet-Pediatric** is collected from Lucile Packard Children's Hospital Stanford, including 7,643 video clips from 1,958 patients aged 0 to 18 years. This dataset consists of either parasternal short axis or apical 4-chamber views, with only the ED and ES frames annotated. • **EchoNet-Dynamic** consists of 10,030 apical 4-chamber view echocardiography videos collected from Stanford University Hospital, with only ED and ES frames annotated for each video. We uniformly sample 10 frames of each video clip from datasets, following previous research [5,21]. The video clips are cropped to ensure that the ED frame is the first and the ES frame is the last, thereby capturing a complete heartbeat cycle. The frame size was adjusted to 256×256, and only the annotations from the ED and ES frames were used for training and evaluation. For the EchoNet-Dynamic dataset, we used the original dataset splits. For the other datasets, we follow related studies [24,26] and split the data into training, validation, and testing sets in an 8:1:1 ratio.

Evaluation Metrics: We report three statistical metrics for left ventricular ejection fraction. The estimation methods vary due to different views provided by the datasets. For the EchoNet-Dynamic and EchoNet-Pediatric datasets, both ground truth and predicted ejection fractions are obtained using the Simpson's single-plane method of disks [9]. For the CAMUS dataset, the Simpson's biplane method of disks [24] is used to calculate ejection fractions. We follow [5,21] and calculate the Pearson correlation coefficient (corr), mean bias (bias), and standard deviation (std) for the predicted and ground truth ejection fractions. Additionally, we employed two widely used segmentation evaluation metrics: the mean Dice coefficient (Dice) and Hausdorff Distance at 95% (HD95).

Implementation Details: Our model is implemented using PyTorch and is trained or inferred on two NVIDIA RTX 4090 GPU. We train our model end-to-end using the Adam optimizer and employ the cosine annealing strategy to adjust the learning rate. The maximum and minimum learning rates are set to 1e-4 and 1e-5, respectively, and the maximum training epoch is set to 120. During training, we apply gamma augmentation, random scaling, random rotation, and random contrast adjustments, each with a probability of 0.5.

3.1 Comparison with State-of-the-Art Methods

Quantitative Comparisons: The quantitative results for the three datasets are presented in Tables 1 and 2. It can be seen that our model outperforms other methods in all datasets. However, image-based methods (The first four) still exhibit competitive performance on certain datasets. These methods focus on

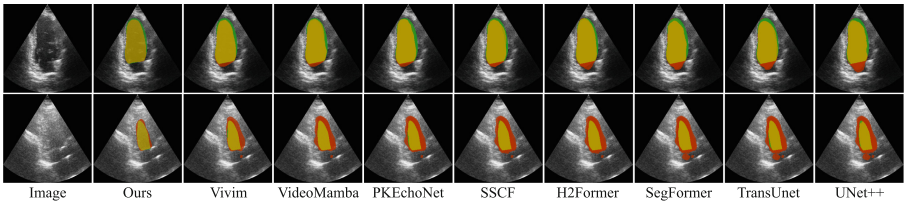


Fig. 4. Visualization of segmentation maps for different models.

Table 3. Quantitative results of ablation studies.

Settings	CAMUS				EchoNet-Dynamic			
	corr	bias ± std	Dice	HD95	corr	bias ± std	Dice	HD95
Image-level	83.48	5.28 ± 6.71	92.93	13.91	74.79	6.00 ± 9.34	91.93	2.97
Video-level	80.67	4.48 ± 7.04	93.04	13.09	78.01	4.62 ± 7.92	92.03	2.91
w/o Temporal	83.73	4.20 ± 6.67	93.02	14.26	78.42	4.80 ± 8.02	92.10	2.90
w/o Spatio	80.44	5.05 ± 7.44	93.26	12.57	77.86	4.76 ± 8.21	92.32	2.79
w/o ST Diagonal	86.69	4.05 ± 5.99	93.21	12.40	79.65	6.03 ± 7.70	92.12	2.84
w/o ST Anti-diagonal	88.09	2.95 ± 5.57	93.27	12.28	81.11	4.85 ± 7.54	92.26	2.82
HSS-Net (Ours)	**90.47**	**2.43 ± 5.02**	**93.89**	**11.29**	**84.50**	**0.95 ± 6.75**	**92.67**	**2.66**

capturing local features within single-frame images, whereas video-based methods (The last four), although proficient in capturing temporal information, may not fully exploit their advantages on datasets with minimal inter-frame differences or subtle dynamic changes. Our model hierarchically processes single-frame and multi-frame information, balancing both detailed features and dynamic changes in echocardiography. This approach is particularly advantageous for ejection fraction estimation. Additionally, it maintains an optimal balance between model performance and computational complexity.

Qualitative Comparisons: We present visualizations of several challenging cases. As shown in Fig. 4, these sample images exhibit artifacts, speckle noise, and blurred boundaries. Such challenging conditions mislead most of the compared models, resulting in missed or misclassified regions. In contrast, our model accurately locates the regions and delineates the boundaries. These visualizations further demonstrate that our approach can achieve better segmentation results and robustly handle poor-quality images.

3.2 Ablation Study

Effectiveness of Hierarchical Design: To validate the effectiveness of the hierarchical design, we perform two sets of ablation experiments. One set processes single-frame images at all stages using only separable convolution blocks (labeled as Image-level). The other set processes multi-frame images at all stages

using only spatio-temporal Mamba blocks (labeled as Video-level). As shown in Table 3, the model performance in both experiments decreased to varying degrees, especially in the key metric of Pearson correlation for ejection fraction estimation. This demonstrates that by extracting fine-grained details at the low level and modeling cross-frame temporal relationships at the high level, the model effectively handles subtle differences across different conditions, providing a more reliable foundation for clinical assessment of cardiac function.

Effectiveness of Spatio-temporal Cross Scan Module: We investigate the effect of each mode in the STCS module. As shown in Table 3, all performance metrics exhibit varying degrees of decline compared to our full method. Temporal scanning captures sequential dynamic information during cardiac motion, which is crucial to accurately modeling heart movement patterns. Spatial scanning aids the model in understanding changes at the same location across different time frames. This enhances the model's ability to perceive spatial consistency features. Spatio-temporal diagonal and anti-diagonal scanning effectively capture complex interactions between temporal and spatial dimensions, enhancing the model's ability to integrate spatio-temporal information.

4 Conclusion

We propose a novel method HSS-Net that employs a hierarchical design. The low-level stages use convolutions to extract local details from single-frame images, while the high-level stages leverage the Mamba architecture to process spatio-temporal information across multiple frames. By handling single-frame and multi-frame information hierarchically, the model's accuracy and robustness are enhanced. Extensive experimental results demonstrate that our method achieves state-of-the-art results on three benchmark datasets.

Acknowledgement. This work was in part supported by the National Science Fund of China under Grant Nos. 62172228, U24A20330, and 62361166670.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akkus, Z., et al.: Artificial intelligence (ai)-empowered echocardiography interpretation: a state-of-the-art review. J. Clin. Med. **10**(7), 1391 (2021)
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: ViViT: a video vision transformer. In: International Conference on Computer Vision, pp. 6836–6846 (2021)
3. Chen, C., et al.: Deep learning for cardiac image segmentation: a review. Front. Cardiovas. Med. **7**, 25 (2020)
4. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

5. Deng, X., Wu, H., Zeng, R., Qin, J.: MemSAM: taming segment anything model for echocardiography video segmentation. In: Computer Vision and Pattern Recognition, pp. 9622–9631 (2024)
6. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
7. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: an efficient hierarchical hybrid transformer for medical image segmentation. IEEE Trans. Med. Imaging **42**(9), 2763–2775 (2023)
8. Leclerc, S., et al.: Lu-net: a multistage attention network to improve the robustness of segmentation of left ventricular structures in 2-d echocardiography. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **67**(12), 2519–2530 (2020)
9. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Trans. Med. Imaging **38**(9), 2198–2210 (2019)
10. Li, K., et al.: VideoMamba: state space model for efficient video understanding. arXiv preprint arXiv:2403.06977 (2024)
11. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision, pp. 565–571 (2016)
12. Moradi, S., et al.: Mfp-unet: a novel deep learning based approach for left ventricle segmentation in echocardiography. Physica Med. **67**, 58–69 (2019)
13. Painchaud, N., Duchateau, N., Bernard, O., Jodoin, P.M.: Echocardiography segmentation with enforced temporal consistency. IEEE Trans. Med. Imaging **41**(10), 2867–2878 (2022)
14. Reddy, C.D., Lopez, L., Ouyang, D., Zou, J.Y., He, B.: Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. J. Am. Soc. Echocardiogr. **36**(5), 482–489 (2023)
15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetv2: inverted residuals and linear bottlenecks. In: Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
16. Thomas, S., Gilbert, A., Ben-Yosef, G.: Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 380–390. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_37
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Wang, D., Zhou, T., Yang, J.: Hybrid-frequency feature evolution network for endoscopic ultrasound image segmentation. In: International Symposium on Biomedical Imaging, pp. 1–5 (2025)
19. Wang, D., Zhou, T., Zhang, Y., Gao, S., Yang, J.: Frequency-aware interaction network for ultrasound image segmentation. IEEE Trans. Circuits Syst. Video Technol. 1 (2024)
20. Wei, H., Ma, J., Zhou, Y., Xue, W., Ni, D.: Co-learning of appearance and shape for precise ejection fraction estimation from echocardiographic sequences. Med. Image Anal. **84**, 102686 (2023)
21. Wu, H., Lin, J., Xie, W., Qin, J.: Super-efficient echocardiography video segmentation via proxy-and kernel-based semi-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 2803–2811 (2023)
22. Wu, H., Liu, J., Xiao, F., Wen, Z., Cheng, L., Qin, J.: Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. Med. Image Anal. **78**, 102397 (2022)

23. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: simple and efficient design for semantic segmentation with transformers, vol. 34, pp. 12077–12090 (2021)
24. Yang, J., Ding, X., Zheng, Z., Xu, X., Li, X.: GraphEcho: graph-driven unsupervised domain adaptation for echocardiogram video segmentation. In: International Conference on Computer Vision, pp. 11844–11853 (2023)
25. Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)
26. Ye, Z., Chen, T.: P-mamba: marrying perona malik diffusion with mamba for efficient pediatric echocardiographic left ventricular segmentation. arXiv preprint arXiv:2402.08506 (2024)
27. Zamzmi, G., Rajaraman, S., Hsu, L.Y., Sachdev, V., Antani, S.: Real-time echocardiography image analysis and quantification of cardiac indices. Med. Image Anal. **80**, 102438 (2022)
28. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging **39**(6), 1856–1867 (2019)

HybridMamba: A Dual-Domain Mamba for 3D Medical Image Segmentation

Weitong Wu[1], Zhaohu Xing[1], Jing Gong[2,3], Qin Peng[2,3], and Lei Zhu[1,4(✉)]

[1] The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
leizhu@ust.hk
[2] Department of Radiology, Fudan University Shanghai Cancer Center, Shanghai, China
[3] Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China
[4] The Hong Kong University of Science and Technology, Hong Kong, China

Abstract. In the domain of 3D biomedical image segmentation, Mamba exhibits superior performance as it addresses the limitations in modeling long-range dependencies inherent to CNNs and mitigates the abundant computational overhead associated with Transformer-based frameworks when processing high-resolution medical volumes. However, attaching undue importance to global context modeling may inadvertently compromise critical local structural information, thus leading to boundary ambiguity and regional distortion in segmentation outputs. Therefore, we propose the HybridMamba, an architecture employing dual complementary mechanisms: 1) a feature scanning strategy that progressively integrates representations both axial-traversal and local-adaptive pathways to harmonize the relationship between local and global representations, and 2) a gated module combining spatial-frequency analysis for comprehensive contextual modeling. Besides, we collect a multi-center CT dataset related to lung cancer. Experiments on MRI and CT datasets demonstrate that HybridMamba significantly outperforms the state-of-the-art methods in 3D medical image segmentation.

Keywords: State space model · Mamba · Frequency and spatial feature modeling · 3D medical image segmentation

1 Introduction

In clinical diagnostic workflows, achieving voxel-level precision in pathological region segmentation from 3D medical imaging modalities (e.g., CT, MRI) constitutes a critical prerequisite for quantitative disease characterization [25]. While deep learning architectures have demonstrated remarkable potential in medical image segmentation, thereby reducing inter-observer variability and clinician workload, there still exist fundamental limitations. Traditional Convolutional Neural Networks (CNNs) suffer from structural constraints in modeling long-range dependencies within medical image segmentation due to their intrinsic

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 279–289, 2026.
https://doi.org/10.1007/978-3-032-04947-6_27

locality [3]. Deep learning architectures have shown great promise in medical image segmentation by minimizing inter-observer variability and clinician workload. However, traditional Convolutional Neural Networks (CNNs) are fundamentally limited by structural constraints that hinder their ability to model long-range dependencies due to their inherent locality [3]. Recent attempts like 3D UX-Net [12] endeavor to mitigate this issue by expanding effective receptive fields through large-kernel convolutions, but still have constraints in modeling global relationships.

The emergence of Transformer architecture [19] has notably enhanced the modeling of global context through self-attention mechanisms, effectively addressing the limitations of CNNs in capturing long-range dependencies. For instance, UNETR [8] merges the Vision Transformer (ViT) [4] into the encoder to capture contextual information and utilizes a convolutional decoder with multi-scale skip connections to learn local features together for 3D medical image segmentation. Similarly, SwinUNETR [7] employs the SwinTransformer [14] in its encoder to enable efficient extraction of features at multiple resolutions. Nonetheless, these Transformer-based methods imposes significant scalability constraints for high-resolution 3D medical imaging analysis. This computational bottleneck manifests particularly in memory-intensive segmentation tasks like biomedical image segmentation, resulting in suboptimal throughput rates during both training and inference phases despite recent hardware advancements.

Emerging from State Space Model, Mamba [5,21,24] presents an innovative approach to long-range dependency learning through selective mechanism and hardware-efficient algorithms. This paradigm achieves CNN-level training stability coupled with RNN-like inference efficiency [6], all maintained within linear computational complexity. Current advances demonstrate growing applications of Mambas in medical imaging analysis: In 2D contexts, U-Mamba [15] augments standard nnUNet frameworks [10] by integrating directional scanning modules into encoder pathways, significantly enhancing the performance in medical image segmentation. Moreover, Swin-UMamba [13] strategically incorporates large-scale ImageNet pretrained representations, thereby enhancing Mamba's efficacy in clinical image segmentation scenarios. For 3D medical image analysis, SegMamba [27] pioneers voxel-level contextual modeling through its tri-orientated Mamba (ToM) module for feature modeling and gated spatial convolution (GSC) module for the representation enhancement of spatial features. However, Existing 2D methodologies fail to address cross-slice correlations inherent in volumetric datasets effectively; meanwhile SegMamba exhibit suboptimal coordination between localized feature retention and global context integration when processing hierarchically partitioned image patches.

In this paper, we propose HybridMamba, a hierarchical architecture addressing two fundamental challenges in 3D medical image analysis: 1) multi-resolution contextual preservation across spatial scales; and 2) synergistic integration of frequency-spatial feature representations. To our best knowledge, this is the first method to deploy the feature extraction strategy in both the frequency domain and the spatial domain to facilitate more accurate and robust representations on

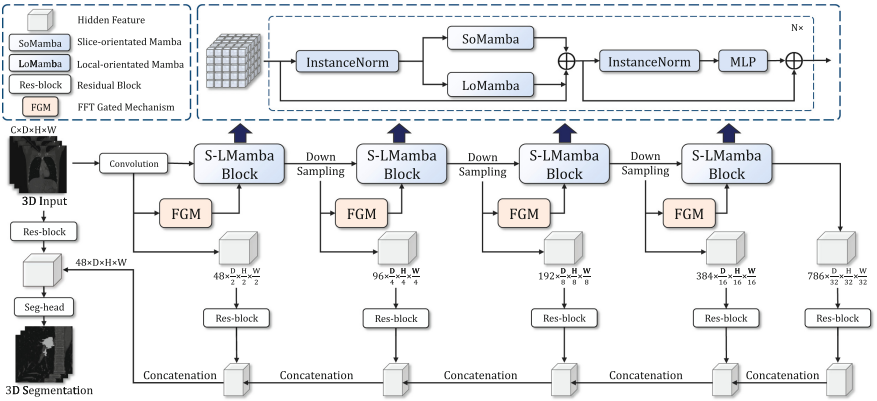


Fig. 1. The overview of the proposed HybridMamba. The encoder consists of a multiple S-LMamba blocks for balancing local and global features and an FFT Gate Mechanism (FGM) for merging the features from spatial and frequency domain dynamically according to the characteristics of different layers.

multiple 3D medical image segmentation tasks. Consider the non-negligibility of relationships between global and local information, we design a fused ergodic mechanism named SoMamba (Slice-oriented Mamba) and LoMamba(Local-oriented Mamba) to increase the sequential modeling of 3D features. Following this, we further propose the gated mechanism that dynamically weights frequency-transformed features against spatially encoded patterns prior to cascaded Mamba processing stages. We verify the superior performance of the proposed HybridMamba on both public MRI dataset and our collected CT dataset. The results of experiments showcase the efficiency of our method.

2 Method

In this section, we introduce the implementation approach of HybridMamba, which is built upon SegMamba [27] with the particular improvements on encoder. We incorporate the frequency domain feature into encoder modeling by utilizing the gated mechanism before transforming from two kinds of traversing pathways to sequence. Figure 1 demonstrates the overview of HybridMamba. The details of the encoder will be further described in this section.

2.1 Slice-Local Mamba (S-LMamba) Block

It is of great necessity to utilize Mamba to model long-range dependencies within 3D high-resolution biomedical image segmentation [20,26,28]. While transformer architectures [22,23] effectively capture global information, they impose substantial computational costs when processing excessively long feature sequences. Accordingly, Mamba is utilized especially for high-resolution biomedical image field. It fully leverages long-distance modeling while being exponentially more

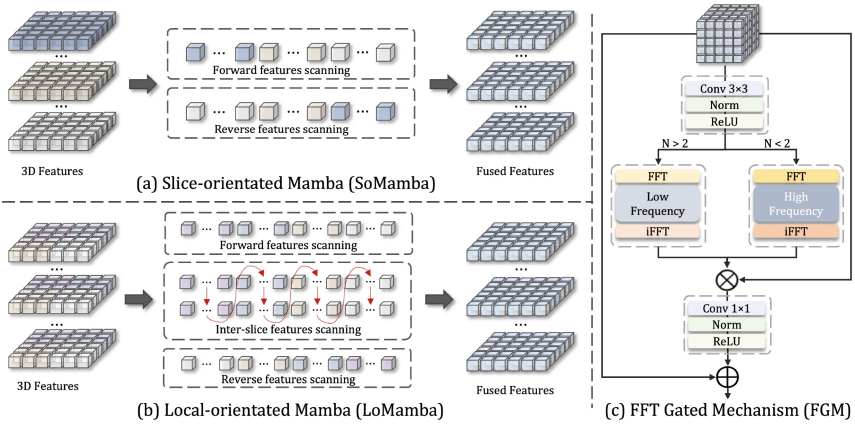


Fig. 2. The left side (a) and (b) shows the detailed implementation in SoMamba and LoMamaba. The right side (c) demonstrates the specific layers in the FGM.

computationally efficient than the Transformer framework. However, prioritizing long-distance dependencies overlooks the extraction of local segmentation information within and between slices, potentially leading to incoherent partitioning and inadequate semantic understanding in the model. Drawing inspiration from the scanning method presented in Local Mamba [9], we design an S-LMamba Block to partition all slices of the medical image, generating windows of a desired size according to different scales in each layer. We also recognize the importance of long-distance dependencies throughout the training process. Therefore, we introduce a slice-oriented traversing strategy to enhance the learning of contextual information within the medical image. This approach allows Mamba to extract more relevant segmentation information from the sequences.

The encoder illustrated in Fig. 1 is composed of multiple S-LMamba blocks and FGM modules. After the initial preprocessing of the input 3D volume $I \in \mathbb{R}^{48\times D\times H\times W}$ busing a large convolutional layer with a kernel size of $7 \times 7 \times 7$, the resulting 3D feature $x_0 \in \mathbb{R}^{48\times \frac{D}{2}\times \frac{H}{2}\times \frac{W}{2}}$ is processed through the S-LMamba blocks and FGM modules, along with subsequent down-sampling layers. For the n^{th} S-LMamba Block, the computational operation can be defined as:

$$\tilde{x}_n^l = \mathrm{MLP}(\mathrm{IN}(\mathrm{SoMamba}\left(\mathrm{LN}\left(\hat{x}_n^l\right)\right) + \mathrm{LoMamba}\left(\mathrm{LN}\left(\hat{x}_n^l\right)\right) + \hat{x}_n^l)) + \tilde{x}_n^l, \quad (1)$$

where the SoMamba and LoMamba refer to the proposed Slice-oriented and Local-oriented Mamba module, respectively, which will be discussed next. $l \in \{0, 1, ..., n-1\}$, LN refers to layer normalization, and IN refers to instance normalization. MLP refers to the multi-layer perceptron for enriching the feature representation.

Slice-Orientated Mamba (SoMamba) and Local-Orientated Mamba (LoMamba). The Mamba layer captures feature dependencies by flattening the 3D features into a 1D sequence. Managing the order of this flattening pro-

cess is crucial, as it directly impacts the model's learning efficiency. To optimize the arrangement of the flattened 1D sequence, we introduce the Slice-orientated Mamba (SoMamba) and Local-orientated Mamba (LoMamba) module illustrated in part (a) and (b) of Fig. 2.

$$\text{SoMamba}(x) = \text{Mamba}(x_f) + \text{Mamba}(x_r), \tag{2}$$

$$\text{LoMamba}(x) = \text{Mamba}(x_{lf}) + \text{Mamba}(x_{lr}) + \text{Mamba}(x_{ls}), \tag{3}$$

where Mamba denotes the Mamba layer to model the global information within the sequence. f and r in Eq. 2 denote forward and reverse direction respectively. lf, lr, and ls in Eq. 3 refer to local-window forward direction, local-window reverse direction, and local-window across slices, correspondingly.

For SoMamba, we design a traversing path that spans the entire slice in both forward and reverse directions to compute feature dependencies, aiming to effectively capture the global information inherent in high-dimensional features. Additionally, we observe the presence of short-distance dependencies within the overall medical image, particularly as lesion areas may constitute a smaller proportion of the total. To address this, we design a Local modeling strategy known as LoMamba to focus on the adjacent pixels with the same semantic region, which empower the model to better aggregate the features because of more compact physical position of key information after flattening into a 1D sequence.

Figure 3 showcases the distinction between normal scanning and local scanning modes using the example of the local flattening sequence with a window size of three. In practice, the size of the local window is determined dynamically by the proportion and location of key information areas within the overall medical image. To be specific, the sequence through the standard flattening method exhibits a considerable distance between neighboring pixels that contain important segmentation information. In contrast, the continuous key segmentation pixels flattened by local windows of size three manifesting a strong distance coherence to a certain extent.

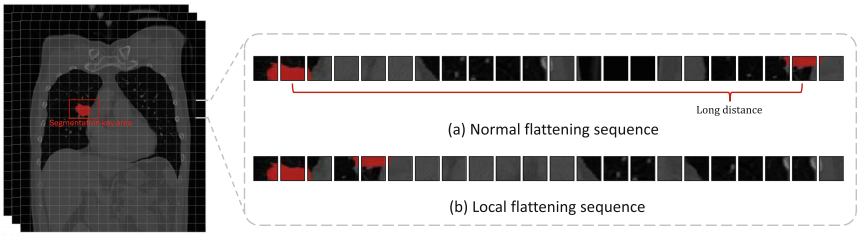


Fig. 3. Take 3×3 local window size around segmentation key area on one slice as an example. Sequence (a) showcases the normal flatten method, which traverse the whole image patches from beginning to end. Sequence (b) demonstrates the local flatten method.

2.2 FFT Gated Mechanism (FGM)

Fast Fourier Transform (FFT) [2] is leveraged to calculate the frequency values from spatial domain. In cases of CT images with poor contrast and high noise, as well as MRI images affected by artifacts, frequency information from high level and low level can correspondingly provide some boundary and shape cues to the model [30], which enables it to capture a wider range of feature representations, thus increasing the robustness. Furthermore, as highlighted in [29], the deeper layers of deep neural network tend to retain more low-frequency information. To capitalize on this, we design the FFT Gated Mechanism (GSM) to integrate the frequency and spatial features to enhance the performance of Mamba model. As illustrated in subgraph (c) of Fig. 2, the input 3D features are first fed into a convolution block, which contains a convolution, a normalization, and an activation layer. Then, the feature are transformed into the Fourier domain according to different layers, being extracted high and low frequency leveraging learnable thresholds within a filter. Subsequently, conducting inverse FFT to transform the required frequency feature back. Finally, a convolution block is used to further fuse the frequency features and spatial features after gated multiplication, while a residual connection is utilized to reuse the input features.

$$x_{\text{fre}} = \text{IFFT}(\,\text{Filter}(\,\text{FFT}\,(x))), \quad x_{\text{out}} = x_s * \text{gate} + x_{\text{fre}} * (1 - \text{gate}), \tag{4}$$

$$\text{Filter} = \begin{cases} x * (|x| < f_{\text{low}}), X \in \text{low-level frequency}, \\ x * (|x| > f_{\text{high}}), X \in \text{high-level frequency}, \end{cases} \tag{5}$$

where $\text{gate} = \text{Conv}_{3\times3\times3}(x_{\text{fft}}, x_s)$ is used to get the prior embedding to realize the coordination of features in spatial domain and frequency domain. f_{low}, f_{high} denote the learnable thresholds of the filter with the initial values set to 0.1 and 0.9 respectively.

3 Experiments

3.1 Datasets and Implementation Details

BraTS2023 Dataset [1,11,16]**.** This dataset comprises 1,251 3D MRI volumes of the brain. Each volume is presented in four different imaging modalities: T1, T1Gd, T2, and T2-FLAIR. Three segmentation targets are identified within each volume: Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC).

Lung Cancer (LC) Dataset. We gathered a collection of 828 3D chest CT scans from multiple centers, focusing on cases of central type lung carcinoma with small lesions. These cases pose inherent challenges in detection and analysis, as they appear as subtle anomalies within the imaging data. Each volume contains a single segmentation target, with the dominant lesion accurately annotated for each case. The visualization of this dataset is depicted in Fig. 4.

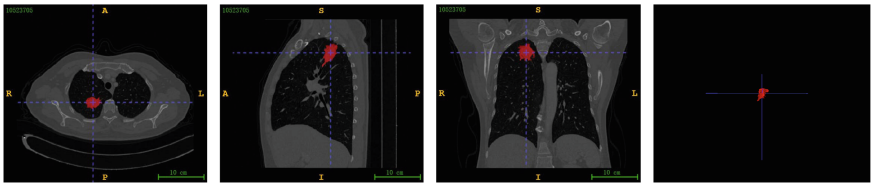


Fig. 4. The data visualization for LC dataset, which highlights the challenges posed by small lesions, which can be difficult to detect and analyze.

Evaluation and Metrics. In line with established approaches [27], we employ the Dice score (Dice) and Hausdorff Distance (HD95) to quantitatively assess our network's performance and compare it with state-of-the-art (SOTA) methods.

Implementation Details. Our model is developed using PyTorch 2.0.1 with CUDA 11.8 and Monai 1.3.0. During training, we apply a random crop size of 128 × 128 × 128 and utilize a batch size of 2 per GPU for each dataset. We employ cross-entropy loss for all experiments, using an SGD optimizer with a polynomial learning rate scheduler (initial learning rate set at 1e−4 and a decay rate of 3e−5). Each dataset undergoes 1000 training epochs, and we incorporate the following data augmentations: additive brightness, gamma correction, rotation, scaling, mirroring, and elastic deformation. All experiments are conducted on a cloud computing platform equipped with four NVIDIA GeForce RTX 4090 GPUs. For each dataset, we randomly assign 70% of the 3D volumes for training, 10% for validation, and the remaining 20% for testing.

3.2 Experimental Results

Compared Methods. We evaluate our network by comparing it to six state-of-the-art (SOTA) 3D image segmentation methods, which includes three CNN-based approaches (SegResNet [17], UX-Net [12], MedNeXt [18]), two

Table 1. Quantitative comparison on BraTS2023 dataset. The bold value denotes the best performance.

Methods	BraTS2023								LC	
	WT		TC		ET		Avg		Lung Cancer	
	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓	Dice↑	HD95↓
SegresNet↑ [17]	92.02	4.07	89.10	4.08	83.66	3.88	88.26	4.01	71.56	55.03
UX-Net↑ [12]	93.13	4.56	90.03	5.68	85.91	4.19	89.69	4.81	72.63	59.06
MedNeXt↑ [18]	92.41	4.98	87.75	4.67	83.96	4.51	88.04	4.72	57.76	111.47
UNETR↑ [8]	92.19	6.17	86.39	5.29	84.48	5.03	87.68	5.49	65.12	101.7
SwinUNETR↑ [7]	92.71	5.22	87.79	4.42	84.21	4.48	88.23	4.70	65.83	89.27
SegMamba↑ [27]	93.61	**3.37**	92.65	3.85	87.71	3.48	91.32	3.56	72.36	60.33
Ours	**94.10**	3.78	**92.84**	**3.30**	**88.83**	**3.35**	**91.92**	**3.48**	**75.34**	**44.52**

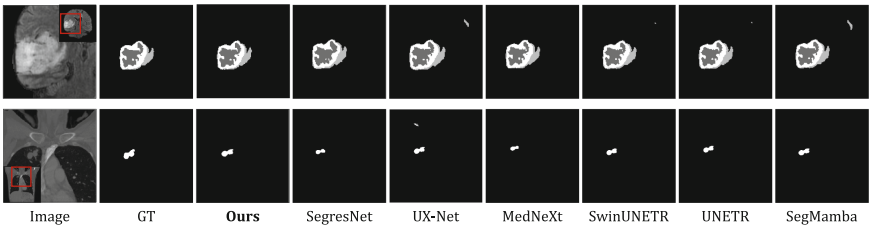


Fig. 5. Visual comparisons of proposed HybridMamba and state-of-the-art methods.

transformer-based methods (UNETR [8] and SwinUNETR [7]), and one Mamba-based technique (SegMamba [27]).

Quantitative Comparisons. Table 1 summarizes the Dice score and HD95 for each modalities on BraTS2023 and LC dataset and the total average scores. For BraTS2023 dataset, SegMamba, the Mamba-based method, achieves the best performance among the comparison methods, with an average Dice of 91.32% and an average HD95 of 3.56. In comparison, our HybridMamba achieves the highest Dice of 94.10%, 92.84%, and 88.83% on WT, TC, and ET, respectively, and the best HD95 with 3.30 on TC and 3.35 on ET except for 3.78 on WT. All in all, the total average scores shows the best segmentation robustness. In addition, our HybridMamba outperforms the SOTA method exceeding 2.98% for Dice and 15.81 for HD95 on LC dataset, getting 75.34% and 44.52% respectively. This proves the most effectiveness of HybridMamba compared to other approaches.

Visual Comparisons. We choose six comparative methods for visual assessment on two datasets to evaluate image segmentation performance. As shown in Fig. 5, our HybridMamba effectively delineates the boundary of each tumor region in the BraTS2023 dataset. Similarly, our approach successfully identifies cancerous areas in the LC dataset. The segmentation results demonstrate enhanced consistency compared to other state-of-the-art techniques.

Ablation Studies. Table 2 confirms the effectiveness of both S-LMamba (M1) and FGM (M2) modules on the LC dataset. S-LMamba improves SegMamba's Dice by 1.47% to 73.83% and reduces HD95 by 9.45 to 69.78. FGM further boosts Dice to 74.98% and reduces HD95 by 18.35 to 51.43. HybridMamba, combining both modules, achieves optimal results with 75.34% Dice and 44.52 HD95.

Table 2. Ablation study for different modules on LC dataset.

Methods	Modules		Dice↑	HD95↓
	S-LMamba	FGM		
SegMamba			72.36	69.78
M1	✓		73.83	60.33
M2		✓	74.98	51.43
Ours	✓	✓	**75.34**	**44.52**

4 Conclusion

In this work, we have developed HybridMamba to enhance the 3D biomedical segmentation task. Specifically, our network makes two primary contributions. First, we devise the S-LMamba block to effectively balance the modeling of global and local dependencies. Second, we aggregate frequency features with spatial features to enhance the representation of the model. Experimental results on two datasets demonstrate that our framework clearly outperforms SOTA methods in terms of 3D medical image segmentation task.

Acknowledgments. This research is supported by the Guangdong Science and Technology Department (No. 2024ZDZX2004), the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things(No.2023B1212010007), the National Natural Science Foundation of China (No. 82001903), the Natural Science Foundation of Shanghai (No. 25ZR1402077 and 21ZR1414200), and the Excellent Young Talents Project of Shanghai Public Health Three-year (2023–2025) Action Plan (No. GWVI-11.2-YQ48).

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**(1), 1–13 (2017)
2. Bergland, G.D.: A guided tour of the fast Fourier transform. IEEE Spectr. **6**(7), 41–52 (1969)
3. Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Gu, A., Dao, T., Ermon, S., Rudra, A., Ré, C.: HiPPO: recurrent memory with optimal polynomial projections. Adv. Neural. Inf. Process. Syst. **33**, 1474–1487 (2020)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) BrainLes 2021. LNCS, vol. 12962, pp. 272–284. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-08999-2_22
8. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
9. Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C.: LocalMamba: visual state space model with windowed selective scan. arXiv preprint arXiv:2403.09338 (2024)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)

11. Kazerooni, A.F., et al.: The brain tumor segmentation (BraTS) challenge 2023: focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). ArXiv pp. arXiv–2305 (2024)
12. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3D UX-net: a large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076 (2022)
13. Liu, J., et al.: Swin-UMamba: mamba-based UNet with ImageNet-based pretraining. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15009, pp. 615–625. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_59
14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
15. Ma, J., Li, F., Wang, B.: U-mamba: enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
16. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
17. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_28
18. Roy, S., et al.: MedNeXt: transformer-driven scaling of convnets for medical image segmentation. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14223, pp. 405–415. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_39
19. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
20. Wang, H., et al.: Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. IEEE Trans. Med. Imaging (2024)
21. Wang, H., et al.: Serp-mamba: advancing high-resolution retinal vessel segmentation with selective state-space model. arXiv preprint arXiv:2409.04356 (2024)
22. Wang, H., et al.: Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. IEEE Trans. Med. Imaging (2024)
23. Wu, H., et al.: Semi-supervised video desnowing network via temporal decoupling experts and distribution-driven contrastive regularization. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV 2024. LNCS, vol. 15068, pp. 70–89. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72684-2_5
24. Wu, H., Yang, Y., Xu, H., Wang, W., Zhou, J., Zhu, L.: RainMamba: enhanced locality learning with state space models for video deraining. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 7881–7890 (2024)
25. Xing, Z., et al.: Diff-UNet: a diffusion embedded network for robust 3D medical image segmentation. Med. Image Anal. 103654 (2025)
26. Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L.: Diff-UNet: a diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326 (2023)
27. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: SegMamba: long-range sequential modeling mamba for 3D medical image segmentation. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15008, pp. 578–588. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72111-3_54
28. Xing, Z., Zhu, L., Yu, L., Xing, Z., Wan, L.: Hybrid masked image modeling for 3D medical image segmentation. IEEE J. Biomed. Health Inform. (2024)

29. Xu, Z.Q.J., Zhang, Y., Luo, T., Xiao, Y., Ma, Z.: Frequency principle: Fourier analysis sheds light on deep neural networks. arXiv preprint arXiv:1901.06523 (2019)
30. Zhou, Y., Huang, J., Wang, C., Song, L., Yang, G.: XNet: wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21085–21096 (2023)



IFRFNet: Iterative Frequency Restoration-Fusion Network for Fast System Matrix Calibration on Magnetic Particle Image

Weixin Xu[1,2,3], Penghua Zhai[1,2,3], Jie Tian[1,2(✉)], and Wei Mu[1,2,3(✉)]

[1] School of Engineering Medicine, Beihang University, Beijing, China
jie.tian@ia.ac.cn, weimu@buaa.edu.cn
[2] Key Laboratory of Big Data-Based Precision Medicine (Beihang University), Ministry of Industry and Information Technology of the People's Republic of China, Beijing 100191, China
[3] Qingdao Research Institute, Beihang University, Qingdao 266100, China

Abstract. Magnetic Particle Imaging (MPI), an emerging technique with high sensitivity and resolution, requires time-consuming calibration for System Matrix (SM)-based reconstruction. Due to the strong locality and redundancy in the frequency domain, sparse sampling can capture sufficient information for rapid SM calibration without full-size SMs. However, it often leads to low-frequency energy leakage due to nonlinear magnetization of nanoparticles, causing the loss of low-frequency components. These components are essential for maintaining the SM's shape, and their absence leads to structural degradation and visible artifacts. Current methods tend to overemphasize high-frequency features, neglecting these low-frequency ones. Besides, single-step upsampling leads to error accumulation, especially with large scaling ratios, degrading reconstruction quality. To address these issues, we propose the Iterative Frequency Restoration-Fusion Network (IFRFNet), which uses an iterative frequency-domain restoration-fusion module. Unlike single-step upsampling, our approach refines, fuses, and upsamples high- and low-frequency features in stages, ensuring continuous optimization. This prevents error accumulation, preserves fine details, and maintains structural integrity. By iteratively recovering low-frequency components and refining high-frequency details, IFRFNet minimizes artifacts and retains crucial information. The Effective Upsampler further enhances the quality of the features, ensuring clear and realistic final SM volumes. Experiments on the OpenMPI dataset show that IFRFNet achieves SOTA performance.

Keywords: MPI · System Matrix · Fast Calibration

This study was supported by the National Natural Science Foundation of China (62176013, 92259205, 62027901), Beijing Natural Science Foundation (L222021), National Key Research and Development Program of China (2022YFC2505100), Shandong Provincial Natural Science Foundation (ZR2024QF010), and the Science, Technology and Innovation Project of Xiongan New Area (2023XAGG0071).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 290–300, 2026.
https://doi.org/10.1007/978-3-032-04947-6_28

1 Introduction

Magnetic Particle Imaging (MPI) is a tomographic imaging technique capable of detecting the magnetism of iron oxide nanoparticles injected into the bloodstream to produce three-dimensional images [9,22]. In recent years, MPI has garnered significant attention for its ability to enhance imaging resolution without the use of radioactive substances. This technology has been extensively explored in various fields, including vascular imaging, cell tracking, tumor imaging, and has achieved remarkable progress in clinical translation. Traditionally, MPI reconstruction relies on two primary methods: X-space-based methods and system matrix (SM)-based methods [4]. Among these, the SM-based approach offers superior image quality but requires a highly complex and time-intensive calibration process. This process involves measuring the voxel-level response of the MPI scanner within the field of view (FOV) to generate a complex-valued SM. For instance, obtaining a system matrix of size $37 \times 37 \times 37$ can take approximately 32 h, whereas a reduced size of $9 \times 9 \times 9$ requires only about 37 min. Sparse sampling has been proposed to accelerate the calibration process, as it can capture enough information for rapid SM calibration without the need for full-size SMs. Despite its advantages, sparse sampling often leads to low-frequency energy leakage due to the nonlinear magnetization of nanoparticles, which enables imaging but concentrates signal energy in high frequencies, leading low-frequency components relatively weak and easily suppressed [19]. Also, sparse sampling cannot fully cover all spatial points [14]. These low-frequency components are essential for preserving the overall shape and structure of the SM. Their absence can cause structural degradation and visible artifacts, compromising the quality of the reconstructed images. Recent approaches have addressed this issue by treating SM calibration as compressed sensing (CS) [10,11,14] and super-resolution (SR) methods [2,12,18] to reconstruct high-resolution images from low-resolution SMs acquired through rapid sparse calibration, accelerating the calibration process. With the rapid development of deep learning, many methods have been proposed for SR tasks. The SRCNN [5] was the pioneering deep learning model applied to image super-resolution (SR) using an end-to-end convolutional neural network (CNN). Since then, various CNN-based approaches [1,6,21] have been developed to address the SR problem. LKFN [3] enhances feature extraction by introducing larger kernel convolutions, thus providing a wider receptive field. SMFANet [20] proposes the SMFA module, which efficiently captures both non-local and local features to extract more representative information. Additionally, some Vision Transformer (ViT) [7] based methods are proposed to complete variational tasks [8,17]. TTVSR [17] presents a framework that incorporates both temporal consistency and spatial attention, significantly improving video super-resolution by capturing temporal and spatial dependencies efficiently. TranSMS [8] introduced a multi-scale transformer-based architecture that captures fine-grained details across multiple scales, enhancing the model's ability to perform 2D SM calibration in MPI.

However, these methods often prioritize high-frequency features at the expense of low-frequency components, leading to the neglect of essential structural information. Furthermore, single-step upsampling methods suffer from error accumulation, especially when scaling ratios are large (e.g., 4×), which results in the loss of fine details, increased artifacts, and difficulties in feature refinement. Finally, although iterative upsampling has proven that can effectively alleviate this problem [15], current methods often ignore further processing of the features after each upsampling step, directly using coarse features. If the features after each upsampling step are not fully optimized, the network may struggle to extract critical information, which in turn affects the clarity and authenticity of the final image. Additionally, errors from earlier stages can accumulate and be magnified in subsequent stages, ultimately degrading the overall super-resolution performance. To address the aforementioned challenges, we propose a novel fast SM calibration framework, named the Iterative Frequency Restoration-Fusion Network (IFRFNet). This framework employs an iterative frequency domain feature refine module that filters both high and low-frequency domain features while simultaneously repairing and fusing them. Subsequently, we utilize the Effective Upsampler to upsample and optimize the features, ensuring that the final calibrated SM volumes retain both clarity and reality, effectively restoring the overall shape.

Our main contributions are summarized as follows:

- We propose a novel attention-based frequency refinement module, termed the Frequency Attention Module, designed to capture both global and local features effectively. This module assigns varying weights to the extracted frequency features, ranging from 0 to 1, reflecting their relative importance.
- We propose a novel frequency domain feature processing module, consisting of the Frequency Filter and Frequency Refine-Fusion Module, to filter, refine, and fusion the features with different frequencies to help better restore the information gap between low-resolution SMs and high-resolution SMs.
- To better generate the high-resolution SMs from the low-resolution ones, we propose an Effective Upsampler module, which could effectively upsample restored frequency fusion features directly.
- Based on the Frequency Filter, Frequency Refine-Fusion Module, and Effective Upsampler, we propose a novel framework, for fast SM calibration on MPI, dubbed IFRFNet, which achieves state-of-the-art (SOTA) performance on the OpenMPI dataset, the widely used public benchmark with extensive resources for MPI research and development.

2 Methods

2.1 Overall Architecture

The overall architecture of our proposed IFRFNet is shown in Fig. 1. Specifically, our proposed IFRFNet consists of the following aspects: 1). The encoder, which is used to extract the shallow feature from low-resolution SM input, consists

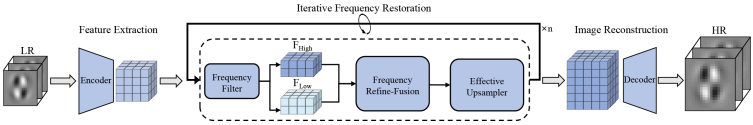


Fig. 1. The framework of our proposed IFRFNet.

of two convolution blocks. 2). The iterative frequency restoration blocks. These blocks are imported to filter high and low frequencies from the previous encoder layer first, then refine those frequencies and fuse the more precise frequencies. Finally, an effective upsampler was proposed to upsample the refined frequencies to the needed sizes. The number of restoration blocks equals $\log_2 U$, in which U represents the upsampling ratio. 3). The decoder block was built by a single convolution block, to reconstruct features to the SMs with high resolution.

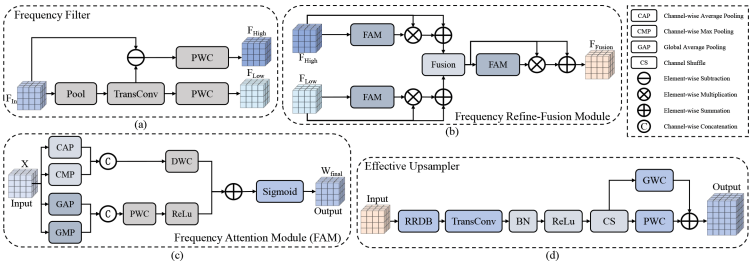


Fig. 2. Details about our proposed modules.

2.2 Frequency Filter

Following the feature extraction block, the Frequency Filter (FF) is proposed to split features with high frequency and low frequency, details of the FF are shown in Fig. 2(a). Specifically, since the low frequency represents the overall shape of the inputs, we proposed to utilize the average pooling module first to extract the global shape information. Then, a transposed convolution is incorporated to recover the spatial resolution of the features. Finally, a point-wise convolution (PWC) is used to get the filtered low-frequency features, F_{Low}. Details of this process are shown in Eq. 1.

$$F_{Low} = PWC(TransConv(Pool(F_{In}))) \tag{1}$$

Moreover, as shown in the Eq. 2, for the high-frequency features that focus on detailed texture information, we use element subtraction to subtract the low-frequency features before PWC from input features, which can effectively remove the overall shape information and separate the high-frequency features.

Afterward, similar to the low-frequency filter, we used a PWC operation to get the final high-frequency features, dubbed F_{High}.

$$F_{High} = PWC(F_{In} - TransConv(Pool(F_{In}))) \tag{2}$$

2.3 Frequency Refine-Fusion Module

To refine and fusion the filtered frequency-domain features, we propose the Frequency Refine-Fusion Module (FRFM), details are shown in Fig. 2(b). This module consists of two main blocks: Frequency Attention Module (FAM), details are shown in Fig. 2(c), and Fusion block. Specifically, for the input frequency-domain features, F_{High} and F_{Low}, a FAM was first used to calculate the weights for each feature. After the weights are calculated, it is multiplied with the input frequency feature to obtain a feature map with different importances, and the skip connection is used to add it to the original feature map to obtain the final refined frequency domain feature map that retains the context information. We adopt the FAM on F_{High}, F_{Low} and the fused coarse features, therefore get F^f_{High}, F^f_{Low}, F^f_{Fusion}.

FAM: To effectively capture both global and local information across channel and spatial dimensions, we propose a two-step attention mechanism on input frequency-domain feature X. First, we introduce channel-wise attention to compute attention weights along the channel dimension. This involves two operations: channel average pooling (CAP) and channel max pooling (CMP). Specifically, the resulting channel-wise vectors, $W_{ca} \in \mathbb{R}^{C\times1\times1\times1}$ and $W_{cm} \in \mathbb{R}^{C\times1\times1\times1}$, represent global and local features, respectively. These two vectors are then concatenated to obtain the combined channel-wise attention weights. Finally, we apply a depth-wise convolution (DWC) to generate the final channel attention output, denoted W_c, details of this process are shown in Eq. 3. Moreover, for the spatial dimensions, we utilize global average pooling (GAP) and global max pooling (GMP) to compute spatial attention vectors, $W_{sa} \in \mathbb{R}^{1\times H\times W\times D}$ and $W_{sm} \in \mathbb{R}^{1\times H\times W\times D}$, which capture global and local spatial features, respectively. These features are concatenated, and a PWC with a ReLU activation is applied to generate the final spatial attention output, W_s, details of this process are shown in Eq. 4.

$$W_c = DWC(Concat(CAP(x), CMP(x))) \tag{3}$$

$$W_s = ReLu(PWC(Concat(GAP(x), GMP(x)))) \tag{4}$$

After obtaining the channel attention weights W_c and spatial attention weights W_s, we combine them through element-wise addition $W = W_c + W_s$, followed by a sigmoid activation $W_{final} = \sigma(W)$ to normalize the values between 0 and 1. This operation ensures a balanced contribution from both channel and spatial attention, improves feature selection by focusing on relevant features, and stabilizes the learning process through normalization.

Fusion Block: To achieve effective feature integration from high-frequency and low-frequency representations, we utilize a simple yet effective fusion block. This block enables the integration of high-frequency and low-frequency features and preserves complementary information from both inputs by splitting the features and performing separate additions. Specifically, given two inputs $F^f_{\text{High}} \in \mathbb{R}^{C\times H\times W\times D}$ and $F^f_{\text{Low}} \in \mathbb{R}^{C\times H\times W\times D}$, we first split each input along the channel dimension into two equal parts:

$$F^{f,1}_{\text{High}}, F^{f,2}_{\text{High}} = F^f_{\text{High}}[:C/2,:,:,:], \quad F^{f,1}_{\text{Low}}, F^{f,2}_{\text{Low}} = F^f_{\text{Low}}[:C/2,:,:,:] \tag{5}$$

Next, we perform element-wise addition on the corresponding parts:

$$F^{f,1} = F^{f,1}_{\text{High}} + F^{f,1}_{\text{Low}}, \quad F^{f,2} = F^{f,2}_{\text{High}} + F^{f,2}_{\text{Low}} \tag{6}$$

Finally, the fused output F^f_{out} is obtained by concatenating $F^{f,1}$ and $F^{f,2}$ along the channel dimension:

$$F^f_{\text{out}} = \text{Concat}(F^{f,1}, F^{f,2}), \quad F^f_{\text{out}} \in \mathbb{R}^{C\times H\times W\times D} \tag{7}$$

2.4 Effective Upsampler

The proposed Effective Upsampler, as depicted in Fig. 2(d), is a carefully designed module aimed at enhancing the resolution of input feature maps while maintaining high-quality feature representation. The process begins with an input feature map passing through a Residual-in-Residual Dense Block (RRDB) [23], which effectively captures both local and global context features. Following this, a transposed convolution layer (TransConv) performs upsampling to increase the spatial resolution. Batch normalization (BN) and ReLU activation are applied to stabilize the training process and introduce non-linearity, respectively. Subsequently, a Channel Shuffle (CS) operation reorganizes feature channels to enhance feature interaction, followed by two parallel pathways. The first pathway applies a Group-Wise Convolution (GWC) layer to assign channel-specific features adaptively, while the second pathway utilizes a PWC to refine spatial details. Finally, the outputs of the two pathways are element-wise summed, resulting in the final upsampled feature map.

3 Experiments

3.1 Dataset

The OpenMPI dataset [13], the first open-source MPI dataset, has become a widely used public benchmark with extensive resources to support MPI research and development. This dataset includes system matrix (SM) calibration data and phantom measurements derived from various magnetic nanoparticle (MNP) formulations. Following standard practices [2], we specifically utilized Experiment #7 of the SM calibration, which features SynomagD MNPs (Micromod GmbH, Germany), to construct our training dataset. For testing, we used Experiment

#6, which includes Perimag MNP (Micromod GmbH, Germany), to assess the generalizability of the model across different types of MNP. To ensure the reliability and robustness of our experimental results, we applied stringent filtering criteria, retaining only SMs with a signal-to-noise ratio (SNR) greater than 3. This preprocessing step produced a training set of 4,129 volumes and a test set of 3,290 volumes.

3.2 Implementation Details

Our experimental setup utilized PyTorch version 1.13.1 for all model training. The training process was executed on a single NVIDIA RTX 4090 GPU equipped with 24 GB of memory. The models were optimized using the Adam optimizer, initialized with a learning rate of 2.5×10^{-4}, a decay factor of 0.2, and a weight decay of 5×10^{-4}. Following the methodology outlined in [8], we adopted the L1 loss function, mathematically expressed as:

$$L_1 = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{8}$$

where N denotes the total number of samples, y_i represents the ground truth values, and $\hat{y}_i$ refers to the predicted outputs.

3.3 Evaluation Metrics

We utilized three widely recognized metrics to quantitatively evaluate the performance of various state-of-the-art (SOTA) models. Following the approach adopted in prior studies [2,8], the normalized Root Mean Square Error (nRMSE) was employed as the sole metric for SM calibration assessment. For image reconstruction evaluation, we used nRMSE alongside Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM).

3.4 Results

We evaluated our approach against seven SOTA models: Bicubic interpolation, Trilinear interpolation, SRCNN, 3dSMRnet, VolumeNet, SMFANet, and LKFN. The evaluation followed a two-step process. In the first step, we measured the accuracy of the reconstructed SMs by comparing them to the ground truth. In the second step, we utilized the calibrated SMs to reconstruct the measurements of the shape and resolution phantoms in OpenMPI dataset. For consistency, the same standard regularization parameter was applied ($\lambda = 0.75$) across three iterations.

SM Calibration Results. The quantitative results of the SM calibration are summarized in Table 1. Notably, with a scaling ratio equals 2, the proposed IFRFNet demonstrates a remarkable improvement by reducing the normalized

Table 1. SM calibration results on OpenMPI dataset, with the best performing values in bold and the second best performing values underlined.

Ratio	2×	4×
Method	nRMSE↓ (%)	nRMSE↓ (%)
Bicubic	6.76	8.76
Trilinear	6.21	8.55
SRCNN [5]	3.88	5.70
3dSMRnet [2]	<u>3.62</u>	<u>4.86</u>
VolumeNet [16]	4.13	6.12
SMFANet [20]	4.04	5.81
LKFN [3]	3.98	6.02
Ours	**3.47**	**4.62**

Table 2. Image reconstruction results of Resolution phantom on calibrated SM on OpenMPI dataset, with the best performing values in bold and the second best performing values underlined.

Ratio	2×			4×		
Method	nRMSE(↓)	PSNR(↑)	SSIM(↑)	nRMSE(↓)	PSNR(↑)	SSIM(↑)
Bicubic	1.93%	34.31	0.8960	8.49%	21.42	0.2181
Trilinear	1.86%	34.60	0.8634	9.64%	20.31	0.1474
SRCNN	0.78%	42.17	0.9553	5.81%	24.71	0.2374
3dSMRnet	0.75%	42.48	0.9569	<u>2.64%</u>	<u>31.58</u>	<u>0.6201</u>
VolumeNet	0.77%	42.28	0.9556	6.54%	23.69	0.233
SMFANet	0.77%	42.24	0.9559	5.74%	24.81	0.2482
LKFN	<u>0.73%</u>	<u>42.71</u>	<u>0.9585</u>	5.66%	24.94	0.2241
Ours	**0.72%**	**42.82**	**0.9589**	**2.24%**	**33.00**	**0.7004**

Root Mean Square Error (nRMSE) by 0.15% compared to prior state-of-the-art (SOTA) approaches. This notable enhancement underscores the accuracy and robustness of IFRFNet in addressing moderate scaling ratios. Furthermore, when the scaling ratio increased to 4, IFRFNet consistently maintains its superior performance, achieving an additional 0.24% reduction in nRMSE. These results highlight the versatility and adaptability of IFRFNet across varying scaling ratios, reaffirming its advantage over existing methodologies.

Image Reconstruction Results. We evaluated the image reconstruction performance using a super-resolution calibrated SM. For image reconstruction, following [2], we selected the shape phantom and resolution phantom from the OpenMPI dataset. Quantitative results of these two phantoms are shown in Table 2 and 3. Similar to results in Table 1, results of image reconstruction

Table 3. Image reconstruction results of Shape phantom on calibrated SM on OpenMPI dataset, with the best performing values in bold and the second best performing values underlined.

Ratio	2×			4×		
Method	nRMSE(↓)	PSNR(↑)	SSIM(↑)	nRMSE(↓)	PSNR(↑)	SSIM(↑)
Bicubic	3.19%	29.93	0.8106	8.45%	21.46	0.3902
Trilinear	3.12%	30.12	0.7968	7.03%	23.06	0.3265
SRCNN	1.51%	36.43	0.8716	4.61%	26.72	0.4813
3dSMRnet	1.57%	36.11	0.8589	<u>2.90%</u>	<u>30.75</u>	<u>0.6815</u>
VolumeNet	1.50%	36.46	0.8742	5.48%	25.22	0.4241
SMFANet	<u>1.48%</u>	<u>36.58</u>	<u>0.8768</u>	5.48%	25.23	0.4367
LKFN	1.49%	36.56	0.8761	5.56%	25.10	0.4385
Ours	**1.45%**	**36.78**	**0.8781**	**2.58%**	**31.78**	**0.7014**

and SM calibration are almost consistent. Specifically, for the resolution phantom, compared with the previous SOTA methods, performance of our proposed IFRFNet can achieve (0.01%, 0.11, 0.0004) improvement on (nRMSE, PSNR, SSIM) metrics when the scaling ratio equals 2 and (0.40%, 1.42, 0.0803) improvement on (nRMSE, PSNR, SSIM) metrics for ratio 4. Moreover, for the shape phantom, compared with the previous SOTA methods, performance of our proposed IFRFNet can achieve (0.04%, 0.21, 0.0013) improvement on (nRMSE, PSNR, SSIM) metrics when the scaling ratio equals 2 and (0.32%, 1.03, 0.0199) improvement on (nRMSE, PSNR, SSIM) metrics for ratio 4.

4 Conclusion

In this paper, we present a novel and efficient SM calibration framework for MPI, termed IFRFNet. The proposed model employs the iterative strategy to enhance the restoration of shape and resolution information in SM by effectively filtering high- and low-frequency features and optimizing their fusion through an attention-based mechanism. To achieve this, we first introduce a Frequency Filter Module, which isolates low- and high-frequency features from the original inputs. Next, we propose the Frequency Refine-Fusion Module, the core of which includes the Frequency Attention Module (FAM) and the Fusion Block. FAM help to refine the high- and low-frequency features extracted in the previous stage, while a dedicated Fusion Block integrates the optimized features more effectively. Finally, an Effective Upsampler is designed to further refine features during the upsampling process, mitigating issues such as uneven feature distribution and cumulative error amplification. Experimental results on the OpenMPI dataset demonstrate that IFRFNet surpasses existing methods, avoiding loss of low-frequency information during upsampling. This work establishes a strong foundation for future advancements and practical applications in MPI.

Disclosure of Interests. We have no conflicts of interest to disclose.

References

1. Ahn, N., Kang, B., Sohn, K.A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 252–268 (2018)
2. Baltruschat, I.M., Szwargulski, P., Griese, F., Grosser, M., Werner, R., Knopp, T.: 3d-SMRnet: achieving a new quality of MPI system matrix recovery by deep learning. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 74–82. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_8
3. Chen, J., Duanmu, C., Long, H.: Large kernel frequency-enhanced network for efficient single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6317–6326 (2024)
4. Chen, X., Jiang, Z., Han, X., Wang, X., Tang, X.: Research of magnetic particle imaging reconstruction based on the elastic net regularization. Biomed. Signal Process. Control **69**, 102823 (2021)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)
6. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part II. LNCS, vol. 9906, pp. 391–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_25
7. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Güngör, A., Askin, B., Soydan, D.A., Saritas, E.U., Top, C.B., Çukur, T.: TranSMS: transformers for super-resolution calibration in magnetic particle imaging. IEEE Trans. Med. Imaging **41**(12), 3562–3574 (2022)
9. Han, X., et al.: The applications of magnetic particle imaging: from cell to body. Diagnostics **10**(10), 800 (2020)
10. Ilbey, S., Top, C.B., Güngör, A., Çukur, T., Saritas, E.U., Güven, H.E.: Fast system calibration with coded calibration scenes for magnetic particle imaging. IEEE Trans. Med. Imaging **38**(9), 2070–2080 (2019)
11. Kaethner, C., Erb, W., Ahlborg, M., Szwargulski, P., Knopp, T., Buzug, T.M.: Non-equispaced system matrix acquisition for magnetic particle imaging based on lissajous node points. IEEE Trans. Med. Imaging **35**(11), 2476–2485 (2016)
12. Kluth, T., Bathke, C., Jiang, M., Maass, P.: Joint super-resolution image reconstruction and parameter identification in imaging operator: analysis of bilinear operator equations, numerical solution, and application to magnetic particle imaging. Inverse Prob. **36**(12), 124006 (2020)
13. Knopp, T., Szwargulski, P., Griese, F., Gräser, M.: OpenMPIData: an initiative for freely accessible magnetic particle imaging data. Data Brief **28**, 104971 (2020)
14. Knopp, T., Weber, A.: Sparse reconstruction of the magnetic particle imaging system matrix. IEEE Trans. Med. Imaging **32**(8), 1473–1480 (2013)
15. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 624–632 (2017)
16. Li, Y., Iwamoto, Y., Lin, L., Xu, R., Tong, R., Chen, Y.W.: VolumeNet: a lightweight parallel network for super-resolution of MR and CT volumetric data. IEEE Trans. Image Process. **30**, 4840–4854 (2021)

17. Liu, C., Yang, H., Fu, J., Qian, X.: Learning trajectory-aware transformer for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5687–5696 (2022)
18. Omer, O.A., Wojtczyk, H., Buzug, T.M.: Simultaneous reconstruction and resolution enhancement for magnetic particle imaging. IEEE Trans. Magn. **51**(2), 1–4 (2015)
19. Panagiotopoulos, N., et al.: Magnetic particle imaging: current developments and future directions. Int. J. Nanomed. 3097–3114 (2015)
20. Ren, B., et al.: The ninth NTIRE 2024 efficient super-resolution challenge report. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6595–6631 (2024)
21. Sun, L., Dong, J., Tang, J., Pan, J.: Spatially-adaptive feature modulation for efficient image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13190–13199 (2023)
22. Sun, S., Chen, Y., Zhao, M., Xu, L., Zhong, J.: Image reconstruction for magnetic particle imaging based on sparse representation and deep learning. IEEE Trans. Instrum. Meas. **73** (2024)
23. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)



Instrument-Splatting: Controllable Photorealistic Reconstruction of Surgical Instruments Using Gaussian Splatting

Shuojue Yang[1], Zijian Wu[2], Mingxuan Hong[1], Qian Li[1], Daiyun Shen[1], Septimiu E. Salcudean[2], and Yueming Jin[1(✉)]

[1] National University of Singapore, Singapore, Singapore
ymjin@nus.edu.sg
[2] The University of British Columbia, Vancouver, Canada

Abstract. Real2Sim is becoming increasingly important with the rapid development of surgical artificial intelligence (AI) and autonomy. In this work, we propose a novel Real2Sim methodology, *Instrument-Splatting*, that leverages 3D Gaussian Splatting to provide fully controllable 3D reconstruction of surgical instruments from monocular surgical videos. To maintain both high visual fidelity and manipulability, we introduce a geometry pre-training to bind Gaussian point clouds on part mesh with accurate geometric priors and define a forward kinematics to control the Gaussians as flexible as real instruments. Afterward, to handle unposed videos, we design a novel instrument pose tracking method leveraging semantics-embedded Gaussians to robustly refine per-frame instrument poses and joint states in a render-and-compare manner, which allows our instrument Gaussian to accurately learn textures and reach photorealistic rendering. We validated our method on 2 publicly released surgical videos and 4 videos collected on *ex vivo* tissues and green screens. Quantitative and qualitative evaluations demonstrate the effectiveness and superiority of the proposed method. Our code is available at https://github.com/jinlab-imvr/Instrument-Splatting.

Keywords: Texture Learning · Surgical Instrument · 3D Gaussian Splatting · Controllable Gaussian Splatting

1 Introduction

In robot-assisted laparoscopic surgery, computer vision-based surgical instrument identification [2,6,7,22,27,29] is of fundamental significance in downstream applications such as augmented reality and autonomous robotic surgery. While deep learning-based methods demonstrated excellent performance in general computer vision domains, adapting these methods to surgical instrument analysis remains challenging due to limited data and annotations.

S. Yang and Z. Wu—Equal contribution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 301–311, 2026.
https://doi.org/10.1007/978-3-032-04947-6_29

For robotic surgery automation, existing public datasets either lack instrument pose labels or only provide sole wrist part pose labels [24]. This limits imitation learning-based methods where robots directly learn instrument motions from human expert surgical videos [10,19]. To address this issue, a feasible solution is Sim2Real transfer learning, which leverages large amounts of trajectories and videos generated from the simulation environment [17]. Conventionally, most studies [4] adopt CAD mesh models of instruments that exhibit visually unrealistic appearances. This significant sim-to-real gap impairs the performance of models trained on such data. Therefore, a highly realistic and controllable 3D asset of surgical instruments is critical.

With the recent progress in 3D reconstruction, new 3D representations such as Neural Radiance Field (NeRF) [16] and 3D Gaussian Splatting (GS) [9] emerged, which serve as powerful tools that can provide reconstructions with high visual fidelity. Driven by these advances, Zeng et al. [28] utilize 3D GS to reconstruct surgical instruments for novel data synthesis assisting surgical instrument detection. However, this method requires a customized system for data collection and cannot dynamically model surgical instruments with joint changes, which degrades the flexibility and controllability of the reconstructed 3D asset. Similarly, [20] produces instrument reconstruction using large reconstruction model, but lacks controllability. Although previous methods [12,13,21,25,26,30] for dynamic surgical scene reconstruction have demonstrated high visual fidelity on deformable tissue reconstruction, they cannot provide controllable 3D GS since the deformation fields are only conditioned on the timestamps in surgical videos. In addition, the reconstruction quality of these methods is degraded with large and complex instrument motions, which commonly occur in surgical scenarios. Meanwhile, there are emerging works that adopt GS to reconstruct the general robotic arm [14,15] or articulated human body [11], these approaches assume known (or well-estimated) poses and joint states of the articulated objects, which is unpractical in robotic-surgery settings where the kinematics of surgical instruments are unavailable/noisy [2].

Driven by these limitations, we propose *Instrument-Splatting*, a real-to-sim framework which leverages real monocular surgical videos to create a surgical instrument digital twin represented by controllable 3D GS that can be used for photorealistic simulation. Given untextured instrument CAD models, our pipeline starts with a geometry pretraining that effectively binds the GS to the mesh models with accurate geometry. Next, we propose a novel pose estimation and tracking strategy based on the pretrained GS to estimate the instrument pose and joint states. To handle large inter-frame instrument motions, we first introduce a correspondence matching module to guide the GS to move toward the current frame pose, followed by a structural-prior-based loose regularization to penalize singular pose estimations. Finally, we evaluate our framework on both intraoperative videos selected from EndoVis2017&2018 and in-house dataset collected using da Vinci robotic system. Experiments indicate the efficacy of our approach, demonstrating superior pose estimation accuracy and visual fidelity.

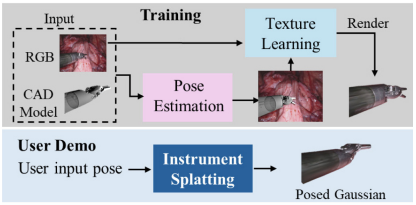


Fig. 1. Instrument-Splatting Pipeline.

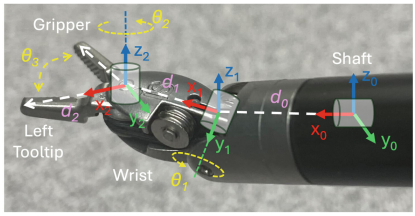


Fig. 2. Forward kinematics of the LND.

2 Method

Instrument-splatting aims to learn a Gaussian Splatting representation of articulated surgical instruments. The overview pipeline is shown in Fig. 1. During training, we first estimate the instrument pose and joint state given an RGB image and instrument CAD model. Then, images with corresponding poses are fed into the texture learning module to optimize the instrument GS. After training, users can input in-ranged instrument pose with joint variables and get the posed mesh along with rendered photorealistic images. In the following, we present the preliminary knowledge including the instrument forward kinematics and 3D GS (Sect. 2.1), GS geometry pretraining (Sect. 2.2), instrument pose estimation & tracking (Sect. 2.3), and the final texture learning (Sect. 2.4).

2.1 Preliminary Knowledge

Forward Kinematics. In this study, we adopt the da Vinci EndoWrist Large Needle Driver (LND) as an example surgical instrument. The CAD model of this instrument is open source [8]. Note that the proposed method can be easily adapted to other types of instruments, given their CAD models.

The definition of coordinate systems and forward kinematics diagram are shown in Fig. 2. We regard shaft frame as the base frame and this frame definition consists of two revolute joint variables θ_1 and θ_2, denoting the rotation of the axis z_1 and y_2, and the gripper open angle θ_3. Let $q = \{\theta_1,\ \theta_2,\ \theta_3\} \in \mathbb{R}^3$ represent the instrument joint state. Given the translation d_0 and d_1 as 215.9 mm and 9 mm, the forward kinematics chains can be defined as ${}^{s}T_{w}(\theta_1)\ {}^{w}T_{g}(\theta_2)\ {}^{g}T_{l}(\theta_3)$, where ${}^{a}T_{b} \in SE(3)$ denotes the rigid transformation from frame a to b, and s, w, g and l denote the frames for shaft, wrist, gripper and left tool tips, respectively. Of note, each of them has only one degree of freedom (DoF) controlled by $\theta_i \in q$. Therefore, we can define a transformation from any joint frame j to the shaft frame as ${}^{s}T_{j\in\{w,\ g\ ,l\}}$ given corresponding joint states q. Next, with another transformation from shaft to camera ${}^{cam}T_{s}$, we define a transform function $\boldsymbol{T_j}(\cdot)$ to transform points $x_j \in \mathbb{R}^3$ in joint frames j to the camera frame as:

$$\boldsymbol{T_j}(x_j;\ q, \xi) = {}^{cam}T_{s}(\xi)\ {}^{s}T_{j}(q)\ x_j, \tag{1}$$

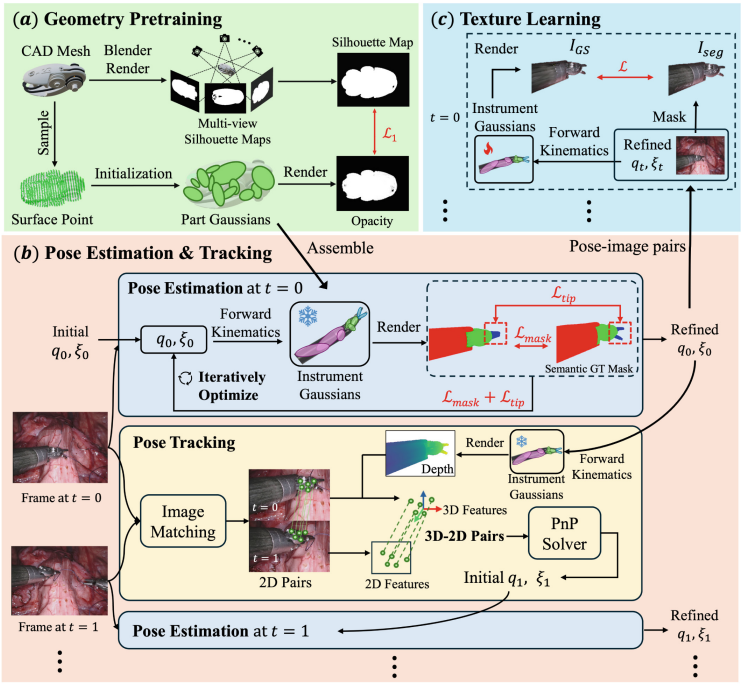


Fig. 3. Overview diagram of Instrument-Splatting methodology.

where we use $\xi \in se(3)$ to parameterize ${}^{cam}T_s$. With Eq. 1, we can define the instrument pose in camera frame given specified $(q,\ \xi)$.

3D Gaussian Splatting. 3D Gaussian Splatting (GS) [9] is a powerful 3D representation that can reconstruct 3D scenes with a group of Gaussian points. Each Gaussian point contains learnable attributes $\Theta = \{\mu \in \mathbb{R}^3, r \in \mathbb{R}^4, s \in \mathbb{R}^3, \alpha \in \mathbb{R}, sh \in \mathbb{R}^{27}\}$ to define the position, rotation, scale, opacity and colors. With camera parameters, α-blending [9] is performed to render the colored images based on the attributes in 3D GS model.

2.2 Geometry Pretraining

To bind the GS points to each rigid part of the instrument and initialize them with accurate geometric priors, we propose a geometry pretraining strategy, as illustrated in Fig. 3(a). Specifically, for each part, we densely sample points on mesh surface by ray-tracing, which effectively avoids the sparse point distribution due to oversized triangles. Then, given the CAD model, we adopt Blender[1] to render the ground-truth silhouette maps from multiple viewpoints. Next, the

[1] https://www.blender.org/.

sampled GS points will be trained to learn the geometry information by minimizing the $\mathcal{L}_1$ loss between the rendered opacity and the ground-truth silhouette. Afterward, part Gaussians are combined to generate semantic-embedded instrument Gasussians where one semantic label (i.e., wrist, shaft, and gripper) $f \in \{1, 2, 3\}$ is added to Gaussian parameters Θ to identify the part the Gaussian point belongs to. Therefore, the instrument GS can be controlled by updating μ and r with Eq. 1 as:

$$\mu'_j = \boldsymbol{T_j}(\mu_j;\ q, \xi); \quad r'_j = \boldsymbol{R_j}(r_j;\ q, \xi), \tag{2}$$

where μ'_j and r'_j respectively denote the updated positions and rotations for Gaussian points in part j, $\boldsymbol{R}_j$ means the rotational component derived from rigid-body transformation $\boldsymbol{T_j}$.

2.3 Pose Estimation and Tracking

An accurate alignment between the rendered instrument and the reference image is crucial for appearance learning. Therefore, we designe a pose estimation and tracking module to estimate per-frame instrument pose and joint changes, as illustrated in the Fig. 3(b). Inspired by [2], the pose estimation adopts the render-and-compare framework to align the rendered semantic silhouette with the segmented instrument. For the render-and-compare method, a good pose initialization is crucial. In the initial frame, we manually pick 2D-3D point correspondences of logo landmarks in the wrist part and use the PnP to solve the pose. Only wrist pose is needed for initialization; shaft and gripper poses are fixed relative to it and only 6 3D-2D pairs on wrist are sufficient. We cast the pose estimation task as an optimization problem that can be formulated as:

$$\widehat{\xi},\ \widehat{q} = argmin\ \mathcal{L}(\xi,\ q) \tag{3}$$

where the loss is $\mathcal{L} = \mathcal{L}_{mask} + \mathcal{L}_{tip}$. The $\mathcal{L}_{mask}$ is $\mathcal{L}_1$ loss. To take advantage of the surgical instrument geometry prior, we propose $\mathcal{L}_{tip} = \mathcal{L}_{dist} + \mathcal{L}_{struct}$ to enforce the tool tip alignment. Given the gripper mask, we use singular value decomposition (SVD) to find the principal axis of the gripper regions, on which the farthest pixel from the wrist part is regarded as the tool tip. We adopt a distance loss $\mathcal{L}_{dist} = ReLU(d - r)$, where d is the Euclidean distance between the rendered tool tip from instrument GS and the estimated tool tip position from segmentation mask, and r is a threshold to determine if the tool tip is aligned well. This loose regularization can alleviate falling into local minima caused by incorrect tool tip detection. In addition, we design a structural loss $\mathcal{L}_{struct} = ReLU(-(\theta_l + \theta_r))$ to constrain the left tip always on the left side of the right tip. θ_l is the left gripper rotation angle, which is a positive value if it rotates leftward and *vice versa* it is a negative value.

To track the instrument pose ξ and joint state q, simply taking the previous frame pose as the initial value struggles to handle the case where the instrument motion between two adjacent frames is large. To address this issue, we propose a novel pose initialization method based on image matching and PnP for pose

tracking. We extract features in the wrist and then find the matching correspondences [5] between two adjacent frames. Given the depth rendered from 3D GS, we can lift the 2D features to 3D. We approximate the frame t's 3D features and frame $t+1$'s 2D features into 2D-3D correspondences and then compute a rough initial pose using a PnP solver for the next frame.

2.4 Texture Learning

After estimating and tracking the pose and joint state over the video, we can control the instrument Gaussian points moving to the target pose in each frame, as shown in Fig. 3 (c). To maintain the geometric details, we freeze μ in this phase, and the other parameters $\{r,\ s,\ \alpha,\ sh\}$ are learnable during this section. Here, we adopt the loss $\mathcal{L} = \mathcal{L}_{mask} + \mathcal{L}_{color}$ to learn the textures, where the $\mathcal{L}_{mask}$ and $\mathcal{L}_{color}$ are the $\mathcal{L}_1$ loss of the semantic silhouette maps and RGB pixels of the instrument, respectively. $\mathcal{L}_{mask}$ is used here to maintain the geometric structure while learning appearances.

3 Experiments and Results

3.1 Datasets and Implementation Details

We establish a new benchmark for the evaluation of surgical instrument reconstruction. It includes two widely used instrument segmentation datasets EndoVis 2017 [3] and 2018 [1]. We select 100 frames containing LND in both datasets. Frames with motion artifacts are excluded to ensure data quality. We split each video into seen and unseen views (7:1) for training and evaluation. Because of the sparse instrument poses in the *in-vivo* datasets, we collect four videos with backgrounds of *ex vivo* tissues and a green screen for fairer validation. Two trajectories with tissue background are to mimic the surgical scenario. The other two trajectories with green screen backgrounds are closer to the camera to capture more texture details. The data collection methodology follows SurgPose [23] using a da Vinci IS1200 system with the da Vinci Research Kit (dVRK) [8]. Regarding annotation, we use SAM 2 [18] to segment the foreground surgical instrument at the part level and obtain corresponding semantic masks. This newly established benchmark, based on our in-house data, will be publicly available.

In this work, we perform geometry pretraining, pose estimation, and texture learning sequentially. We first train the part Gaussian for 10K iterations per part. Then, we adopt 2K iterations to iteratively optimize the pose for each frame, and an early-stop signal will be triggered once the loss remains stable for over 100 iterations. The texture learning lasts for 10K iterations for each video. Only training requires manual initialization; In real inference, we have instrument GS and input user-defined pose for controllable data generation. Therefore, we estimate the instrument poses and use these poses as user input to control the instrument GS for quantitative evaluation. All the experiments are based on the PyTorch framework and conducted with a single NVIDIA RTX A5000 GPU.

Table 1. Quantitative Results of the Key Components

Method	Rendered Semantics Dice↑			Reconstruction Quality		
	Shaft	Wrist	Gripper	PSNR↑	SSIM↑	LPIPS↓
w/o $\mathcal{L}_{tip}$	94.625	83.94	63.165	21.77	91.85	0.091
w/o Reg	95.515	78.4	63.39	22.89	92.63	0.079
w/o GP	–	–	–	14.6	89.54	0.16
Ours	**96.83**	**86.65**	**73.32**	**23.84**	**93.10**	**0.071**

3.2 Quantitative Evaluation of Key Components

First, we verify the effectiveness of the key components of Instrument-Splatting on the EndoVis 2017 and 2018 datasets. As shown in Table 1, we quantify the pose accuracy of our pose estimation and tracking module with 2D projection error similar to [2]. Specifically, we adopt the difference (i.e., Dice score) between rendered semantics maps from posed instrument Gaussians and the corresponding segmentation maps to evaluate the pose accuracy. We test our method by removing the $\mathcal{L}_{tip}$ and the loose regularization strategy, denoted as *w/o* $\mathcal{L}_{tip}$ and *w/o Reg*, respectively. As shown in Table 1, both $\mathcal{L}_{tip}$ and the loose regularization lead to significantly improved pose estimation accuracy, particularly for the gripper. We also report the metrics of the reconstruction quality in Table 1, where it is noticeable that an improved pose estimation can enhance the visual fidelity. Table 1 indicates the importance of Geometry Pretraining (GP), which significantly boosts the reconstruction quality. As shown in Fig. 4, six examples are selected from EndoVis18 and in-house dataset for visualization, which indicates that our method can robustly track the instrument pose.

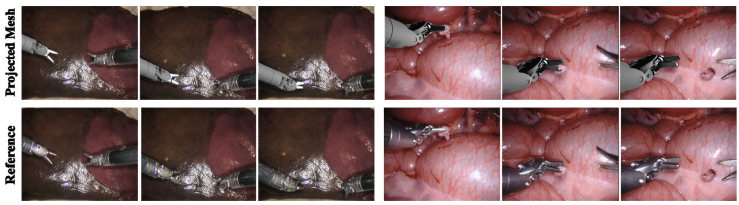


Fig. 4. Visualization of the projected CAD mesh with the estimated poses and corresponding original images.

3.3 Evaluation on Reconstruction Quality

In this work, we are targeting a novel setting that reconstructs articulated surgical instruments with monocular videos and CAD models only. We compare the reconstruction performance of our method with current state-of-the-art reconstruction methods in the surgical domain: EndoGaussian [13] and

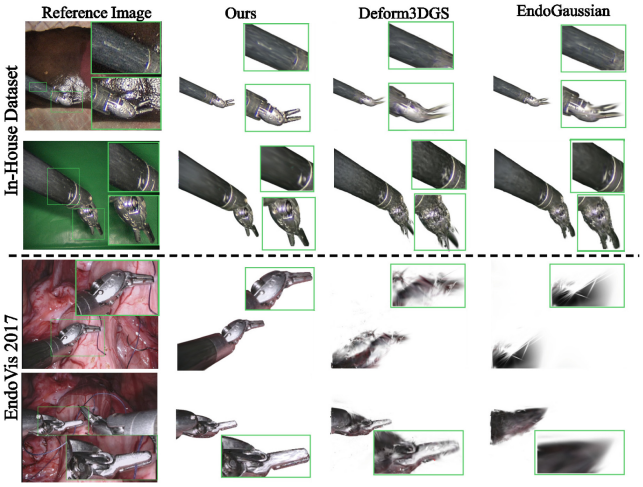


Fig. 5. Visualization of the Novel-view Rendering Results of Different Methods.

Deform3DGS [26]. Although both methods cannot provide controllable 3D assets, they demonstrate high visual fidelity in the reconstruction of deformable tissues.

Table 2. Quantitative evaluation of the reconstruction quality on EndoVis2017&2018 and In-House Dataset.

Dataset	Method	Wrist & Gripper			Shaft			Overall Instrument		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
EndoVis17&18	Deform3DGS	19.65	92.55	0.165	22.23	94.33	0.066	17.44	86.49	0.213
	EndoGaussian	18.90	93.45	0.145	20.91	94.30	0.078	16.45	87.75	0.192
	Ours	**25.81**	**96.87**	**0.064**	**31.39**	**96.52**	**0.035**	**23.84**	**93.10**	**0.071**
In-House	Deform3DGS	28.95	97.60	0.046	**34.52**	97.08	**0.023**	**28.39**	94.76	0.048
	EndoGaussian	25.34	96.94	0.059	31.59	**97.23**	0.033	27.47	94.94	0.053
	Ours	**30.44**	**98.27**	**0.038**	32.49	97.14	0.032	27.94	**95.45**	**0.041**

As shown in Table 2, we show the results of wrist&gripper, shaft, and the whole instrument separately. The metrics for the whole instrument are lower than the other two since they are computed based on all the pixels, following [13]. We visualize the rendered images of different methods in Fig. 5. Both EndoGaussian and Deform3DGS fail to reconstruct instruments in real surgical videos (EndoVis2017&2018) due to long-range motions and large time intervals. As a comparison, benefiting from the pose estimation module, our method can effectively capture the significant inter-frame pose change for real surgical videos, leading to a largely improved accuracy. On in-house data with smoother

and slower motions, although Deform3DGS/EndoGaussian have comparable or even higher PSNR scores, their reconstruction outputs suffer from significant artifacts on wrist and gripper parts and are far from being photorealistic, as shown in the visualization (Fig. 5). This can be explained by the motion patterns in some clips of our in-house data where the shaft part only undergoes limited translation, and Deform3DGS/EndoGaussian has a strong capability of capturing the appearance of these regions. Also, Table 2 shows the two methods exhibit an inferior reconstruction quality on wrist and gripper parts, which indicates the existing methods on surgical scene reconstruction cannot handle articulated instrument parts with high dexterity and complex motions.

4 Conclusion

In this paper, we propose Instrument-Splatting, a novel Real2Sim pipeline to reconstruct articulated robotic surgical instruments with high visual fidelity. With only monocular videos and CAD models of the instruments, our Instrument-Splatting can effectively track the per-frame instrument pose and learn the texture through RGB videos. Different from prior works in the surgical domain, our instrument Gaussian is articulated and fully controllable, which allows for manipulation as flexible as real instruments. We believe this digital twin of the instrument can strongly benefit the community working on surgical AI or surgical autonomy. The experimental results demonstrate the superior advances of our proposed method.

Acknowledgement. This work was supported by Ministry of Education Tier 1 Start up grant, NUS, Singapore (A-8001267-01-00); Ministry of Education Tier 1 grant, NUS, Singapore (A-8001946-00-00); Charles Laszlo Chair in BME; student support funds, UBC, School of BME.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allan, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Allan, M., Ourselin, S., Hawkes, D.J., Kelly, J.D., Stoyanov, D.: 3-D pose estimation of articulated instruments in robotic minimally invasive surgery. IEEE Trans. Med. Imaging **37**(5), 1204–1213 (2018)
3. Allan, M., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
4. Barragan, J.A., Zhang, J., Zhou, H., Munawar, A., Kazanzides, P.: Realistic data generation for 6D pose estimation of surgical instruments. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 13347–13353. IEEE (2024)

5. Chen, H., et al.: ASpanFormer: detector-free image matching with adaptive span transformer. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13692, pp. 20–36. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19824-3_2
6. Chen, Y., Wu, Z., Schmidt, A., Salcudean, S.E.: A-MFST: adaptive multi-flow sparse tracker for real-time tissue tracking under occlusion. arXiv preprint arXiv:2410.19996 (2024)
7. Du, X., et al.: Articulated multi-instrument 2-D pose estimation using fully convolutional networks. IEEE Trans. Med. Imaging **37**(5), 1276–1287 (2018)
8. Kazanzides, P., Chen, Z., Deguet, A., Fischer, G.S., Taylor, R.H., DiMaio, S.P.: An open-source research kit for the da Vinci® surgical system. In: 2014 IEEE international conference on robotics and automation (ICRA), pp. 6434–6439. IEEE (2014)
9. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4) (2023)
10. Kim, J.W., et al.: Surgical robot transformer (SRT): imitation learning for surgical tasks. arXiv preprint arXiv:2407.12998 (2024)
11. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: HUGS: human gaussian splats. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 505–515 (2024)
12. Liu, H., Liu, Y., Li, C., Li, W., Yuan, Y.: LGS: a light-weight 4D gaussian splatting for efficient surgical scene reconstruction. arXiv preprint arXiv:2406.16073 (2024)
13. Liu, Y., Li, C., Yang, C., Yixuan, Y.: EndoGaussian: Gaussian splatting for deformable surgical scene reconstruction. Arxiv (2024)
14. Lou, H., et al.: Robo-GS: a physics consistent spatial-temporal model for robotic arm with hybrid representation. arXiv preprint arXiv:2408.14873 (2024)
15. Lu, G., Zhang, S., Wang, Z., Liu, C., Lu, J., Tang, Y.: ManiGaussian: dynamic Gaussian splatting for multi-task robotic manipulation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV 2024. LNCS, vol. 15093, pp. 349–366. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72761-0_20
16. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2021)
17. Munawar, A., Wu, J.Y., Fischer, G.S., Taylor, R.H., Kazanzides, P.: Open simulation environment for learning and practice of robot-assisted surgical suturing. IEEE Robot. Autom. Lett. **7**(2), 3843–3850 (2022)
18. Ravi, N., et al.: SAM 2: segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
19. Reiley, C.E., Plaku, E., Hager, G.D.: Motion generation of robotic surgical tasks: learning from expert demonstrations. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 967–970. IEEE (2010)
20. Wang, X., Zhang, S., Huang, B., Stoyanov, D., Mazomenos, E.B.: EndoLRMGS: complete endoscopic scene reconstruction combining large reconstruction modelling and gaussian splatting (2025)
21. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 431–441. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_41

22. Wu, Z., Schmidt, A., Kazanzides, P., Salcudean, S.E.: Augmenting efficient real-time surgical instrument segmentation in video with point tracking and segment anything. Healthc. Technol. Lett. (2024)
23. Wu, Z., et al.: SurgPose: a dataset for articulated robotic surgical tool pose estimation and tracking. arXiv preprint arXiv:2502.11534 (2025)
24. Xu, H., et al.: SurgRIPE challenge: benchmark of surgical robot instrument pose estimation. arXiv preprint arXiv:2501.02990 (2025)
25. Yang, C., Wang, K., Wang, Y., Yang, X., Shen, W.: Neural LerPlane representations for fast 4D reconstruction of deformable tissues. arXiv preprint arXiv:2305.19906 (2023)
26. Yang, S., Li, Q., Shen, D., Gong, B., Dou, Q., Jin, Y.: Deform3DGS: flexible deformation for fast surgical scene reconstruction with gaussian splatting. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15006, pp. 132–142. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72089-5_13
27. Ye, M., Zhang, L., Giannarou, S., Yang, G.-Z.: Real-time 3D tracking of articulated tools for robotic surgery. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016, Part I. LNCS, vol. 9900, pp. 386–394. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46720-7_45
28. Zeng, T., Loza Galindo, G., Hu, J., Valdastri, P., Jones, D.: Realistic surgical image dataset generation based on 3D gaussian splatting. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15006, pp. 510–519. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72089-5_48
29. Zhan, B., et al.: Tracking everything in robotic-assisted surgery. arXiv preprint arXiv:2409.19821 (2024)
30. Zhu, L., Wang, Z., Jin, Z., Lin, G., Yu, L.: Deformable endoscopic tissues reconstruction with gaussian splatting. arXiv preprint arXiv:2401.11535 (2024)

Is Hyperbolic Space All You Need for Medical Anomaly Detection?

Alvaro Gonzalez-Jimenez[2,3(✉)], Simone Lionetti[2], Ludovic Amruthalingam[2], Philippe Gottfrois[1,3], Fabian Gröger[1,2], Marc Pouly[2], and Alexander A. Navarini[1,3]

[1] University of Basel, Basel, Switzerland
[2] Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland
[3] University Hospital of Basel, Basel, Switzerland
alvaro.gonzalezjimenez@usb.ch

Abstract. Medical anomaly detection has emerged as a promising solution to challenges in data availability and labeling constraints. Traditional methods extract features from different layers of pre-trained networks in Euclidean space; however, Euclidean representations fail to effectively capture the hierarchical relationships within these features, leading to suboptimal anomaly detection performance. We propose a novel yet simple approach that projects feature representations into hyperbolic space, aggregates them based on confidence levels, and classifies samples as healthy or anomalous. Our experiments demonstrate that hyperbolic space consistently outperforms Euclidean-based frameworks, achieving higher AUROC scores at both image and pixel levels across multiple medical benchmark datasets. Additionally, we show that hyperbolic space exhibits resilience to parameter variations and excels in few-shot scenarios, where healthy images are scarce. These findings underscore the potential of hyperbolic space as a powerful alternative for medical anomaly detection. The project website can be found at https://hyperbolic-anomalies.github.io.

Keywords: Hyperbolic Learning · Anomaly Detection · Hierarchical

1 Introduction

Anomaly detection and localization plays a critical role in various domains, particularly in medical imaging, where distinguishing and localizing between normal and anomalous samples is crucial. A widely adopted approach involves training models exclusively on healthy images, identifying any deviation from this learned distribution as anomalous [32]. This strategy mitigates challenges associated with the scarcity of annotated lesion images while reducing annotation costs and biases inherent in training Artificial Intelligence (AI) models.

A. A. Navarini and M. Pouly—These authors are joint last authors.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 312–322, 2026.
https://doi.org/10.1007/978-3-032-04947-6_30

Among the most effective anomaly detection techniques are projection-based methods, which leverage pre-trained networks to map data into abstract representations, thereby enhancing the separation between normal and anomalous samples. One-class classification [23,29] defines a compact, closed distribution for normal samples, treating any deviations as anomalies. The teacher-student framework [9,35] employs a student network to learn normal sample representations from a teacher, using their representation discrepancy to identify anomalies. Memory Bank methods [8,22,28] store normal sample prototypes and apply statistical modeling or distance metrics to detect anomalies.

A common feature across these methods is the extraction of representations from specific layers of a pre-trained network. Each layer encodes hierarchical attributes, but conventional approaches rely on Euclidean space, which may not be the best option to capture hierarchical relationships [27,30]. This can lead to suboptimal feature representations and reduced anomaly identification performance.

Hyperbolic space, the geometry of constant negative curvature, is well suited to represent hierarchical structures due to its exponential expansion properties [6]. Recent advances have demonstrated the effectiveness of hyperbolic embeddings in domains such as few-shot learning [18], representation learning [11,12,21], and Out-of-Distribution (OOD) detection [14,33]. Given the hierarchical nature of medical image structures, including disease organization, progression, and anatomical relationships, we hypothesize that hyperbolic space can effectively embed these spatial relationships to enhance anomaly localization. This work aims to answer the following research question: **Can hyperbolic space effectively represent hierarchical features and improve anomaly localization performance?**

To address this question, we propose a novel framework that generate synthetic anomalies, extracts multi-layer features from a pre-trained network and projects them into hyperbolic space. These hyperbolic embeddings are aggregated by weighting features based on their confidence, specifically considering their distance from the origin [13,18], which encodes hierarchical depth. Finally, we construct a hyperplane in hyperbolic space to distinguish between normal and anomalous samples.

We validate our framework on multiple medical benchmark datasets including different imaging modalities such as Magnetic Resonance Imaging (MRI), Computer Tomography (CT), Optical Coherence Tomography (OCT) and X-Ray. Our results demonstrate that hyperbolic space consistently outperforms Euclidean space for anomaly detection and localization. Additionally, we find that hyperbolic space exhibits robustness to parameter tuning by adaptively learning the optimal curvature, further improving performance. Notably, our approach achieves state-of-the-art results in few-shot settings, where healthy images are scarce or unavailable.

The paper is organized as follows: In Sect. 2, we introduce the motivation behind our framework and provide its mathematical formulation. Section 3 details the datasets, implementation, and training specifics, along with evalua-

tion metrics. Section 4 presents experimental findings, comparing hyperbolic and Euclidean-based methods and analyzing performance under few-shot conditions. Finally, Sect. 5 summarizes our contributions and discusses broader implications.

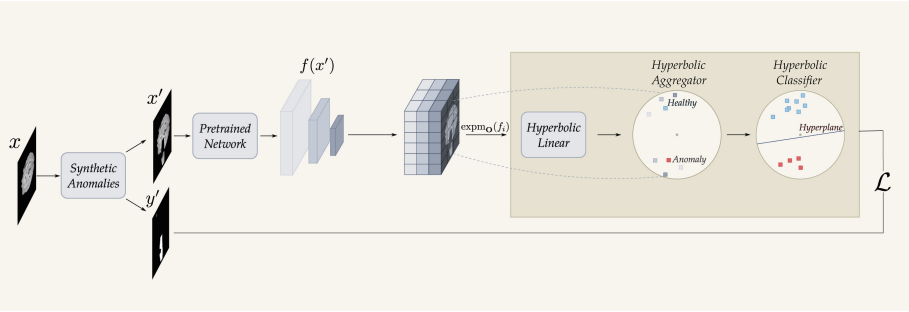


Fig. 1. Overview of the anomaly localization methodology in the hyperbolic space, from medical anomaly synthesis to classification.

2 Methodology

This section details our framework to anomaly localization in hyperbolic space, illustrated in Fig. 1. In Sect. 2.1 we describe our method for synthesizing medical anomalies. Section 2.2 describes how we obtain patchified features from a pre-trained network. Section 2.3 outlines the mapping of patchified Euclidean features to hyperbolic space, followed by the hierarchical aggregation process. Finally, in Sect. 2.4, we present our hyperbolic classifier, which leverages these aggregated features for classification.

2.1 Synthesis Anomalies

Given a training set of normal images $x_i \in \mathbb{R}^{H_0 \times W_0 \times C_0}$, we generate images x_i' with synthetic anomalies. These include

- **CutPaste** [23], random patches extracted and blended in another location with Poisson image editing [3,31];
- **Gaussian Intensity** [37], intensity variations introduced via Gaussian filtering to simulate anomalies such as tumors or cysts;
- **Source Deformation** [3], geometric deformations applied by shifting points within a mask, controlled by a scaling parameter.

2.2 Feature Extraction

Features are extracted from the anomaly image x'_i using a pre-trained network, typically a ResNet-like backbone. We select a subset of feature levels L, corresponding to different layers in the network hierarchy. At each level, local features are computed by aggregating patch-wise neighborhoods using adaptive average pooling. The resulting feature maps are then upsampled to the highest spatial resolution among them for later aggregation to give a feature map $f_{i,l} \in \mathbb{R}^C$.

2.3 Euclidean to Hyperbolic Features

Hyperbolic geometry, characterized by constant negative curvature, effectively models hierarchical structures [27,30]. We employ the Lorentz model due to its simple expression for geodesics [21] and numerical robustness [26]. Minkowski space is the space of vector $\mathbf{z} = (z_0, \boldsymbol{z}) \in \mathbb{R} \times \mathbb{R}^n$ equipped with the Lorentz inner product $\langle \mathbf{z}, \mathbf{z}' \rangle_{\mathbb{L}} = \boldsymbol{z} \cdot \boldsymbol{z}' - z_0 z'_0$.

The Lorentz hyperboloid model $\mathbb{L}^n_c$ of n-dimensional hyperbolic space with curvature c is the manifold that satisfies $\langle \mathbf{z}, \mathbf{z} \rangle_{\mathbb{L}} = -1/c$ with $z_0 > 0$. As the feature vectors lie on the Euclidean space we use the exponential map to projects them onto the hyperboloid

$$\exp\mathrm{m}_{\mathbf{z}}(\mathbf{v}) = \cosh\left(\sqrt{c}\,\|\mathbf{v}\|_{\mathbb{L}}\right)\mathbf{z} + \frac{\sinh\left(\sqrt{c}\,\|\mathbf{v}\|_{\mathbb{L}}\right)}{\sqrt{c}\,\|\mathbf{v}\|_{\mathbb{L}}}\mathbf{v}, \tag{1}$$

so $\mathbf{f}_{i,l} = \exp\mathrm{m}_{\mathbf{O}}(f_{i,l})$, where $\mathbf{O} = (1/\sqrt{c}, \mathbf{0})$ is the hyperboloid origin.

We project the hyperbolic features to a lower-dimensional hyperbolic space, and adapt the features to the target domain with a hyperbolic linear layer [4], as the network is biased from the pre-training dataset *i.e.*, *ImageNet* [10] which is suboptimal in medical contexts. We aggregate features from different hierarchical levels to a single point in hyperbolic space $\mathbf{z}_i$ using a weighted Lorentzian centroid [21]:

$$\mathbf{z}_i = \sqrt{c}\frac{\mathbf{z}'_i}{\left|\|\mathbf{z}'_i\|_{\mathbb{L}}\right|} \quad \text{with} \quad \mathbf{z}'_i = \sum_{l \in L} w_{i,l}\mathbf{f}_{i,l}. \tag{2}$$

The weights $w_{i,l}$ are the Euclidean L_2 norms of features $\mathbf{f}_{i,l}$ after transformation to the Poincaré ball, which are connected to model confidence [13,14,18].

2.4 Hyperbolic Classifier

We classify anomalous features using distances to hyperplanes in the Lorentz model. The hyperplane in $\mathbb{L}^n_c$ perpendicular to $\mathbf{w}$ is given by

$$H_{\mathbf{w}} = \{\mathbf{y} \in \mathbb{L}^n_c \,|\, \langle \mathbf{w}, \mathbf{y} \rangle_{\mathbb{L}} = 0\}, \tag{3}$$

and the distance of a point $\mathbf{z}$ from the hyperplane reads

$$d_{\mathbb{L}}\left(\mathbf{z}, H_{\mathbf{w}}\right) = \frac{1}{\sqrt{c}}\left|\sinh^{-1}\left(\sqrt{c}\frac{\langle \mathbf{w}, \mathbf{z} \rangle_{\mathbb{L}}}{\|\mathbf{w}\|_{\mathbb{L}}}\right)\right|. \tag{4}$$

The patch-wise logit and probability for an image x_i with representation $\mathbf{z}_i$ to be anomalous are then given by

$$\ell_{\mathbf{w}}(\mathbf{z}_i) = \operatorname{sign}\left(\langle \mathbf{w}, \mathbf{z}_i \rangle_{\mathbb{L}}\right) \|\mathbf{w}\|_{\mathbb{L}} d_{\mathbb{L}}\left(\mathbf{z}_i, H_{\mathbf{w}}\right), \qquad p_{\mathbf{w}}(\mathbf{z}_i) = \left[1 + \exp\left(\ell_{\mathbf{w}}(\mathbf{z}_i)\right)\right]^{-1}. \tag{5}$$

The model constructs a hyperplane for robust class discrimination by optimizing it through binary cross-entropy

$$\mathcal{L} = -\mathbb{E}_{x_i \sim \mathcal{A}}[\log\left(p_{\mathbf{w}}(\mathbf{z}_i)\right)] - \mathbb{E}_{x_i \sim \mathcal{N}}[\log\left(1 - p_{\mathbf{w}}(\mathbf{z}_i)\right)] \tag{6}$$

where $\mathcal{A}$ and $\mathcal{N}$ are sets of anomalous and normal pixel centroids computed via Eq. (2), respectively.

3 Experiments

This section describes how hyperbolic space is evaluated for anomaly detection and localization.

3.1 Datasets

We follow BMAD [2], a recent benchmark for medical anomaly detection and localization spanning different imaging modalities. It features defined dataset splits to facilitate reproducibility and prevent leakage. We only deviate by excluding the pathology dataset Camelyon16 due to known difficulties with memory, and by resizing all images to 224 × 224 pixels. The five datasets used in this work are summarized in Table 1.

BraTS2021 [1] is a widely used dataset for brain tumor segmentation and classification in MRI, BMAD considers the FLAIR sequences for anomaly detection.

BTCV [20] and **Liver Tumor Segmentation (LiTs)** [5] focus on liver CT imaging. BMAD uses the anomaly-free BTCV set for training and LiTs for evaluation.

The **Retinal Edema Segmentation Challenge (RESC)** [16] provides OCT images for retinal pathology analysis.

OCT2017 [17] is a large-scale OCT dataset for retinal disease classification, comprising one normal category and three medical conditions. The latter are treated as a single abnormal class.

RSNA [34] contains chest X-rays labeled with one normal category and eight conditions, all of which are treated as a single abnormal class.

Table 1. Count of normal and anomalous samples across BMAD dataset splits.

Dataset	BraTS2021		BTCV+LiTs		RESC		OCT2017		RSNA	
Split ↓	norm.	anom.	norm.	anom.	norm.	anom.	norm.	anom.	norm.	anom.
Train	7,500	0	1,542	0	4,297	0	26,315	0	8,000	0
Valid	39	44	93	73	45	70	8	24	70	1,420
Test	640	3,075	833	660	1,041	764	242	726	781	16,413

3.2 Experimental Setup

We use a pre-trained WideResNet50 [36] as feature extractor in all experiments. To ensure a fair comparison with baseline methods, we refrain from using data augmentation, applying only ImageNet-based normalization [10]. The pre-trained network is frozen, and only the hyperbolic components are trained. We extract features from `layer_2` and `layer_3`, with a dimensionality of 1024, which are subsequently patchified using a patch size of 3.

The curvature parameter is trainable and initialized to $c = 1$. Training is conducted for 50 epochs across all datasets using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 32. All experiments are performed on a single NVIDIA Tesla V100 GPU with 32 GB of memory.

We evaluate both image-level (detection) and pixel-level (localization) performance using Image-AUROC (I_{AUROC}) and Pixel-AUROC (P_{AUROC}) in percentage, respectively. In addition to our hyperbolic approach, we benchmark against several state-of-the-art Euclidean anomaly detection and localization models, including RD4AD [9], STFPM [35], PaDiM [8], PatchCore [28], and CFA [22].

Finally, we test for statistical significance using the Mann-Whitney U test to compare the AUROC distributions between two models. We assume statistical significance for $p < 0.05$ and denote this with **bold**.

4 Results

Table 2. Comparison of anomaly detection and localization performance across medical datasets. The values represent the mean, the minimum (subscript), and maximum (superscript) over 5 different random seeds.

Methods	BraTS2021		BTCV + LiTs		RESC		OCT2017	RSNA
	I_{AUROC}	P_{AUROC}	I_{AUROC}	P_{AUROC}	I_{AUROC}	P_{AUROC}	I_{AUROC}	I_{AUROC}
RD4AD	$89.52^{90.19}_{88.85}$	$96.36^{96.48}_{96.24}$	$59.14^{64.45}_{53.83}$	$91.40^{91.50}_{91.30}$	$88.25^{90.25}_{86.25}$	$96.18^{96.38}_{95.98}$	$94.88^{97.58}_{92.17}$	$67.63^{68.73}_{66.53}$
STFPM	$84.25^{86.63}_{81.87}$	$96.03^{96.43}_{95.63}$	$61.48^{63.15}_{59.81}$	$96.26^{96.40}_{96.12}$	$87.26^{87.49}_{87.03}$	$94.96^{95.02}_{94.90}$	$91.88^{93.21}_{90.55}$	$69.31^{70.4}_{68.22}$
PaDiM	$79.62^{80.96}_{78.28}$	$94.22^{94.45}_{93.99}$	$50.91^{51.24}_{50.58}$	$90.48^{90.63}_{90.33}$	$75.15^{76.57}_{73.73}$	$91.22^{91.59}_{90.85}$	$90.17^{90.78}_{89.56}$	$74.48^{74.74}_{74.22}$
PatchCore	$92.02^{92.13}_{91.91}$	$95.53^{95.58}_{95.48}$	$59.33^{59.47}_{59.19}$	$95.00^{95.01}_{94.99}$	$90.54^{90.64}_{90.44}$	$95.87^{95.91}_{95.83}$	$97.45^{98.10}_{96.80}$	$75.67^{75.87}_{75.47}$
CFA	$84.99^{85.15}_{84.83}$	$96.61^{96.65}_{96.57}$	$53.89^{58.13}_{49.65}$	$97.40^{97.46}_{97.34}$	$72.47^{74.74}_{70.20}$	$92.49^{93.57}_{91.41}$	$79.10^{79.66}_{78.54}$	$66.65^{66.80}_{66.50}$
Ours	$92.49^{93.02}_{91.96}$	$95.56^{95.63}_{95.49}$	$\mathbf{65.94^{67.99}_{63.89}}$	$96.49^{99.11}_{93.87}$	$90.71^{91.28}_{90.14}$	$95.32^{95.56}_{95.08}$	$97.85^{98.12}_{97.58}$	$\mathbf{79.46^{80.20}_{78.72}}$

Table 2 presents the experimental results for anomaly detection and localization. We observe acceptable agreement of the methods based on Euclidean geometry with the results reported by the BMAD in Table 2 [2]. Among these, PatchCore achieves the most consistent performance across datasets, although the somewhat lower performance for BTCV+LiTs could be interpreted as sensitivity to a distribution shift. In contrast, the proposed hyperbolic framework shows the best performance on whole images across all datasets, even if this is not always statistically significant, and it remains robust even for BTCV+LiTs. At the pixel level, the hyperbolic approach remains competitive with other methods, even though different Euclidean baselines outperform it in specific cases. However, in medical practice, misdiagnosing an entire image is generally more problematic than minor pixel-wise mismatches.

4.1 Ablation Study on Model Parameters

We conduct ablation studies on curvature, patch size, and dimensionality of the hyperbolic space using the BraTS dataset. Figure 2 presents the impact of these variations on performance.

We first investigate the role of curvature by fixing it to $c = \{0.01, 0.1, 1, 10, 100\}$. The first plot indicates that constraining the curvature leads to a decline in performance, with better results observed at lower curvature values. This underscores the advantage of a learnable curvature, which allows the model to adaptively optimize the geometry of the representation space for anomaly identification.

Next, we analyze the effect of the patch size $\{1, 2, 3, 4, 5, 6\}$ when aggregating local features. Increasing the patch size negatively impacts both I_{AUROC} and P_{AUROC}. This suggests that fine-grained feature extraction is preferable for capturing subtle anomalies, whereas overly large patches may dilute local information critical for accurate anomaly localization.

Lastly, hyperbolic space has been shown to efficiently encode representations in lower-dimensional embeddings, making it advantageous for memory-constrained scenarios [14,19]. To evaluate this, we reduce the feature dimensionality to $\{512, 128, 16, 8, 2\}$. The last plot reveals that while I_{AUROC} is more sensitive to extreme dimensionality reduction, P_{AUROC} remains relatively stable.

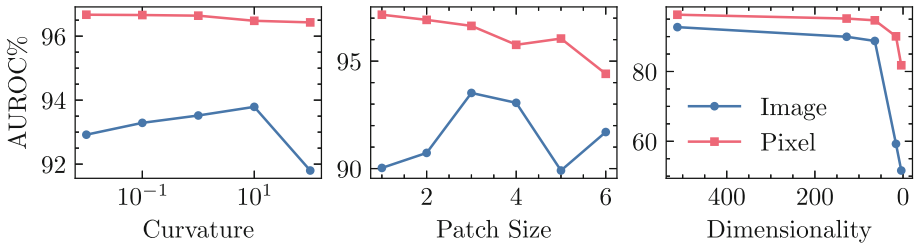


Fig. 2. Ablation study on key components of our hyperbolic framework: fixed curvature, patch size variations, and hyperbolic layer dimensionality.

4.2 Few-Shot Anomaly Detection and Localization

We further evaluate the robustness of our framework in a few-shot setting, where only a limited number of normal images are available for training. We experiment with $\{1, 3, 5, 10, 25\}$ normal images and compare our performance against PaDiM [8] and PatchCore [28]. The results in Fig. 3 demonstrate that our hyperbolic model significantly outperforms both baselines, particularly in extreme data scarcity scenarios.

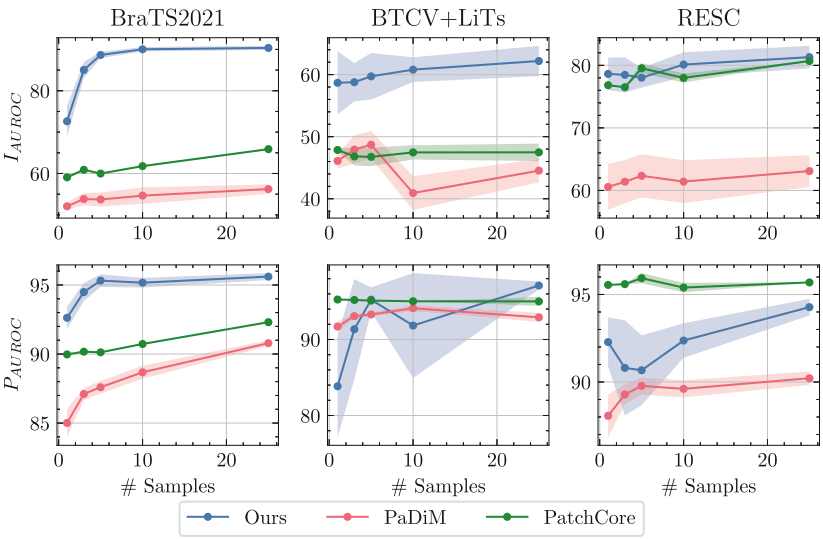


Fig. 3. Few-shot evaluation with varying normal image counts $\{1, 3, 5, 10, 25\}$. Our hyperbolic model outperforms PaDiM and PatchCore in scarce data scenarios. Error bands are obtained with five different random seeds, without changing the training set.

5 Conclusions

In this work, we introduced a hyperbolic anomaly detection and localization framework that leverages the unique geometric properties of hyperbolic space to enhance both classification and localization of medical anomalies.

Our evaluation across multiple medical imaging datasets demonstrates that our method consistently outperforms state-of-the-art anomaly identification approaches in terms of I_{AUROC}, and is competitive with the best ones for localization as demonstrated by P_{AUROC}. Additionally, we show that hyperbolic embeddings retain strong performance in low dimensions enabling efficient deployment in resource-constrained environments, and consistently outperforms Euclidean baselines in few-shot data regimes.

One key area for future investigation is the incorporation of features from earlier layers of the model, which could help leverage the hierarchical information embedded throughout the network. Although fully hyperbolic networks have been shown to outperform hybrid architectures [4,7,33], this remains a developing research area, and challenges related to stability and reproducibility persist. Furthermore, while our study focused on feature-based anomaly baselines, expanding the comparison to reconstruction-based [24,25] or gradient-based [15] with other performance metrics could provide a more comprehensive assessment of the model's effectiveness.

Additionally, integrating multi-modal data, such as radiology reports or genomic information, could provide a richer anomaly characterization, improving interpretability and clinical utility.

Our work contributes to the advancement of AI-driven medical anomaly detection and localization, with a particular emphasis on improving the accuracy and localization of anomalies, especially in few-shot settings. These findings have the potential to significantly enhance medical image quality assessment and facilitate the quantitative analysis of rare diseases, ultimately leading to more precise and data-efficient diagnostic models for clinical applications.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baid, U., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification (2021)
2. Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., Li, X.: BMAD: benchmarks for medical anomaly detection. In: CVPRW (2024)
3. Baugh, M., Tan, J., Müller, J.P., Dombrowski, M., Batten, J., Kainz, B.: Many tasks make light work: learning to localise medical anomalies from multiple synthetic tasks. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14220, pp. 162–172. Springer, Cham. (2023). https://doi.org/10.1007/978-3-031-43907-0_16
4. Bdeir, A., Schwethelm, K., Landwehr, N.: Fully hyperbolic convolutional neural networks for computer vision. In: ICLR (2023)
5. Bilic, P., et al.: The liver tumor segmentation benchmark (LiTS). Med. Image Anal. (2023)
6. Bridson, M.R., Haefliger, A.: Metric spaces of non-positive curvature (1999)
7. Chen, W., et al.: Fully hyperbolic neural networks. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2022)
8. Defard, T., Setkov, A., Loesch, A., Audigier, R.: PaDiM: a patch distribution modeling framework for anomaly detection and localization. In: Del Bimbo, A., Cucchiara, R., Sclaroff, S., Farinella, G.M., Mei, T., Bertini, M., Escalante, H.J., Vezzani, R. (eds.) ICPR 2021. LNCS, vol. 12664, pp. 475–489. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68799-1_35
9. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: CVPR (2022)

10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
11. Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., Vedantam, S.R.: Hyperbolic image-text representations. In: ICML (2023)
12. Ganea, O.E., Bécigneul, G., Hofmann, T.: Hyperbolic entailment cones for learning hierarchical embeddings. In: ICML (2018)
13. GhadimiAtigh, M., Schoep, J., Acar, E., van Noord, N., Mettes, P.: Hyperbolic image segmentation. In: CVPR (2022)
14. Gonzalez-Jimenez, A., et al.: Hyperbolic metric learning for visual outlier detection. In: Del Bue, A., Canton, C., Pont-Tuset, J., Tommasi, T. (eds.) ECCV 2024. LNCS, vol. 15639, pp. 327–344. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-91585-7_20
15. Gonzalez-Jimenez, A., Lionetti, S., Pouly, M., Navarini, A.A.: SANO: score-based diffusion model for anomaly localization in dermatology. In: CVPRW, pp. 2988–2994 (2023)
16. Hu, J., Chen, Y., Yi, Z.: Automated segmentation of macular edema in OCT using deep neural networks. Med. Image Anal. (2019)
17. Kermany, D.S., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell (2018)
18. Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., Lempitsky, V.: Hyperbolic image embeddings. In: CVPR (2020)
19. Kusupati, A., et al.: Matryoshka representation learning. In: NeurIPS (2024)
20. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Multi-atlas labeling beyond the cranial vault - workshop and challenge. In: MICCAI (2015)
21. Law, M., Liao, R., Snell, J., Zemel, R.: Lorentzian distance learning for hyperbolic representations. In: ICML (2019)
22. Lee, S., Lee, S., Song, B.C.: CFA: coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. IEEE Access (2022)
23. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: CutPaste: self-supervised learning for anomaly detection and localization. In: CVPR (2021)
24. Liu, J., et al.: A survey on diffusion models for anomaly detection. arXiv preprint arXiv: 2501.11430 (2025)
25. Livernoche, V., Jain, V., Hezaveh, Y., Ravanbakhsh, S.: On diffusion modeling for anomaly detection. In: ICLR (2023)
26. Mishne, G., Wan, Z., Wang, Y., Yang, S.: The numerical stability of hyperbolic representation learning. In: ICML (2023)
27. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. In: NeurIPS (2017)
28. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: towards total recall in industrial anomaly detection. In: CVPR (2022)
29. Ruff, L., et al.: Deep one-class classification. In: ICML (2018)
30. Sala, F., Sa, C.D., Gu, A., Re, C.: Representation tradeoffs for hyperbolic embeddings. In: ICML (2018)
31. Schlüter, H.M., Tan, J., Hou, B., Kainz, B.: Natural synthetic anomalies for self-supervised anomaly detection and localization. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13691, pp. 474–489. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19821-2_27
32. Tschuchnig, M.E., Gadermayr, M.: Anomaly detection in medical imaging - a mini review. In: Data Science – Analytics and Applications (2022)
33. van Spengler, M., Berkhout, E., Mettes, P.: Poincare ResNet. In: ICCV (2023)

34. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR (2017)
35. Yamada, S., Hotta, K.: Reconstruction student with attention for student-teacher pyramid matching (2022)
36. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
37. Zhang, X., Xu, M., Qiu, D., Yan, R., Lang, N., Zhou, X.: MediCLIP: adapting CLIP for few-shot medical image anomaly detection. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15011, pp. 458–468. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72120-5_43

LDDMEm: Large Deformation Diffeomorphic Metric Embedding

Decoupling Shape Analysis from Image Registration

Greg M. Fleishman[1(✉)] and P. Thomas Fletcher[2]

[1] HHMI, Janelia Research Campus, Ashburn, USA
fleishmang@janelia.hhmi.org
[2] University of Virginia, Charlottesville, USA

Abstract. We present a method, open-source software, and experiments which embed arbitrary deformation vector fields produced by any method (e.g., ANTs or VoxelMorph) in the Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework. This decouples formal diffeomorphic shape analysis from image registration, which has many practical benefits. Shape analysis can be added to study designs without modification to already chosen image registration methods and existing databases of deformation fields can be reanalyzed within the LDDMM framework without repeating image registrations. Pairwise time series studies can be extended to full time series regression with minimal added computing. The diffeomorphic rigor of image registration methods can be compared by embedding deformation fields and comparing projection distances. Finally, the added value of formal diffeomorphic shape analysis can be more fairly evaluated when it is derived from and compared to a baseline set of deformation fields. In brief, the method is a straightforward use of geodesic shooting in diffeomorphisms with a deformation field as the target, rather than an image. This is simpler than the image registration case which leads to a faster implementation that requires fewer user derived parameters.

Keywords: LDDMM · Diffeomorphisms · Image Registration · Shape Analysis · Computational Anatomy · Manifold Statistics

1 Motivation

The Large Deformation Diffeomorphic Metric Mapping (LDDMM) framework [5] is a beautiful example of advanced mathematics applied to a practical problem. In addition to registering two images it provides a rigorous guarantee of transform smoothness. Additionally, it can model real continuum mechanical properties of imaged materials [15], shape interpolate image time series of arbitrary length [14], and provide a rigorous setting in which to conduct shape statistics [8]. While LDDMM has been substantially developed in the literature, its adoption for practical use has not kept pace due to several barriers. For practical studies, robust and accurate estimation of correspondence between noisy

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 323–332, 2026.
https://doi.org/10.1007/978-3-032-04947-6_31

images is of primary importance. Non-LDDMM methods are less encumbered by advanced theoretical foundations, are less computationally complex, and their software development has focused on features that strengthen the capacity to accurately register noisy image data. Such features include multiple transformation models, image matching functionals, regularization schemes, multi-scale optimization, and fast deep learning models [2,18]. Even for studies that may benefit from using LDDMM, the greater availability, open-source developer mind share, and documentation of non-LDDMM methods may be more appealing than the benefits of formal shape analysis. Furthermore, deep learning has provided the community with a new set of image registration tools that are often very fast which is an appealing property for practical applications [3,6]. However, when test data does not match training data these methods can perform poorly with little ability to change their behavior other than more training, which may require a substantial investment. Deep learning methods also rarely come with theoretical guarantees or modeling properties such as those provided by LDDMM. This paper proposes a method to connect the theoretical benefits of LDDMM to the practical benefits of non-LDDMM image registration including deep learning.

Large Deformation Diffeomorphic Metric Embedding (LDDMEm) takes any displacement vector field and solves for a nearest approximation within the LDDMM framework. This capability effectively decouples image registration of noisy data from shape analysis. Images can be registered with any method. Later, shape analysis can be added to the study design without repeating the comparatively costly and challenging to parameterize image registrations. The effects of shape analysis can also be more fairly evaluated, as rather than requiring a different method entirely, they are derived from the existing study deformations. After a brief background, the theoretical framework for LDDMEm is developed. Synthetic and real data experiments are presented. Finally, open source LDDMEm code is described.

2 Background: LDDMM

LDDMM has been thoroughly described in the literature [7,12,13]. The most relevant background materials for the present work are Vialard et al. [17], which describes LDDMM from a Hamiltonian physics perspective and formulates an optimal control algorithm on the initial conditions of the dynamical system, and Singh et al. [16], which equivalently describes LDDMM from a Lie group perspective and relaxes the Vialard shooting algorithm to optimize vector instead of scalar valued initial momenta. LDDMM models the deformation of image $I \in L^2(\Omega, \mathbb{R})$ onto image J as a time dependent flow of diffeomorphisms. The following augmented Hamiltonian describes the entire system, and we will go

through it term by term:

$$\widetilde{\mathcal{E}} = \frac{1}{2}||v_0||_V^2 + \frac{1}{2\sigma^2}||I(\phi_{1,0}) - J||_{L^2}^2 + \int_0^1 \langle \hat{v}, v - Km \rangle_{L^2} dt + \int_0^1 \langle \hat{m}, \dot{m} + \mathrm{ad}_v^* m \rangle_{L^2} dt + \int_0^1 \langle \hat{I}, \dot{I} + \nabla I \cdot v \rangle_{L^2} dt \tag{1}$$

It is appropriate to think of the first two terms as kinetic and potential energy thus comprising the Hamiltonian itself. The kinetic energy is the norm of velocity vector field v_0 in the space V, which is defined by its inner product: $||v_0||_V = \langle v_0, Lv_0 \rangle_{L^2} = \langle v_0, m_0 \rangle_{L^2} = \langle Km_0, m_0 \rangle_{L^2}$. Here, $m_0 = Lv_0$ is the momentum corresponding to velocity v_0. We have also used the common notation convention: $L^{-1} = K$. The operator L describes the assumed continuum mechanical properties of the spatial domain. It can be derived from real physical properties (e.g., elasticity or viscosity) or assumed, and typically has the form $L = (a\nabla^2 + b\nabla(\nabla\cdot) + c)^d$; the only requirement is that it be self-adjoint. Intuitively, L penalizes the derivatives of v_0, meaning smoother fields have smaller magnitudes.

The potential energy term quantifies the disagreement between the image I warped by a transform $\phi_{1,0}$, and the image J. Here, we write the sum of squared differences, but any differentiable measure is acceptable. The transform $\phi_{1,0}$ arises from a time dependent velocity flow, the integral of which establishes particle position relationships both forward and backward in time:

$$(a)\ \frac{d\phi_{0,t}}{dt} = v(\phi_{0,t}) \qquad (b)\ \frac{d\phi_{t,0}}{dt} = -\nabla\phi_{t,0} \cdot v \tag{2}$$

The image I is said to be observed at time $t = 0$. Its spatial domain flows according to $v(t)$. Consequently, the appearance of I at time $t = 1$, at the spatial positions upon which J is sampled, is given by $\phi_{1,0}$, which maps positions from time $t = 1$ to $t = 0$. The Hamiltonian terms are what we wish to optimize. We would like the velocity flow $\tilde{v}(t)$ that yields the transform $\tilde{\phi}_{1,0}$ such that the sum of kinetic and potential energy terms is minimal. Soon, we will constrain the problem such that only the initial velocity $\tilde{v}_0$ must be found.

The terms on the second line of Eq. (1) are augmentations that enforce constraints placed upon the system. The Lagrange multipliers $\hat{v}, \hat{m}$, and $\hat{I}$ are unconstrained. To recover the Hamiltonian, the right hand side of each inner product must be zero at all points in space and time, i.e., those equations must be satisfied. The first constraint is $v = Km$, which is always true by definition, but including this term simplifies later calculations so it is added for convenience. The second constraint, $\partial_t m = -\mathrm{ad}_v^* m$, is the most interesting and is a defining characteristic of LDDMM. Suppose the transform $\tilde{\phi}_{1,0}$ optimally matches image I to image J. We would like the arc length $\int_0^1 ||\tilde{v}(t)||_V dt$, which via equation (2b) terminates at the point $\tilde{\phi}_{1,0}$, to be minimal. This is precisely the definition of a geodesic, and a geodesic is uniquely determined by its tangent vector at $t = 0$. Thus, we are only required to specify $\tilde{v}_0$, so long as we constrain $\tilde{v}(t)$ to evolve without extrinsic acceleration. Naively, one might assume that the correct

constraint would be $\partial_t v = 0$. However, recall that the flow is assumed to be in a continuum with mechanical properties determined by L. As particles displace, intrinsic strain or viscous forces develop and the particles accelerate. Such forces are proportional to the spatial derivatives of the displacement and the velocity. The operator is thusly defined $\mathrm{ad}^*_v m = (\nabla v)^T m + (\nabla m)v + (\nabla \cdot v)m$, and recall $m = Lv$. Intuitively, we can see that $\partial_t m = -\mathrm{ad}^*_v m$ accounts for such strain or viscous forces through the spatial derivatives of v and m. More formally, the definition of $\mathrm{ad}^*_v m$ is a special case of the Euler-Poincaré equation for the Lie group of diffeomorphisms [1], the derivation of which is beyond the scope of this paper. The third constraint is $\partial_t I = -\nabla I \cdot v$. This is a consequence of Eq. (2b) and describes how the given image I flows along the geodesic $\phi_{t,0}$.

In summary, the augmented Hamiltonian equation (1) describes the energy of a dynamical system parameterized by an initial velocity v_0. The system is constrained to evolve along a geodesic path ending at the transform $\phi_{1,0}$. The stationary points of this functional optimally balance the magnitude of v_0 (with respect to L) with the image matching accuracy of $\phi_{1,0}$. Given I, J, and choosing values (a, b, c, d, σ^2), which determine L and the balance between kinetic and potential energies, we require an algorithm to determine $\tilde{v}_0$, or equivalently $\tilde{m}_0$, which minimizes (1). A gradient descent algorithm can be obtained by taking the variation of (1) with respect to the flows v, I, $\hat{v}$, $\hat{m}$, and $\hat{I}$ with the following result:

$$\partial_{m_0} \tilde{\mathcal{E}} = K m_0 - \hat{m}_0 \tag{3}$$

$$\begin{cases} \dot{\hat{m}} = (\nabla v)\hat{m} - (\nabla \hat{m})v + \hat{v} & \hat{m}_1 = 0 \\ \dot{\hat{I}} = -\nabla \cdot (\hat{I} v) & \hat{I}_1 = I(\phi_{1,0}) - J \\ L\hat{v} = \hat{I}\nabla I - \mathrm{ad}^*_{\hat{m}} m \end{cases} \tag{4}$$

Equation (3) requires $\hat{m}_0$ which can be obtained by integrating the system (4) backward in time.

3 Methods: LDDMEm

LDDMEm replaces image matching with displacement vector field matching. The potential energy term $||I(\phi_{1,0}) - J||^2_{L^2}$ is replaced with a similar term acting directly on vector fields: $||\phi_{0,1} - \phi^*||^2_{L^2}$ for the given field ϕ^*. Also, the third Lagrangian constraint $\int_0^1 \langle \hat{I}, \dot{I} + \nabla I \cdot v \rangle_{L^2} dt$ is replaced with the construction of the flow of transforms itself, namely equation (2a): $\int_0^1 \langle \hat{\phi}, \dot{\phi} - v(\phi) \rangle_{L^2} dt$. This results in the augmented Hamiltonian:

$$\begin{aligned} \widetilde{\mathcal{E}} \;&=\; \frac{1}{2}||v_0||^2_V \;+\; \frac{1}{2\sigma^2}||\phi_{0,1} - \phi^*||^2_{L^2} \\ &+ \int_0^1 \langle \hat{v}, v - Km \rangle_{L^2} dt + \int_0^1 \langle \hat{m}, \dot{m} + \mathrm{ad}^*_v m \rangle_{L^2} dt \;+\; \int_0^1 \langle \hat{\phi}, \dot{\phi} - v(\phi) \rangle_{L^2} dt \end{aligned} \tag{5}$$

With these changes we proceed taking variations as before to obtain a gradient descent algorithm with the following result:

$$\partial_{m_0}\tilde{\mathcal{E}} = Km_0 - \hat{m}_0 \tag{6}$$

$$\begin{cases} \dot{\hat{m}} = (\nabla v)\hat{m} - (\nabla \hat{m})v + \hat{v} & \hat{m}_1 = 0 \\ \dot{\hat{\phi}} = -(\nabla v)^T \hat{\phi} & \hat{\phi}_1 = \phi_{0,1} - \phi^* \\ L\hat{v} = -|\nabla \phi_{t,0}| \cdot \hat{\phi}(\phi_{t,0}) - \mathrm{ad}^*_{\hat{m}} m \end{cases} \tag{7}$$

Comparing (7) to (4) we see the differences. The adjoint variable $\hat{I}$ is replaced with $\hat{\phi}$ whose backward integration is an advection along v. In the $\hat{v}$ equation, $\hat{I}$ is replaced with $\hat{\phi}$ pulled from $t = 0$ to the current time. This pull of $\hat{\phi}$ is a reflection of the choice to model ϕ^* as $\phi_{0,1}$ instead of $\phi_{1,0}$. Since LDDMM entails the construction of both the forward and inverse transform flows, but we are given only one target transform, we are free to choose between the forward and inverse transform to compare to the given field. We chose $\phi_{0,1}$ because equation (2a) is more numerically stable than Eq. (2b). This system is optimized in the same way as the image matching case returning an initial momentum $\tilde{m}_0$ which determines a geodesic of transforms that passes through ϕ^*. With this embedding, the geodesic trajectory implied by ϕ^* can be interpolated, extrapolated, or averaged with other geodesics to study the deformation(s) in a formal shape space.

One interesting property of this formulation is that, unlike the noisy image matching case, the use of sum of squared differences to compare $\phi_{0,1}$ and ϕ^* is not only the simplest option but also sufficiently general since by definition the values of $\phi_{0,1}$ and ϕ^* are directly comparable. Another interesting property is the comparison to point matching formulations [11]. If ϕ^* is viewed as a dense set of coordinate matches, then LDDMEm is the generalization of point match based LDDMM methods to an infinite number of points, or more practically, a point match for every sample location.

4 Experiments

4.1 Synthetic

Two-dimensional synthetic experiments were designed to validate the ability of LDDMEm to embed displacement vector fields from multiple sources. Additionally, these experiments demonstrate one potential advantage of formal shape modeling: extrapolation accuracy. We begin with a smoothed image of an ellipse. A synthetic initial momentum field was generated and integrated to produce a geodesic of shape deformation sampled at time points $t = \{0, 1, 2, 16\}$. The shape at $t = 0$ (the ellipse) was registered as the moving image to the shape at $t = 2$ using three different commonly used packages: simpleITK [18], ANTs [2], and the pre-trained VoxelMorph SynthMorph shapes model [9]. These transforms were naively interpolated/extrapolated to the unseen time points $t = \{1, 16\}$ by multiplying the whole field by the appropriate scaling factor. The transforms

were also embedded in the LDDMM framework with LDDMEm yielding initial momenta that integrate to close approximations of the given transforms at $t = 2$. These geodesics were integrated to the unseen time points $t = \{1, 16\}$ and compared to the naive interpolation/extrapolation results from scaling the given transforms directly. The results are presented in Fig. 1.

All three registration methods aligned the ellipse to the $t = 2$ shape accurately, with only a slight residual remaining in the VoxelMorph case. LDDMEm embedded all transforms accurately, reproducing them at sub-voxel accuracy. For interpolation to $t = 1$, both the naive scaling and LDDMM geodesic integration methods were accurate. However, for extrapolation to $t = 16$, integration of the initial momentum produced a closer shape match than naive scaling in all three cases.

4.2 ADNI 2

To test LDDMEm on real samples we used a set of 350 subjects from the Alzheimer's Disease Neuroimaging Initiative phase 2 dataset (ADNI-2) [4]. For each subject we have four scans: screening, and 6-month, 12-month, and 24-month follow up times. The same three registration methods were used, however for these experiments we switched to the dense-brain-T1-3d-mse-32feat pretrained VoxelMorph model, which was specifically trained for registration of T1 MRI volumes like the ADNI-2 data we study here. For all samples, and all registration methods, we registered the screening scan as the moving image to the 12-month follow up scan. These transforms were naively interpolated/extrapolated by multiplying by the appropriate scaling factor to estimate the anatomy at the 6-month and 24-month time points. The transforms were also embedded using LDDMEm to obtain initial momenta parameterizing geodesic flows of transforms. These geodesics were integrated to estimate the anatomy at the 6-month and 24-month time points. The results are presented in Fig. 2.

Curiously, for all three registration methods, the LDDMEm embedded transform matched the screening image to the 12-month image more accurately than the given transform, despite having never seen either image. We hypothesize this to be the effect of additional regularization, but have not yet determined an exact reason. The extrapolation effect observed in the synthetic experiments is only present for the ANTs transforms, albeit with a very small effect size. It may be that the time intervals used are too short to observe the improved extrapolation of geodesic modeling. We are interested to see how a registration to the 3-month image extrapolates with LDDMEm to longer time points such as 36-months. These experiments are ongoing and are not presented here.

5 Software

LDDMEm is available as an open source Python package at: github.com/GFleishman/lddmem. At the time of writing, the package contains three modules: `main`, `epdiff`, and `cli`. The `main` module contains the

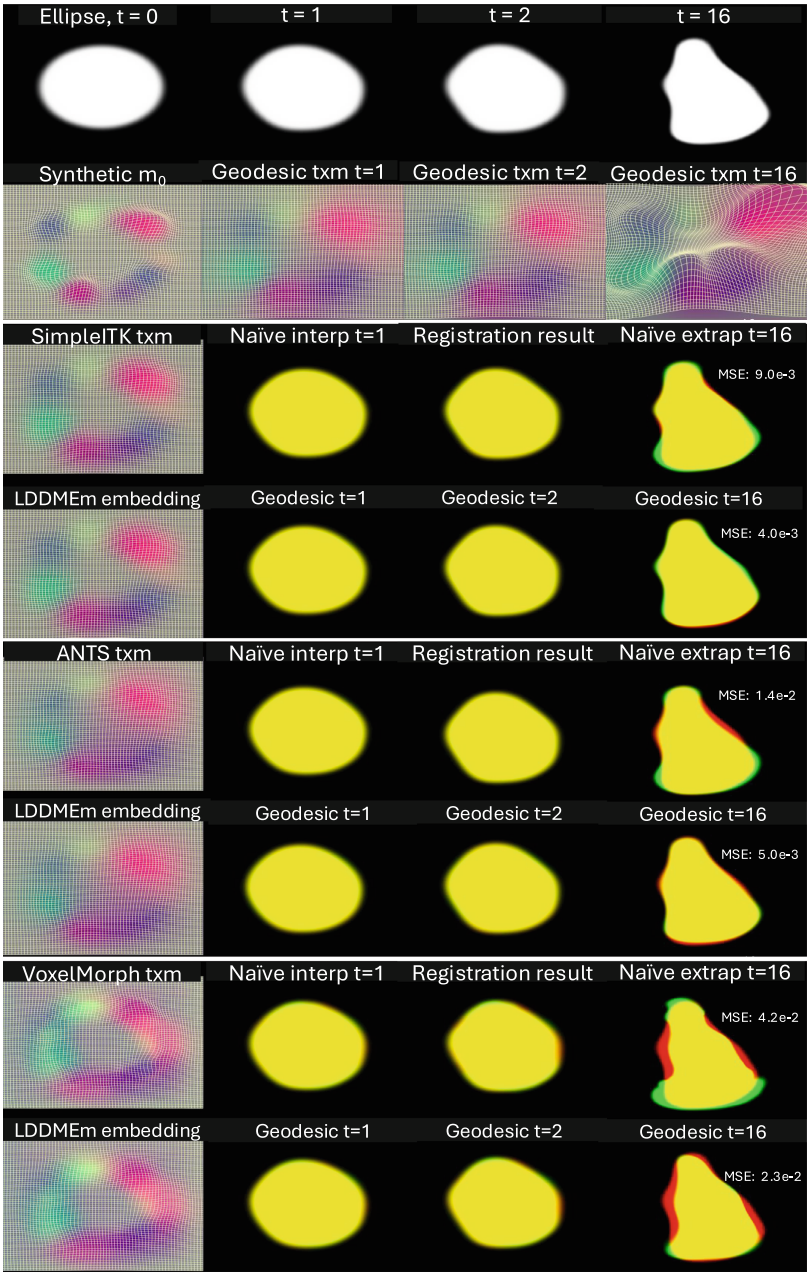


Fig. 1. Synthetic experiment results: LDDMEm successfully embeds transforms from three commonly used image registration packages and extrapolates shape change with greater accuracy. Target images in red, aligned images in green. (Color figure online)

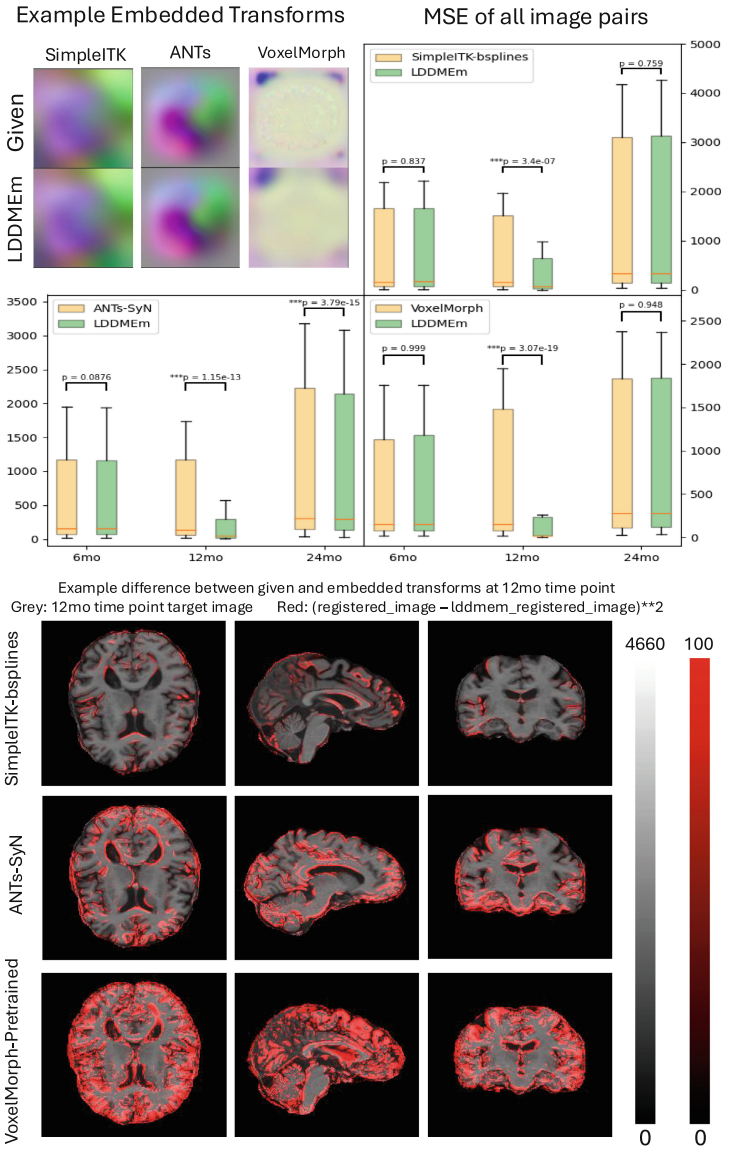


Fig. 2. ADNI2 experiment results: LDDMEm successfully embedded transforms generated from real data (upper left panel). Curiously, for all three registration methods, the LDDMEm embedded transforms align the original two images more accurately, with the largest differences observed for VoxelMorph (perhaps due to increased regularization, which can be seen visually in the upper left panel). The improved geodesic extrapolation observed for the synthetic data is only observed here for ANTs, albeit with a very small effect size.

function `lddmem` which is the primary method for embedding a transform. The function supports multi-resolution optimization in both space and time, masking to ignore regions of fields that should not be matched, automatic convergence criteria, and other helpful features. The `main` module also contains the function `forward_integration` which can be used to integrate an initial momentum to any time point. The `epdiff` module contains implementations of essential operators including L, K, $\mathrm{ad}_v m$, and $\mathrm{ad}^*_v m$. Finally, the `cli` module is a command line interface for calling `main.lddmem`. Future plans include offering Simple Geodesic Regression [10] and Principal Geodesic Analysis [8].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arnold, V.: Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. Annales de l'Inst. Fourier **16**(1), 319–361 (1966). https://doi.org/10.5802/aif.233
2. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. **12**(1), 26–41 (2008). https://doi.org/10.1016/j.media.2007.06.004
3. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Dalca, A.V., Guttag, J.: An unsupervised learning model for deformable medical image registration. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9252–9260 (2018). https://doi.org/10.1109/CVPR.2018.00964, http://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00964
4. Beckett, L.A., Donohue, M.C., Wang, C., Aisen, P., Harvey, D.J., Saito, N.: The Alzheimer's disease neuroimaging initiative phase 2: increasing the length, breadth, and depth of our understanding. Alzheimers Dement. **11**(7), 823–831 (2015). https://doi.org/10.1016/j.jalz.2015.05.004
5. Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int. J. Comput. Vision **61**(2), 139–157 (2005). https://doi.org/10.1023/B:VISI.0000043755.93987.aa
6. Ding, Z., Fleishman, G., Yang, X., Thompson, P., Kwitt, R., Niethammer, M.: Fast predictive simple geodesic regression. Med. Image Anal. **56**, 193–209 (2019). https://doi.org/10.1016/j.media.2019.06.003
7. Dupuis, P., Grenander, U., Miller, M.: Variational problems on flows of diffeomorphisms for image matching. Q. Appl. Math. **56**(3), 587–600 (1998). https://doi.org/10.1090/qam/1632326
8. Fletcher, P., Lu, C., Pizer, S., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imaging **23**(8), 995–1005 (2004). https://doi.org/10.1109/TMI.2004.831793
9. Hoffmann, M., Billot, B., Greve, D.N., Iglesias, J.E., Fischl, B., Dalca, A.V.: Synthmorph: learning contrast-invariant registration without acquired images. IEEE Trans. Med. Imaging **41**(3), 543–558 (2022). https://doi.org/10.1109/TMI.2021.3116879

10. Hong, Y., Shi, Y., Styner, M., Sanchez, M., Niethammer, M.: Simple geodesic regression for image time-series. In: Dawant, B.M., Christensen, G.E., Fitzpatrick, J.M., Rueckert, D. (eds.) WBIR 2012. LNCS, vol. 7359, pp. 11–20. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31340-0_2
11. Joshi, S., Miller, M.: Landmark matching via large deformation diffeomorphisms. IEEE Trans. Image Process. **9**(8), 1357–1370 (2000). https://doi.org/10.1109/83.855431
12. Miller, M.I., Trouve, A., Younes, L.: On the metrics and Euler-Lagrange equations of computational anatomy. Annu. Rev. Biomed. Eng. **4**, 375–405 (2002). https://doi.org/10.1146/annurev.bioeng.4.092101.125733
13. Miller, M.I., Trouvé, A., Younes, L.: Geodesic shooting for computational anatomy. J. Math. Imaging Vis. **24**(2), 209–228 (2006). https://doi.org/10.1007/s10851-005-3624-0
14. Niethammer, M., Huang, Y., Vialard, F.-X.: Geodesic regression for image time-series. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011. LNCS, vol. 6892, pp. 655–662. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23629-7_80
15. Shen, Z., Vialard, F.X., Niethammer, M.: Region-specific diffeomorphic metric mapping. Adv. Neural. Inf. Process. Syst. **32**, 1098–1108 (2019)
16. Singh, N., Hinkle, J., Joshi, S., Fletcher, P.T.: A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction. In: 2013 IEEE 10th International Symposium on Biomedical Imaging, pp. 1219–1222 (2013). https://doi.org/10.1109/ISBI.2013.6556700
17. Vialard, F.X., Risser, L., Rueckert, D., Cotter, C.J.: Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation. Int. J. Comput. Vision **97**(2), 229–241 (2012). https://doi.org/10.1007/s11263-011-0481-8
18. Yaniv, Z., Lowekamp, B.C., Johnson, H.J., Beare, R.: SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. J. Digit. Imaging **31**(3), 290–303 (2018). https://doi.org/10.1007/s10278-017-0037-8

MDAA-Diff: CT-Guided Multi-dose Adaptive Attention Diffusion Model for PET Denoising

Xiaolong Niu[1,4,5], Zanting Ye[1,4,5], Xu Han[2], Yanchao Huang[3], Hao Sun[1,4,5], Hubing Wu[3], and Lijun Lu[1,4,5,6(✉)]

[1] School of Biomedical Engineering, Southern Medical University, Guangzhou, China
[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] Nanfang PET Center, Nanfang Hospital Southern Medical University, Guangzhou, China
[4] Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China
[5] Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou, China
[6] Pazhou Lab, Guangzhou, China
ljlubme@gmail.com

Abstract. Acquiring high-quality Positron Emission Tomography (PET) images requires administering high-dose radiotracers, which increases radiation exposure risks. Generating standard-dose PET (SPET) from low-dose PET (LPET) has become a potential solution. However, previous studies have primarily focused on single low-dose PET denoising, neglecting two critical factors: discrepancies in dose response caused by inter-patient variability, and complementary anatomical constraints derived from CT images. In this work, we propose a novel CT-Guided Multi-dose Adaptive Attention Denoising Diffusion Model (MDAA-Diff) for multi-dose PET denoising. Our approach integrates anatomical guidance and dose-level adaptation to achieve superior denoising performance under low-dose conditions. Specifically, this approach incorporates a CT-Guided High-frequency Wavelet Attention (HWA) module, which uses wavelet transforms to separate high-frequency anatomical boundary features from CT images. These extracted features are then incorporated into PET imaging through an adaptive weighted fusion mechanism to enhance edge details. Additionally, we propose the Dose-Adaptive Attention (DAA) module, a dose-conditioned enhancement mechanism that dynamically integrates dose levels into channel-spatial attention weight calculation. Extensive experiments on ^{18}F-FDG and ^{68}Ga-FAPI datasets demonstrate that MDAA-Diff outperforms state-of-the-art approaches in preserving diagnostic quality under reduced-dose conditions. Our code is publicly available at https://github.com/Long0121/MDAA-Diff.

X. Niu, Z. Ye and X. Han—Co-first authors.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 333–343, 2026.
https://doi.org/10.1007/978-3-032-04947-6_32

Keywords: PET denoising · Multi-dose adaptive · High-frequency CT-guided · Diffusion model

1 Introduction

Positron Emission Tomography (PET) is a molecular imaging technique used to evaluate metabolic activity and functional states of human tissues and organs. PET facilitates precise functional assessments of tissues and organs through tracer-based metabolic visualization imaging, demonstrating indispensable clinical significance in tumor diagnosis, neurodegenerative disease research, and cardiovascular functional analysis [2,9,29]. However, PET imaging quality remains highly dependent on radiotracer dose. While low-dose PET (LPET) imaging mitigates the risk of ionizing radiation (particularly for sensitive populations such as children and pregnant women [3,5,12]), it inevitably increases image noise, compromising diagnostic accuracy. This inherent trade-off between radiation reduction and diagnostic efficacy remains a major challenge in contemporary PET imaging technology.

Deep learning has emerged as a pivotal approach for radiation dose reduction in PET imaging, enabling diagnostic-quality standard-dose PET (SPET) denoising from LPET. Compared with traditional methods, such as local mean filtering and frequency-domain filtering [1,4,24], which often oversmooth critical details, deep neural networks learn complex nonlinear mappings between LPET and SPET through end-to-end training [7,10,20,22,25–27], suppressing noise while preserving lesion metabolic activity features. Generative model-based denoising architectures have become a focal point. For instance, Sanaat et al. [17] utilized CycleGAN to achieve cross-domain mapping from LPET to SPET. Zhou et al. [28] combined StyleGAN with segmentation network to generate denoised images with realistic textures. Nevertheless, these deterministic mapping models struggle to quantify uncertainties inherent in the reconstruction process. Denoising Diffusion Probabilistic Model (DDPM) [11] addresses this limitation by iteratively denoising images through a Markov chain, enabling uncertainty modeling while retaining fine-grained details. However, existing approaches predominantly assume single LPET denoising, ignoring the variations in dose response caused by inter-patient variability in clinical practice, such as physiology, metabolism, and tracer uptake patterns.

Preliminary progress has been achieved in research for multi-dose PET denoising. Xie et al. [23] proposed a Unified Noise-aware Network that performs PET denoising across different dose levels. But its weight allocation mechanism compromises recovery performance under high-dose conditions. Tang et al. [19] developed a High-Frequency-Guided Residual Diffusion Model which uses sinusoidal positional encoding to encode the dose, and then input it into the denoising network along with the embedded timesteps. However, the simple concatenation of dose embedding and time embedding may fail to enable effective interaction between dose information and the diffusion process, potentially limiting the model's adaptability to feature variations across different dose levels. More-

over, this method inadequately exploits the advantages of multimodal imaging in clinical applications. In multimodal PET/CT imaging, the high spatial coherence between CT and PET images provides critical anatomical constraints that enhance denoising performance [8]. Chen et al. [6] concatenate multimodal images as input, ignoring varying modalities contribute differently across image locations. Wang et al. [21] proposed a local adaptive fusion method with position-specific weights, which operates only at the input level and does not exploit deep semantic features, potentially limiting the effectiveness of the multimodal fusion.

In this work, we propose a CT-Guided Multi-Dose Adaptive Attention Diffusion Model (MDAA-Diff) for PET denoising. Specifically, we design a CT-Guided High-frequency Wavelet Attention (HWA) module to extract CT high-frequency anatomical boundary features using multi-scale wavelet decomposition, which are then adaptively fused with PET imaging. Moreover, we construct a Dose-adaptive Attention (DAA) module that utilizes dose factors to guide channel-wise and spatial feature weighting, enhancing the model's perception capability and robustness under different dose conditions.

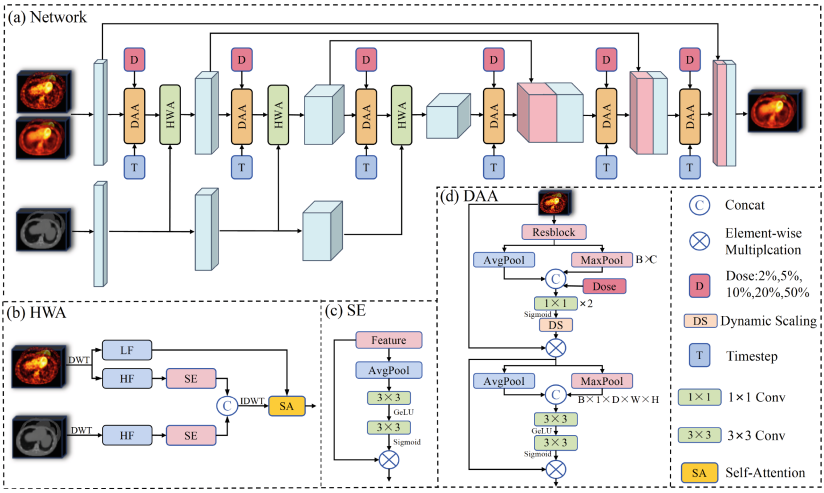


Fig. 1. Overview of the MDAA-Diff. (a) is the overall network architecture of the diffusion model; (b) is a CT-guided modal fusion module; (c) is designed to perform channel-wise attention; and (d) is a multi-dose adaptive attention module.

2 Methodology

2.1 Improved Denoising Diffusion Probabilistic Models

Our model builds upon an Improved Denoising Diffusion Probabilistic Models (IDDPM) [15], which significantly reduces the sampling timesteps compared to

DDPM. While DDPM models the reverse denoising process as a conditional probability distribution, defined as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\, \mu_\theta(x_t, t),\, \Sigma_\theta(x_t, t)), \tag{1}$$

where $\mu_\theta(x_t, t)$ represents the learnable value, and the variance $\Sigma_\theta(x_t, t)$ is fixed as $\sigma^2 I$ and cannot be optimized as a learnable parameter. IDDPM addresses this limitation by parameterizing $\Sigma_\theta(x_t, t)$ modeling through the following formulation:

$$\Sigma_\theta(x_t, t) = \exp\left(v \log \beta_t + (1 - v) \log \tilde{\beta}_t\right), \tag{2}$$

where v is a weighting parameter, β_t represents the noise intensity during the forward diffusion process, and $\tilde{\beta}_t$ denotes the adjusted noise intensity in the reverse process. This parameterization enhances both the model's flexibility and the quality of the generated samples. Additionally, it optimizes noise scheduling by selecting specific timesteps $S = \{s_1, s_2, \ldots, s_K\}$, which further improves sampling efficiency.

2.2 MDAA-Diff Model

Our MDAA-Diff model is shown in Fig. 1. First, the noise-perturbed SPET and LPET images are concatenated along the channel dimension and fed into the input model. Multi-scale PET/CT features are extracted by a dedicated encoder to achieve cross-modal information fusion. The dose factor is dynamically integrated into PET feature enhancement through the DAA module, enabling dose-adaptive adjustments.

High-Frequency Wavelet Attention. Current PET and CT feature fusion methods often rely on channel concatenation or Cross-Attention (CA) mechanisms [14]. The quadratic-complexity operation of CA imposes prohibitive computational demands for high-resolution medical imaging. To address this issue, we propose the HWA module, a fusion module based on 3D wavelet transform (with bior4.4 wavelet) and Squeeze-and-Excitation (SE), to achieve multi-scale cooperative enhancement of cross-modal features. As shown in Fig. 1b, PET and CT features are decomposed into low-frequency and high-frequency components: PET low-frequency features capture global metabolic patterns, while high-frequency features encode lesion edges. CT high-frequency components features capture fine anatomical details, including bone boundaries and organ morphology, which provide accurate spatial localization and structural constraints for PET denoising.

On this basis, the SE module (as shown in Fig. 1c) is utilized to perform dynamic channel weights for both PET and CT high-frequency components, emphasizing key features and reducing redundancy. The weighted PET and CT high-frequency features are concatenated along the channel dimension and processed through convolutional layers for cross-modal interaction and dimensionality reduction. To enable multi-scale information complementarity, HWA modules are embedded across all encoder levels (from shallow local details to deep

semantic features) to establish a progressive fusion architecture. Finally, the fused features are reconstructed from the PET low-frequency components and the convolved high-frequency components through inverse wavelet transform.

Dose-Adaptive Attention Module. The DAA module adaptively enhances multi-dose features through the synergistic mechanism of the channel enhancement branch and the spatial enhancement branch. The details of the HWA module are shown in Fig. 1d. The module dynamically adjusts the weights of the input features to accommodate different dose-level distributions and improves the robustness of the model.

The Channel Feature Enhancement Branch. The branch primarily focuses on dynamically adjusting the channel weights of the input features by capturing global and local information, while integrating the dose factor D to achieve adaptive optimization. Specifically, D is projected through a multi-layer perceptron (MLP) to generate the dose embedding. Global average pooling and max pooling operations are applied to the input features, and the results are concatenated with the dose embedding. The channel weights are then generated through a convolutional network, defined as:

$$W_{\text{channel}} = \text{sigmoid}\Big(W_2 \cdot \text{ReLU}\big(W_1 \cdot \text{Concat}(\text{Pool}(X), D_{\text{emb}})\big)\Big) \tag{3}$$

where $\text{Pool}(X) \in \{\text{AvgPool}(X), \text{MaxPool}(X)\}$, W_1 and W_2 are 1×1 convolutional weight matrices. The channel weights are further scaled by the factor α and shifted by the bias factor β, which are used to normalize and adjust the input features, achieving dynamic channel enhancement, defined as:

$$X' = \text{Norm}(X) \odot (1 + \alpha) + \beta \tag{4}$$

where $\text{Norm}(X)$ represents a normalization operation, and $\odot$ denotes element-wise multiplication.

The Spatial Feature Enhancement Branch. The branch aims to extract the structural detail and local texture information of the input features along the spatial dimension. Specifically, the input features are pooled along the channel dimension using global average pooling and max pooling. The results are concatenated and passed through a convolutional layer to generate the spatial attention map, defined as:

$$X' = X \odot \text{sigmoid}\left(\text{Conv}\left(\text{Concat}(\text{Mean}(X), \text{Max}(X))\right)\right), \tag{5}$$

where Conv is a convolutional block composed of a 3×3 kernel and a dilated convolution (dilation rate set to 2). Finally, the spatial attention map is multiplied element-wise with the input features to achieve context enhancement along the spatial dimension.

3 Experiments

3.1 Datasets

We evaluated the performance of MDAA-Diff using two independent PET/CT datasets comprising 51 patients with ^{18}F-FDG tracer and 60 patients with ^{68}Ga-FAPI tracer. The datasets were provided by the Nanfang PET Center, Nanfang Hospital, Southern Medical University. SPET images were reconstructed from 300 s (100%) acquisitions obtained 60 min post-injection. LPET images were reconstructed based on time windows resampled during acquisition, simulating reduced acquisition durations of 6 s (2%), 15 s (5%), 30 s (10%), 60 s (20%), and 150 s (50%). All PET images were reconstructed with resolution of $192 \times 192 \times 673$. To address PET-CT mismatches, CT images were resampled and registered to the PET coordinate system using mutual information optimization.

3.2 Implementation Details

Our MDAA-Diff model was implemented in PyTorch and trained on an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. A five-fold cross-validation strategy was employed in the experimental design. The number of diffusion steps during the training phase was set to 1000, and the number of sampling steps during the inference phase was set to 50. We used an AdamW optimizer with an initial learning rate of 1e−4, utilizing the cosine annealing scheduling strategy. The generated image quality was evaluated using the Peak Signal-to-Noise Ratio

Table 1. Quantitative comparison of MDAA-Diff with five state-of-the-art methods. PSNR and SSIM are reported for both tracers at different dose levels.

^{18}F-FDG					
PSNR↑/SSIM↑	6 s	15 s	30 s	60 s	150 s
CNN [18]	39.858/0.969	40.082/0.972	40.269/0.974	40.459/0.977	40.614/0.978
SwinUnetr [16]	44.408/0.976	44.979/0.979	45.417/0.981	46.059/0.983	47.115/0.986
MambaMIR [13]	44.349/0.978	45.801/0.982	45.830/0.983	46.043/0.985	46.774/0.987
UNN [23]	44.504/0.972	45.185/0.977	45.663/0.980	46.071/0.982	46.646/0.984
HF-ResDiff [19]	44.728/0.979	45.243/0.982	45.677/0.984	46.174/0.986	47.004/0.984
Ours	**46.035/0.984**	**46.523/0.987**	**46.889/0.988**	**47.315/0.989**	**48.042/0.990**
^{68}Ga-FAPI					
PSNR↑/SSIM↑	6 s	15 s	30 s	60 s	150 s
CNN [18]	39.570/0.967	39.832/0.961	40.049/0.963	40.276/0.965	40.680/0.968
SwinUnetr [16]	44.119/0.966	44.560/0.969	44.866/0.971	45.171/0.973	45.783/0.976
MambaMIR [13]	44.028/0.968	44.429/0.970	44.810/0.972	45.207/0.974	45.844/0.977
UNN [23]	44.759/0.969	45.019/0.973	45.432/0.975	45.765/0.977	46.496/0.980
HF-ResDiff [19]	44.219/0.964	44.632/0.968	45.081/0.971	45.493/0.974	46.287/0.977
Ours	**46.689/0.980**	**47.009/0.981**	**47.160/0.981**	**47.362/0.982**	**47.921/0.983**

(PSNR) and Structural Similarity Index (SSIM) metrics. Visual comparisons were conducted to demonstrate the model's ability to improve anatomical fidelity and effectively suppress noise in critical regions.

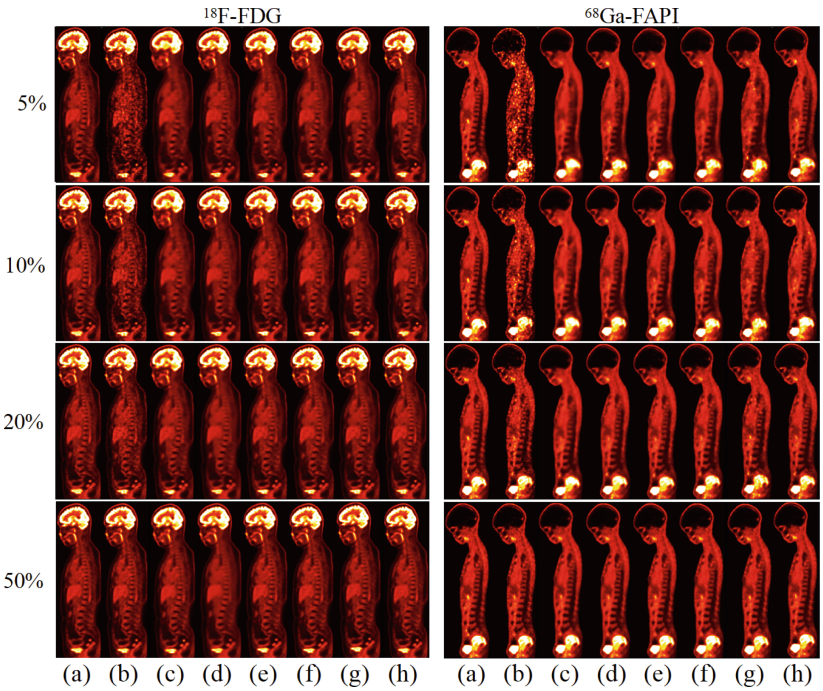


Fig. 2. The comparison results of MDAA-Diff with several state-of-the-art methods at different dose levels. (a): SPET; (b): LPET; (c): CNN; (d): UNN; (e): SwinUnetr; (f): MambaMIR; (g): HF-ResDiff; (h): Ours.

3.3 Comparison with State-of-the-Art Methods

We compared MDAA-Diff with five state-of-the-art methods, including single-dose approaches (CNN [18], SwinUnetr [16], MambaMIR [13]) and multi-dose approaches (HF-ResDiff [19], UNN [23]). All comparison methods take concatenated PET and CT images as input. Quantitative analysis (as shown in Table 1) demonstrated that MDAA-Diff outperformed all comparison models across both PSNR and SSIM metrics. Moreover, its performance improved significantly at lower dose levels (e.g., 2% and 5% doses). For instance, in 2% dose imaging for the ^{18}F-FDG tracer, MDAA-Diff achieved an SSIM of 98.4%, which is 0.5% points higher than the second-best model (HF-ResDiff: 97.9%). This finding highlights its robustness to complex noise distributions.

Qualitative analysis further validated the clinical value of the proposed method. As illustrated in Fig. 2, MDAA-Diff exhibited significant improvements in boundary sharpness for the spine region. Particularly in the 2% dose experiments, its structural fidelity was noticeably superior to that of other comparison methods. For denoising tasks of small lesions, MDAA-Diff effectively suppressed noise artifacts (as indicated by the blue arrows in Fig. 3) while preserving the fine texture details of the lesions.

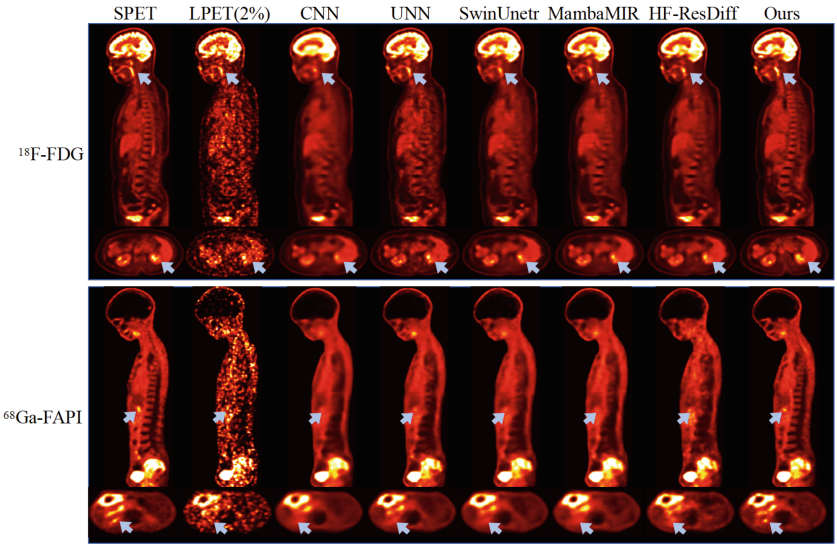


Fig. 3. Quantitative comparison with state-of-the-art methods at 2% dose level.

3.4 Ablation Study

To validate the effectiveness of each proposed module, we conducted ablation studies with three sub-models: (1) IDDPM: the standard diffusion model; (2) IDDPM-HWA: IDDPM integrated with the CT-guided HWA; (3) MDAA-Diff (Ours): IDDPM incorporating both HWA and DAA module. All approaches were kept consistent in the experimental setup, ensuring a fair comparison (as shown in Table 2). The comparison results showed that IDDPM-HWA significantly improved denoising performance over IDDPM by leveraging CT fusion. MDAA-Diff further enhanced denoising capability through multi-dose adaptation and CT guidance, achieving clearer spinal and pulmonary boundaries with enhanced structural fidelity. These findings confirmed that the HWA and DAA module effectively improve PET denoising performance, validating their critical role in modeling multi-dose characteristics and anatomical constraints.

Table 2. Quantitative ablation results on the ^{18}F-FDG and ^{68}Ga-FAPI tracer datasets.

^{18}F-FDG					
PSNR↑/SSIM↑	6 s	15 s	30 s	60 s	150 s
LPET	43.151/0.931	43.997/0.944	44.807/0.954	45.674/0.962	47.096/0.969
IDDPM	44.944/0.978	45.108/0.981	45.623/0.983	46.232/0.985	47.341/0.987
IDDPM+HWA	45.724/0.981	46.068/0.983	46.273/0.985	46.739/0.987	47.805/0.988
Ours	**46.035/0.984**	**46.523/0.987**	**46.889/0.988**	**47.315/0.989**	**48.042/0.990**
^{68}Ga-FAPI					
PSNR↑/SSIM↑	6 s	15 s	30 s	60 s	150 s
LPET	40.443/0.873	41.382/0.892	42.302/0.908	43.224/0.922	44.689/0.935
IDDPM	44.790/0.973	45.363/0.974	45.825/0.976	46.154/0.977	46.796/0.979
IDDPM+HWA	45.992/0.976	46.187/0.978	46.369/0.980	46.648/0.980	47.287/0.982
Ours	**46.689/0.980**	**47.009/0.981**	**47.160/0.981**	**47.362/0.982**	**47.921/0.983**

4 Conclusion

In this study, we propose a CT-Guided Multi-Dose Adaptive Attention Diffusion Model (MDAA-Diff) for LPET denoising. It utilizes high-frequency wavelet transforms from CT images to integrate multimodal information. This approach address detail blurring limitations in traditional single-modal denoising models. Additionally, MDAA-Diff incorporates DAA module to explicitly model dose-dependent relationships, providing a robust and reliable solution for LPET imaging in clinical applications.

Acknowledgments. This research was funded by the National Natural Science Foundation of China (62371221, 12326616 and 62201245), National High end Foreign Experts Recruitment Plan (G2023030025L), the Science and Technology Program of Guangdong Province 2022A0505050039.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arabi, H., Zaidi, H.: Improvement of image quality in PET using post-reconstruction hybrid spatial-frequency domain filtering. Phys. Med. Biol. **63**(21), 215010 (2018)
2. Boellaard, R.: Standards for PET image acquisition and quantitative data analysis. J. Nucl. Med. **50**(Suppl 1), 11S-20S (2009)
3. Catana, C.: The dawn of a new era in low-dose PET imaging (2019)
4. Chan, C., Fulton, R., Barnett, R., Feng, D.D., Meikle, S.: Postreconstruction nonlocal means filtering of whole-body pet with an anatomical prior. IEEE Trans. Med. Imaging **33**(3), 636–650 (2013)

5. Chawla, S.C., et al.: Estimated cumulative radiation dose from PET/CT in children with malignancies: a 5-year retrospective review. Pediatr. Radiol. **40**, 681–686 (2010)
6. Chen, K.T., et al.: Ultra-low-dose 18F-Florbetaben amyloid PET imaging using deep learning with multi-contrast MRI inputs. Radiology **290**(3), 649–656 (2019)
7. Cui, J., Wang, Y., Zhou, L., Fei, Y., Zhou, J., Shen, D.: 3D point-based multi-modal context clusters GAN for low-dose PET image denoising. IEEE Trans. Circuits Syst. Video Technol. (2024)
8. Fallahpoor, M., et al.: Deep learning techniques in PET/CT imaging: a comprehensive review from sinogram to image space. Comput. Methods Programs Biomed. **243**, 107880 (2024)
9. Fletcher, J.W., et al.: Recommendations on the use of 18F-FDG PET in oncology. J. Nucl. Med. **49**(3), 480–508 (2008)
10. Fu, Y., et al.: MPGAN: multi pareto generative adversarial network for the denoising and quantitative analysis of low-dose PET images of human brain. Med. Image Anal. **98**, 103306 (2024)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 6840–6851 (2020)
12. Huang, B., Law, M.W.M., Khong, P.L.: Whole-body PET/CT scanning: estimation of radiation dose and cancer risk. Radiology **251**(1), 166–174 (2009)
13. Huang, J., et al.: MambaMIR: an arbitrary-masked mamba for joint medical image reconstruction and uncertainty estimation. arXiv preprint arXiv:2402.18451 (2024)
14. Jiang, C., et al.: PET-diffusion: unsupervised PET enhancement based on the latent diffusion model. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14220, pp. 3–12. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43907-0_1
15. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR (2021)
16. Sanaat, A., et al.: A deep learning model for generating [18F] FDG PET images from early-phase [18F] florbetapir and [18F] flutemetamol PET images. Eur. J. Nucl. Med. Mol. Imaging **51**(12), 3518–3531 (2024)
17. Sanaat, A., Shiri, I., Arabi, H., Mainta, I., Nkoulou, R., Zaidi, H.: Deep learning-assisted ultra-fast/low-dose whole-body PET/CT imaging. Eur. J. Nucl. Med. Mol. Imaging **48**, 2405–2415 (2021)
18. Schaefferkoetter, J., et al.: Convolutional neural networks for improving image quality with noisy PET data. EJNMMI Res. **10**, 1–11 (2020)
19. Tang, Z., Jiang, C., Cui, Z., Shen, D.: HF-ResDiff: high-frequency-guided residual diffusion for multi-dose PET reconstruction. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15007, pp. 372–381. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72104-5_36
20. Wang, Y., et al.: 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. Neuroimage **174**, 550–562 (2018)
21. Wang, Y., et al.: 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis. IEEE Trans. Med. Imaging **38**(6), 1328–1339 (2018)
22. Xiang, L., et al.: Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI. Neurocomputing **267**, 406–416 (2017)
23. Xie, H., et al.: Unified noise-aware network for low-count PET denoising with varying count levels. IEEE Trans. Radiat. Plasma Med. Sci. **8**(4), 366–378 (2023)
24. Xu, J., Gong, E., Pauly, J., Zaharchuk, G.: 200x low-dose PET reconstruction using deep learning. arXiv preprint arXiv:1712.04119 (2017)

25. Xue, S., et al.: A cross-scanner and cross-tracer deep learning method for the recovery of standard-dose imaging quality from low-dose PET. Eur. J. Nucl. Med. Mol. Imaging 1–14 (2022)
26. Zhou, B., Tsai, Y.J., Chen, X., Duncan, J.S., Liu, C.: MDPET: a unified motion correction and denoising adversarial network for low-dose gated PET. IEEE Trans. Med. Imaging **40**(11), 3154–3164 (2021)
27. Zhou, L., Schaefferkoetter, J.D., Tham, I.W., Huang, G., Yan, J.: Supervised learning with CycleGAN for low-dose FDG PET image denoising. Med. Image Anal. **65**, 101770 (2020)
28. Zhou, Y., Yang, Z., Zhang, H., Chang, E.I.C., Fan, Y., Xu, Y.: 3D segmentation guided style-based generative adversarial networks for PET synthesis. IEEE Trans. Med. Imaging **41**(8), 2092–2104 (2022)
29. Zhuang, H., Codreanu, I.: Growing applications of FDG PET-CT imaging in non-oncologic conditions. J. Biomed. Res. **29**(3), 189 (2015)

Medical-Knowledge Driven Multiple Instance Learning for Classifying Severe Abdominal Anomalies on Prenatal Ultrasound

Huanwen Liang[1], Jingxian Xu[1], Yuanji Zhang[1], Yuhao Huang[1], Yuhan Zhang[1], Xin Yang[1], Ran Li[2], Xuedong Deng[3], Yanjun Liu[4], Guowei Tao[5], Yun Wu[6], Sheng Zhao[7], Xinru Gao[8(✉)], and Dong Ni[1,9,10,11(✉)]

[1] Medical Ultrasound Image Computing (MUSIC) Lab, School of Biomedical Engineering, Medical School Shenzhen University, Shenzhen, China
[2] Shenzhen RayShape Medical Technology Co., Ltd., Shenzhen, China
[3] Center for Medical Ultrasound, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Gusu School, Nanjing Medical University, Suzhou, China
[4] The First Hospital of China Medical University, Shenyang, China
[5] Qilu Hospital of Shandong University, Jinan, China
[6] Nanjing Women and Children's Healthcare Hospital, Women's Hospital of Nanjing Medical University, Nanjing, China
[7] Maternal and Child Health Hospital of Hubei Province, Wuhan, China
[8] Northwest Women's and Children's Hospital, Xian, China
gxr_ca@126.com
[9] School of Artificial Intelligence, Shenzhen University, Shenzhen, China
[10] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China
[11] School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China
nidong@szu.edu.cn

Abstract. Fetal abdominal malformations are serious congenital anomalies that require accurate diagnosis to guide pregnancy management and reduce mortality. Although AI has demonstrated significant potential in medical diagnosis, its application to prenatal abdominal anomalies remains limited. Most existing studies focus on image-level classification and rely on standard plane localization, placing less emphasis on case-level diagnosis. In this paper, we develop a case-level multiple instance learning (MIL)-based method, free of standard plane localization, for classifying fetal abdominal anomalies in prenatal ultrasound. Our contribution is three-fold. First, we adopt a mixture-of-attention-experts module (MoAE) to weight different attention heads for various planes. Secondly, we propose a medical-knowledge-driven feature selection module (MFS) to align image features with medical knowledge, performing self-supervised image token selection at the case-level. Finally, we propose a prompt-based prototype learning (PPL) to enhance the MFS. Extensively validated on a large prenatal abdominal ultrasound dataset containing 2,419 cases, with a total of 24,748 images and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 344–354, 2026.
https://doi.org/10.1007/978-3-032-04947-6_33

6 categories, our proposed method outperforms the state-of-the-art competitors. Codes are available at: https://github.com/LL-AC/AAcls.

Keywords: Prenatal Abdominal Ultrasound · Multi-Instance Learning · Anomalies Classification

1 Introduction

Fetal abdominal malformations encompass a range of significant congenital anomalies that affect fetal health and are often associated with syndromes or aneuploidy [15]. The most severe and common abdominal anomalies include abdominal wall defects (such as omphalocele and gastroschisis), renal agenesis (RA), duodenal atresia (DA), and multicystic dysplastic kidney (MCDK), with an incidence of 1–4 per 10,000 live births [9,18,26]. Accurate diagnosis is critical for pregnancy decisions and treatment plans to reduce fetal morbidity and mortality. Ultrasound (US) remains the standard tool for prenatal abdominal evaluation, used to acquire standard planes and assess anatomical structures for diagnosing abdominal anomalies. However, abdominal anomalies are diverse in type and presentation, often occurring at earlier gestational ages. The detection of these anomalies heavily relies on the expertise and experience of sonographers, leading to varying detection rates and underscoring the ongoing challenges in prenatal diagnosis. To the best of our knowledge, artificial intelligence has demonstrated strong proficiency in diagnosing complex medical conditions [4,14,22]. However, research on its application to the diagnosis of prenatal abdominal anomalies remains limited.

Fetal Anomalies Detection. Currently, most studies for fetal anomalies detection primarily involve: (1) localizing standard planes from unordered ***case-level*** ultrasound images or video sequences and (2) detecting anomalies on the located ***image-level*** standard planes. Xie et al. [25] classified the fetal brain standard planes, and then the lesions in anomaly images can be localized by activation map. Lin et al. [12] developed an intelligent system, using the YOLO3 to detect fetal brain standard planes and brain structure to classify abnormal images. Qi et al. [16] constructed a deep learning model for nervous system anomalies classification in a multi-center fetal brain dataset. Most recently, Ciobanu et al. [2] proposed the first study to categorize fetal abdominal planes, achieving precise subsequent biomarker measurement and anomaly detection.

Although effective, the above studies require standard plane localization, which necessitates fine-grained image-level annotations. Meanwhile, they face challenges in cases such as abdominal wall defects (e.g., gastroschisis and omphalocele, shown in Fig. 1), where organ displacement or external exposure leads to the loss of abdominal structures, resulting in the failure to locate standard planes. To address these issues, we propose a Multiple Instance Learning (MIL)-based method that only requires case-level annotations (coarse-grained annotations from an ultrasound image pool taken from a single examination) and does not rely on localizing standard planes.

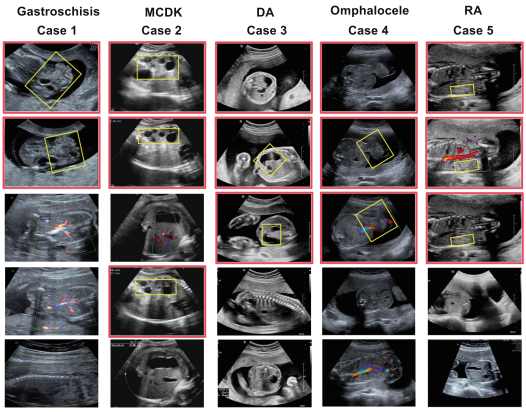


Fig. 1. Dataset of Multi-Center Prenatal Abdominal Ultrasound. Each case contains several ultrasound images taken during a single examination, including both abnormal and normal images. Abnormal images are annotated with red boxes. Abnormal structures are annotated with yellow boxes. (Color figure online)

Multiple Instance Learning (MIL) represents a powerful tool for weakly supervised classification tasks. Specifically, it aims to aggregate multiple instances (image-level) into bag-level labels (case-level) [5–7,10,11,13,19,27,30]. Zhuang et al. [31] proposed a multi-scale feature fusion-based attention MIL model for case-level classification of thyroid nodules. Kaito et al. [20] proposed a selective aggregated MIL for case-level severity in ulcerative colitis diagnosis. Wang et al. [23] proposed a deep semi-supervised MIL framework for classifying diabetic macular edema from optical coherence tomography images.

The goal of this paper is to develop an MIL-based method that does not rely on localizing standard planes for case-level prenatal abdominal ultrasound anomaly classification. Our contribution is three-fold. (1) We adopt a mixture-of-attention-experts module (MoAE): By adopting MoAE, we can assign specific attention experts for different planes (i.e., sagittal, coronal, and transverse planes). (2) We propose a medical-knowledge-driven feature selection module: A self-supervised feature selection mechanism is designed based on medical knowledge, reducing the dependency on standard planes. (3) We propose a prompt-based prototype learning method to differentiate features across different categories. Experiments comparing our method with SOTA MIL methods show significant improvements in abdominal anomaly classification, providing a reliable and automated solution for prenatal ultrasound screening.

2 Methodology

Figure 2 shows an overview of our proposed framework. First, a case image pool is fed into transformer layers for image integration. We adopt a mixture-of-attention-experts module (MoAE), which replaces the vanilla multi-head attention mechanism to control different attention experts, enabling feature

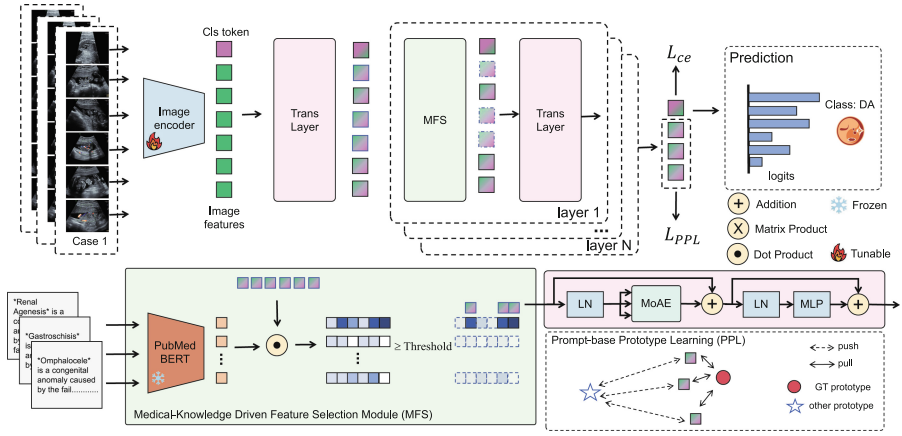


Fig. 2. Overview of our proposed framework.

integration across different planes. Secondly, a medical-knowledge-driven feature selection module (MFS) leverages medical knowledge to select significant images from the case image pool. Finally, the prompt-based prototype learning (PPL) module involves a loss function to bring paired medical knowledge and images closer in the feature space, ensuring the distinctiveness of the selected image.

TransMIL Definition for Case-Level Analysis. The vanilla TransMIL [19] was designed for Whole Slide Imaging (WSI), including two main components: the image encoder and the transformer layers. Due to the large size, WSIs are usually divided into multiple non-overlapping patches as instances. The image encoder f first extracts features X' from these instances X, which are then fed into the transformer layers to generate the final prediction y. In our task, we treat a image pool X obtained from a single examination of a case as instances, each instance x_i represents one image, formulated by:

$$X' = f(X), X = (x_1, x_2, \ldots, x_n), \tag{1}$$

$$y = TransLayer(h_{cls}, X'), X' = (x'_1, x'_2, \ldots, x'_n). \tag{2}$$

Mixture of Attention Experts Module for Integrating Planes. For the transformer layer, multi-head attention is crucial, and different attention heads have different functions, including positional function [21]. Here, we consider different planes to possess spatial position information. We adopt the mixture-of-attention-experts module (MoAE) [8] as shown in Fig. 3, treating attention heads as experts within the MoE mechanism, enabling feature integration across different planes, formulated by:

$$MoAE(X') = \sum_{i=1}^{h} g_i H^i W_O^i, \tag{3}$$

$$H^i = Attention(X'W^Q, X'W^k, X'W^v), \tag{4}$$

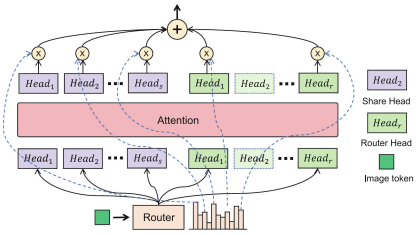


Fig. 3. Mixture of Attention Experts Module.

$$g_i = \begin{cases} \alpha_1 \text{Softmax}(W_s x_i')_j, & \text{if } 1 \le j \le h_s, \\ \alpha_2 \text{Softmax}(W_r x_i')_j, & \text{if } (W_r x_i')_j \in \text{Top-K}(\{(W_r x_i')_j | h_s + 1 \le j \le h\}), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

where H^i represents the output of the vanilla Attention mechanism. g_i represents the router score. Each token activates the top-K most relevant attention heads via routers. When the i_{th} attention head is activated, $g_i > 0$. $Softmax(W_r x_i')$ represents the router score. W represents a trainable projection matrix. This module also retains shared heads h_s to integrate common information across tokens. And the coefficients α_1 and α_2 balance the contributions of the shared and routed heads. This method enables each token to select the most relevant attention heads and produce a weighted summation based on router scores.

Medical-Knowledge Driven Feature Selection Module. Recent multi-modal pre-training models, such as CLIP [17], MedCLIP [24] and BioMedCLIP [29], have learned associations between images-text by training on large-scale language-image pairs using contrastive learning, aligning them in the feature space and enabling zero-shot or few-shot learning capabilities. Additionally, prompt learning has utilized a pre-trained Vision-Language Model (VLM) to adapt to new downstream tasks by incorporating task-specific prompts into the model's input, demonstrating strong performance [1,28].

Inspired by these, we propose a medical-knowledge-driven feature selection module (MFS). Guided by medical-knowledge prompts, The MFS performs self-supervised selection of abnormal images from the case-level image pool without relying on standard plane localization.

First, medical-knowledge are fed into a pre-trained model PubMedBERT [3] to extract the sentence feature as the *prompt*. The input format of the *prompts* is: $Prompt = \{Cls, definition, signs\}$. Secondly, we compute the similarity between each image token and the medical-knowledge *prompts* to obtain an activation score. Then, we use an adaptive threshold to filter out irrelevant images. The β increases progressively at each stage, the formula of *Threshold* can be written by:

$$Threshold = \beta * \frac{1}{n} \sum_{i=1}^{n} max(Softmax(x_i' \cdot Prompts), dim = C). \quad (6)$$

Prompt-Base Prototype learning for Enhancing MFS. To ensure that the images selected by MFS are representative, we propose a loss function based on the image tokens X' as follows:

$$L_{PPL}(X', Prompts, Cls) = \sum^{N} (\frac{1}{n} \sum_{j=1}^{n} \frac{exp(x_j' \cdot Prompts_{Cls})}{\sum_{i=1}^{C} exp(x_j' \cdot Prompts_i)}), \quad (7)$$

where $Prompts$ represents the set of all category prompts, $C = len(Prompts)$ and Cls denotes the ground truth. N represents the size of the dataset, while n represents the number of images in a single case. The image token x_j' selected by MFS and the category prompt are used as a positive pair, while other prompts form negative pairs. Treating prompts as prototypes, we pull the image tokens of the positive pair closer to the prototype center and push them away from other centers, further enhancing the effectiveness of the selected features in MFS. We further add cross-entropy loss, and the overall loss function is:

$$L(y, Cls, Prompt) = L_{ce}(y, Cls) + L_{PPL}(X', Prompts, Cls). \quad (8)$$

3 Experimental Results

Ethical Approval. This study was approved by the Clinical Research Ethics Committee of the Affiliated Suzhou Hospital of Nanjing Medical University (K-2023-067-H01) and conducted following the Declaration of Helsinki. It was registered at www.chictr.org.cn (ChiCTR2300071300).

Materials and Implementation Details. In this paper, we collected a multi-center prenatal abdominal ultrasound dataset. The age range of the pregnant women was 17 to 48 years, and the gestational age ranged from 10 to 39 weeks. Our dataset contains 2,419 cases and a total of 24,748 images, including abdominal wall defect and several common urological, digestive system anomalies: Duodenal Atresia (DA) (476), Gastroschisis (154), Omphalocele (760), Renal Agenesis (RA) (197), Multicystic Dysplastic Kidney (MCDK) (277), and Normal (555). Each case contains 1 to 65 images. The dataset was randomly split into training (1,516), validation (449), and test (454) sets at the case-level, ensuring that the distribution of categories was maintained across all subsets.

We implemented our method in PyTorch, using an NVIDIA 4090 GPU. The batch size, learning rate, and total epochs were set to 64, 1e−3, and 100, respectively. We used the Adam optimizer for model optimization. We used pre-trained model ResNet18 as image encoder. During training, we employed a warm-up learning rate strategy, starting with an initial learning rate of 1e−8, reaching

Table 1. Method comparison. The best results are shown in bold.

	DA		Gastroschisis		Omphalocele		RA		MCDK		Normal		Acc
	Sen	F1	Sen	F1	Sen	F1	Sen	F1	Sen	F1	Sen	F1	
Features Max	87.50%	91.30%	46.88%	50.00%	82.89%	84.56%	75.00%	74.07%	82.14%	79.31%	**97.44%**	89.94%	78.64%
Features Mean	92.71%	92.71%	43.75%	46.67%	78.95%	81.91%	70.00%	74.67%	**87.50%**	82.35%	96.15%	88.76%	78.18%
TransMIL	93.75%	94.74%	53.12%	65.38%	**93.42%**	86.59%	67.50%	72.00%	73.21%	82.83%	96.15%	91.46%	79.53%
AbMIL	91.67%	89.34%	50.00%	57.14%	87.50%	**87.79%**	77.50%	**79.49%**	85.71%	88.89%	94.87%	89.16%	81.21%
DSMIL	**97.92%**	95.92%	46.88%	58.82%	90.79%	87.34%	72.50%	79.45%	85.71%	88.89%	93.59%	89.02%	81.23%
AdditiveMIL	90.62%	92.06%	50.00%	54.24%	83.55%	85.52%	67.50%	69.23%	**87.50%**	83.76%	94.87%	88.10%	79.01%
IBMIL	93.75%	92.78%	43.75%	53.85%	86.18%	85.62%	80.00%	72.73%	76.79%	81.90%	92.31%	88.34%	78.80%
Mamba2MIL	94.79%	96.3%	53.12%	56.67%	78.95%	83.33%	77.50%	61.39%	73.21%	81.19%	93.59%	86.39%	78.53%
SAMIL	92.71%	89.90%	40.62%	44.83%	79.61%	81.48%	80.00%	79.01%	80.36%	79.65%	87.18%	84.47%	76.75%
Our	**97.92%**	**97.92%**	**62.50%**	**67.80%**	87.50%	**87.79%**	**85.00%**	73.91%	80.36%	**89.11%**	**97.44%**	**94.41%**	**85.12%**

Table 2. Ablation studies. The best results are shown in bold.

MFS	MoAE	PPL	DA		Gastroschisis		Omphalocele		RA		MCDK		Normal		Acc
			Sen	F1	Sen	F1	Sen	F1	Sen	F1	Sen	F1	Sen	F1	
✗	✗	✗	93.75%	94.74%	53.12%	65.38%	**93.42%**	86.59%	67.50%	72.00%	73.21%	82.83%	96.15%	91.46%	79.53%
✓	✗	✗	96.88%	97.38%	56.25%	58.06%	82.89%	84.85%	72.50%	69.05%	**85.71%**	88.89%	96.15%	90.36%	81.73%
✓	✗	✓	95.83%	96.34%	**62.50%**	59.70%	85.53%	86.09%	75.00%	**78.95%**	83.93%	87.85%	**1.00%**	**94.55%**	83.80%
✓	✓	✓	**97.92%**	**97.92%**	**62.50%**	**67.80%**	87.50%	**87.79%**	**85.00%**	73.91%	80.36%	**89.11%**	97.44%	94.41%	**85.12%**

the maximum learning rate at epoch 20, and then gradually decreasing using a half-cosine schedule to a minimum of 1e−8.

Result. Table 1 presents a comparison of our method with nine strong competitors, including Feature Max [6], Feature Mean [6], TransMIL [19], AbMIL [5], DSMIL [10], AdditiveMIL [7], IBMIL [13], Mamba2MIL [30], and SAMIL [20]. In anomalies classification, we primarily focus on the sensitivity for detecting abnormal fetuses. Therefore, we show sensitivity as the performance metric for each category, while also presenting the F1 score. Further, we use weighted accuracy as the overall performance metric to account for class imbalance in the dataset. Results show that our method outperforms all other methods in most categories. Regarding sensitivity improvement, compared to the second-best model, the improvements are DA (3.13%), Gastroschisis (9.38%), RA (5%), and Normal (1.29%), respectively. In terms of F1 score, the improvements are DA (1.62%), Gastroschisis (2.42%), Omphalocele (0.45%), MCDK (0.22%), and Normal (2.95%), respectively. Additionally, our method achieves the best overall accuracy 85.12% among all the methods, demonstrating its effectiveness.

We also preformed ablation studies to evaluate the contribution of each proposed module in Table 2. Specifically, we treated the TransMIL as the baseline and added our proposed modules. When using the MFS module alone, it can be seen that MFS contributes about 2% improvement to the overall accuracy. Furthermore, the combination of MFS and PPL further enhances performance, resulting in an additional 2% improvement. As the modules were progressively added, the overall accuracy increased significant of 5.59%, highlighting the effectiveness of the proposed modules in enhancing model performance.

Figure 4 displays typical cases using our modules. The images annotated with red boxes indicate those retained after applying our MFS, while the rest are

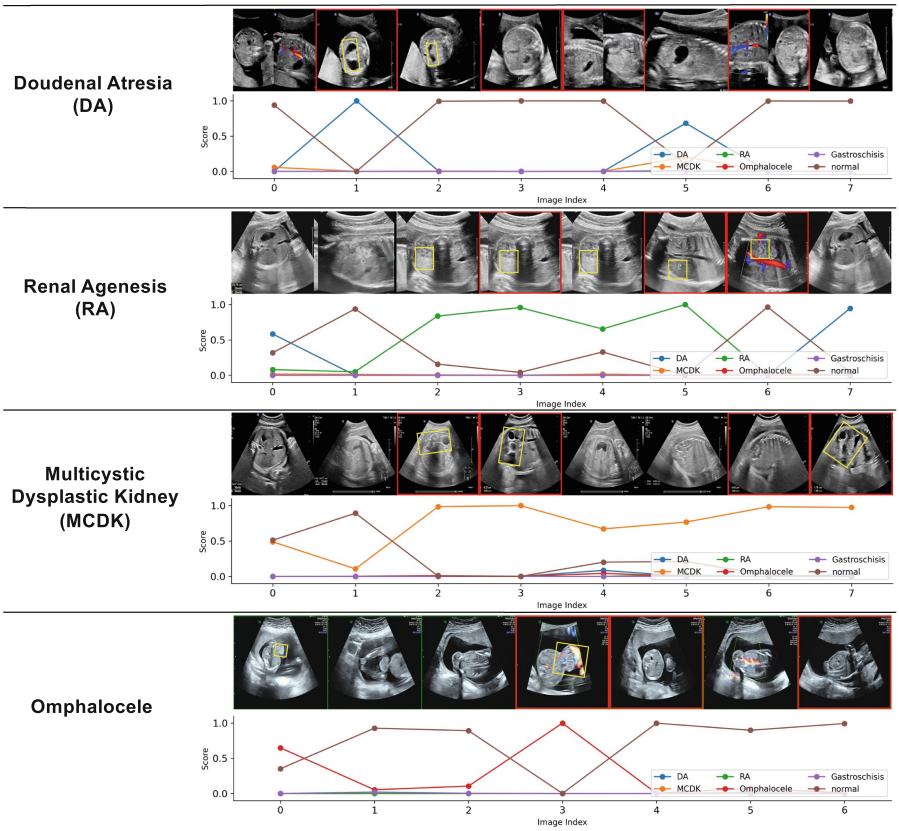


Fig. 4. Visualization of typical cases. Abnormal structures are annotated with yellow boxes. The images activated by MFS are annotated with red boxes. The line chart represents the activation scores of the corresponding image categories.

discarded. As the number of image tokens decreases, the model becomes more focused on anomaly classification. The line chart shows that the images activated by MFS generally receive higher scores. This demonstrates that even if the model does not select all abnormal images, the chosen ones remain highly representative as shown in Fig. 4 (RA and Omphalocele). These typical cases demonstrate that our method can perform case-level predictions effectively without relying on standard plane localization.

4 Conclusion

In this paper, we propose a medical-knowledge driven MIL method for improving case-level prenatal abdominal ultrasound anomalies classification. We first adopt MoAE, assigning different attention experts to specific planes enabling feature

integration. Next, we propose the MFS, which enables the model to perform self-supervised selection of significant images from case-level ultrasound image pool based on medical knowledge. Finally, we propose a loss function PPL to enhance the MFS. Extensive experiments show the effectiveness of our method. In the future, we will extend this framework to more congenital anomalies classification in fetus.

Acknowledgments. This work was supported by the grant from National Natural Science Foundation of China (12326619, 62171290); Science and Technology Planning Project of Guangdong Province (2023A0505020002); Frontier Technology Development Program of Jiangsu Province (BF2024078); Science and Technology Research and Development Program of Shaanxi Province (2011K12-05-11); Key Research and Development Program of Shaanxi (2025SF-YBXM-156).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, W., Si, C., et al.: Semantic prompt for few-shot image recognition. In: Proceedings of the IEEE/CVF Conference on CVPR, pp. 23581–23591 (2023)
2. Ciobanu, S.G., Enache, I.A., Iovoaica-Rămescu, C., Berbecaru, E.I.A., et al.: Automatic identification of fetal abdominal planes from ultrasound images based on deep learning. J. Imaging Inform. Med. 1–8 (2025)
3. Gu, Y., Tinn, R., Cheng, H., Lucas, M., et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthc. (HEALTH) **3**(1), 1–23 (2021)
4. Huang, Y., et al.: Fourier test-time adaptation with multi-level consistency for robust classification. In: Greenspan, H., et al. (eds.) MICCAI 2023. LNCS, vol. 14222, pp. 221–231. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43898-1_22
5. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning, pp. 2127–2136. PMLR (2018)
6. Ilse, M., Tomczak, J.M., Welling, M.: Deep multiple instance learning for digital histopathology. In: Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 521–546. Elsevier (2020)
7. Javed, S.A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., Prakash, A.: Additive mil: intrinsically interpretable multiple instance learning for pathology. Adv. Neural. Inf. Process. Syst. **35**, 20689–20702 (2022)
8. Jin, P., Zhu, B., Yuan, L., Yan, S.: MoH: multi-head attention as mixture-of-head attention. arXiv preprint arXiv:2410.11842 (2024)
9. Laurichesse Delmas, H., et al.: Congenital unilateral renal agenesis: prevalence, prenatal diagnosis, associated anomalies. data from two birth-defect registries. Birth Defects Res. **109**(15), 1204–1211 (2017)
10. Li, B., et al.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF conference on CVPR, pp. 14318–14328 (2021)

11. Li, H., et al.: DT-MIL: deformable transformer for multi-instance learning on histopathological image. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12908, pp. 206–216. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87237-3_20
12. Lin, M., et al.: Use of real-time artificial intelligence in detection of abnormal image patterns in standard sonographic reference planes in screening for fetal intracranial malformations. Ultrasound Obstetr. Gynecol. **59**(3), 304–316 (2022)
13. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19830–19839 (2023)
14. Liu, Z., et al.: Mitral regurgitation recogniton based on unsupervised out-of-distribution detection with residual diffusion amplification. In: Xu, X., Cui, Z., Rekik, I., Ouyang, X., Sun, K. (eds.) MLMI 2024. LNCS, vol. 15241, pp. 52–62. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73284-3_6
15. Pechriggl, E., Blumer, M., et al.: Embryology of the abdominal wall and associated malformations–a review. Front. Surg. **9**, 891896 (2022)
16. Qi, Y., et al.: Multi-center study on deep learning-assisted detection and classification of fetal central nervous system anomalies using ultrasound imaging. arXiv preprint arXiv:2501.02000 (2025)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
18. Sarhan, O., Alghanbar, M., Alsulaihim, A., et al.: Multicystic dysplastic kidney: impact of imaging modality selection on the initial management and prognosis. J. Pediatr. Urol. **10**(4), 645–649 (2014)
19. Shao, Z., et al.: TransMIL: transformer based correlated multiple instance learning for whole slide image classification. Adv. Neural. Inf. Process. Syst. **34**, 2136–2147 (2021)
20. Shiku, K., Nishimura, K., Suehiro, D., Tanaka, K., Bise, R.: Ordinal multiple-instance learning for ulcerative colitis severity estimation with selective aggregated transformer. arXiv preprint arXiv:2411.14750 (2024)
21. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can be pruned. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5797–5808 (2019)
22. Wang, J., et al.: Auto-weighting for breast cancer classification in multimodal ultrasound. In: Martel, A.L., et al. (eds.) MICCAI 2020, Part VI. LNCS, vol. 12266, pp. 190–199. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59725-2_19
23. Wang, X., Tang, F., Chen, H., Cheung, C.Y., Heng, P.A.: Deep semi-supervised multiple instance learning with self-correction for DME classification from oct images. Med. Image Anal. **83**, 102673 (2023)
24. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
25. Xie, B., et al.: Computer-aided diagnosis for fetal brain ultrasound images using deep convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. **15**(8), 1303–1312 (2020). https://doi.org/10.1007/s11548-020-02182-3
26. Xie, X., et al.: Global birth prevalence of major congenital anomalies: a systematic review and meta-analysis. BMC Public Health **25**, 449 (2025)

27. Yang, S., et al.: MambaMIL: enhancing long sequence modeling with sequence reordering in computational pathology. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15004, pp. 296–306. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72083-3_28
28. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF Conference on CVPR, pp. 6757–6767 (2023)
29. Zhang, S., Xu, Y., Usuyama, N., et al.: BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
30. Zhang, Y., Zhang, X., Wang, J., Yang, Y., Peng, T., Tong, C.: Mamba2MIL: state space duality based multiple instance learning for computational pathology. arXiv preprint arXiv:2408.15032 (2024)
31. Zhuang, L., Ivezic, V., et al.: Patient-level thyroid cancer classification using attention multiple instance learning on fused multi-scale ultrasound image features. In: AMIA Annual Symposium Proceedings, vol. 2023, p. 1344 (2024)

MGG-Net: A Multi-modal Feature Extraction and Global-Aware Feature Graph-Based Deep Learning Network for MGMT Status Classification in Glioma

Haoyang Liu[1], Yuwen Zeng[1], Xiaoyong Zhang[2], Wentong Zhou[1], Arata Nagai[1], Masayuki Kanamori[1], Hidenori Endo[1], and Noriyasu Homma[1(✉)]

[1] Tohoku University, Sendai, Japan
liu.haoyang.r4@dc.tohoku.ac.jp, homma@tohoku.ac.jp
[2] National Institute of Technology, Sendai College, Sendai, Japan

Abstract. Gliomas, especially high-grade gliomas, have a high mortality rate. O6-Methylguanine-DNA Methyltransferase (MGMT) status is crucial for gliomas treatment and prognosis. Traditional diagnosis relies on invasive tissue analysis, which is often infeasible for high-risk patients. While machine learning and deep learning methods using multi-sequence Magnetic Resonance Imaging (MRI) images and radiomics provides a non-invasive alternative, existing methods suffer from low accuracy and poor generalization due to challenges in extracting features from the integrated multi-sequence representation. To address this issue, we propose a Multi-modal feature extraction and Global-aware feature Graph-based deep learning network (MGG-Net), integrating convolutional neural network (CNN) and graph convolutional network (GCN) for multi-modal and multi-scale feature learning. Specifically, MGG-Net consists of multiple CNN-GCN stages, responsible for processing MRI image features and radiomic features at different scales. CNN blocks are used to extract fine-grained and sequence-specific local features from each MRI sequence. These features are then fed into a GCN, which models long-range dependencies and extracts high-level global representations. Finally, the fused multi-scale features extracted are used for classification. Experimental results demonstrate that MGG-Net outperforms previous approaches, effectively leveraging multi-scale and multi-modal information for improved MGMT status classification.

Keywords: Deep learning · Graph convolutional network · Glioma · Multi-sequence MRI

1 Introduction

Gliomas [1] are primary brain tumors derived from neural glial stem or progenitor cells. Among them, World Health Organization (WHO) grade IV glioblastoma (GBM) is highly malignant [2]. A crucial prognostic biomarker, O6-methylguanine-DNA methyltransferase (MGMT), plays a key role in treatment

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 355–364, 2026.
https://doi.org/10.1007/978-3-032-04947-6_34

response–its methylation-positive status suppresses gene expression [3], increasing sensitivity to temozolomide therapy [4]. Thus, MGMT methylation testing is essential for clinical decision-making. However, current detection methods require tissue samples obtained through surgery or biopsy, posing significant risks. Therefore, a safe and effective non-invasive detection strategy is urgently needed.

Multi-sequence magnetic resonance imaging (MRI) [5] and radiomics [6] offer non-invasive methods for MGMT status diagnosed by capturing tumor anatomy and function. Multi-sequence MRI provides structural and physiological details, while radiomics extracts high-throughput features reflecting tumor heterogeneity. Although multi-sequenced MRI and radiomics have been widely explored, existing methods struggle to fully utilize complementary information. Prior studies fall into machine learning and deep learning approaches. Machine learning [7–9] relies on handcrafted features, limiting scalability and adaptability. Deep learning, including convolution neural networks (CNNs) [10–12] and Transformers [13], address these issues but have their own limitations. CNNs process images as structured grids, enabling the extraction of single-sequence MRI features but struggling to capture fused-sequence MRI representations [14]. Transformers process images as linear sequences, computing token similarities. However, different MRI sequences, despite representing the same medical entity, may encode information differently, making multi-modal fusion challenging. These limitations highlight the need for a more effective fusion strategy.

To address these limitations, we propose a Multi-Modal Feature Extraction and Global-Aware Feature Graph-Based Deep Learning Network (MGG-Net). MGG-Net consists of two key components: Multi-modal Feature Extraction (MFE) and Global-Aware Feature Graph (GAFG). The MFE module leverages CNNs to extract features from individual MRI sequences, overcoming the limitation of CNNs in capturing fused sequence representations. Meanwhile, the GAFG module employs graph convolutional networks (GCNs) [15] to construct adaptive feature graphs that encode spatial relationships and intra-tumor dependencies. Unlike Transformers, which compute token similarities in a fixed sequence, GAFG dynamically models relationships between MRI sequences, adapting to their differing information representations. By integrating CNN-driven local feature extraction with GCN-based global aggregation, MGG-Net effectively addresses the fusion challenges faced by prior methods. This dynamic feature refinement across graph nodes enhances cross-modality interactions and intra-tumor feature propagation, allowing the model to focus on the most informative tumor regions, such as necrotic cores, which are crucial for MGMT status classification. Ultimately, MGG-Net overcomes the rigid grid constraints of CNNs and the limitations of sequence-based modeling in Transformers, leading to improved classification performance.

Our main contributions are summarized as follows:

– In each MFE module, we process the four MRI sequence separately and incorporate their radiomic features, capturing local representations at each scale before downsampling for the next stage.

– The GAFG module applies dynamic axial graph convolution at each scale to aggregate global contextual relationships, ultimately providing a novel and non-invasive approach for MGMT status classification.

2 Methodology

This study proposes MGG-Net, as shown in Fig. 1 consisting of MFE and GAFG modules. The MFE module extracts local information by independently processing four MRI sequences (T1c, FLAIR, DWI, SWI) using feature extractor and integrating radiomic features for each sequence. The extracted sequence-specific features are passed to the GAFG module, where they are further integrated to capture inter-modality dependencies. MGG-net adopts a multi-scale architecture with four cascaded stages, where features are progressively refined. Finally, Global Average Pooling (GAP) fuses multi-scale features, and a lightweight 3D convolution serves as the MGMT classification head, determining whether MGMT is methylated.

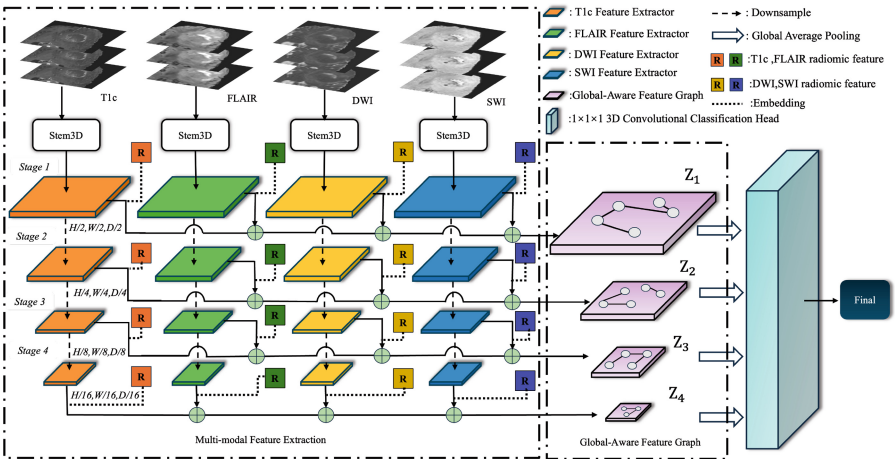


Fig. 1. The MGG-Net framework consists of the Multi-modal Feature Extraction (MFE) and Global-Aware Feature Graph (GAFG) modules. MRI sequences (T1c, FLAIR, DWI, SWI) are processed independently, with radiomic features embedded into each sequence at every stage. Extracted features are fused and structured into a graph to model inter-modal and spatial relationships, enabling MGMT classification.

2.1 Multi-modal Feature Extraction (MFE)

The MFE module consists of four parallel Feature Extraction Streams, each processing one MRI sequence (T1c, FLAIR, DWI, SWI). As shown in Fig. 2, (a)

illustrates 3D stem layer (Stem3D) used for initial feature extraction, while (c) presents the Feature Extractor. The feature extractor consists of a Depthwise Separable 3D Convolution (DWSConv) [16] for local feature extraction and an Inverted Residual (IR) connection for feature refinement. Given an input MRI volume $\mathbf{X}_m \in \mathbb{R}^{C_{\text{in}} \times H \times W \times D}$, where m denotes different MRI sequences, Stem3D extracts preliminary features and downsamples the input, producing $\mathbf{X}'_m$. This is then processed by DWSConv, which sequentially applies pointwise, depthwise, and pointwise convolutions. A pointwise convolution $\mathbf{W}_1 \in \mathbb{R}^{1\times1\times1\times C_{\text{in}} \times C_{\text{mid}}}$ expands the channel dimension, followed by batch normalization and GeLU activation. A depthwise convolution $\mathbf{W}_d \in \mathbb{R}^{K\times K\times K\times C_{\text{mid}}}$ captures spatial features while maintaining channel independence. Finally, another pointwise convolution $\mathbf{W}_2 \in \mathbb{R}^{1\times1\times1\times C_{\text{mid}} \times C_{\text{out}}}$ reduces the channel dimension, followed by batch normalization. The IR connection adds the Stem3D output directly to the stage final result, refining features while preserving spatial consistency.

At the end of each MFE stream, the radiomic features $\mathbf{R}_m \in \mathbb{R}^{B\times C_r}$, which are directly extracted from the tumor regions in the original images, are projected into an embedding space and then transmitted to match the spatial dimensions of $\mathbf{Y}_m$. Specifically, a sequence-specific transformation layer, MLP_m, maps $\mathbf{R}_m$ to $\mathbb{R}^{B\times C_{\text{mid}}}$. The resulting features are then expanded along spatial dimensions to $\mathbb{R}^{B\times C_{\text{mid}} \times D\times H\times W}$ to ensure compatibility with $\mathbf{Y}_m$. Finally, the expanded radiomic features are combined with $\mathbf{Y}_m$ via element-wise addition, producing the stage final feature map $\mathbf{Z}_m$.

After feature extraction, spatial resolution is progressively reduced, allowing the network to capture high-level representations. The extracted individual feature maps from all four sequences are then fused into a integrated representation and processed by the Global-Aware Feature Graph (GAFG) module.

2.2 Global-Aware Feature Graph (GAFG)

The GAFG module, focuses on capturing long-range dependencies by modeling global spatial relationships in the fused multi-modal feature representation, as shown in Fig. 2(c). First, a fusion layer concatenates the four individual feature maps from the MFE module along the channel dimension and applies a $1 \times 1 \times 1$ convolution. Next, a depthwise $5 \times 5 \times 5$ conditional position encoding (CPE) integrates positional information by convolving and then adding the original features back. To preserve the initial fused representation, a residual connection is introduced, where the input to CPE is added back to the final output.

The position-aware feature is then passed to the dynamic axial graph convolution network (DA-GCN). To estimate the overall mean and standard deviation of the position-aware feature, it is first divided into eight quadrants. The L_1 distances between corresponding voxels in diagonally opposite quadrants are computed, forming a distance distribution. The mean and standard deviation of this distribution are then used as an approximation for the global statistics of the position-aware feature. Following this, a rolling operation is performed along the depth, height, and width dimensions based on the predefined K value. As illustrated in Fig. 2(e), when K is set to half of the depth, height, and width,

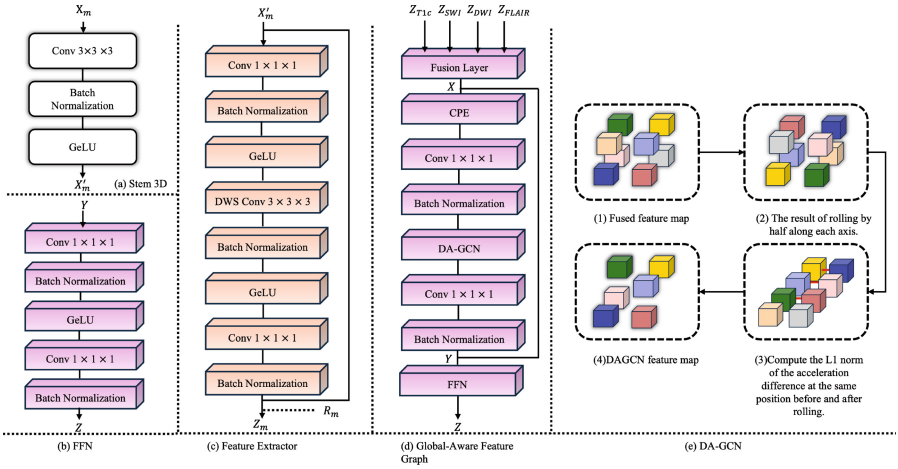


Fig. 2. Detailed structure of MFE and GAFG. (a) illustrates Stem3D, which performs preliminary feature extraction on MRI images. (b) presents the structure of the final FFN layer in GAFG. (c) shows the feature extractor structure for the four MRI sequences, incorporating radiomic features at the final step. (d) depicts the architecture of GAFG. (e) demonstrates the workflow of DA-GCN using an example where the image is rolled by half of its depth, height, and width.

the rolling process of DA-GCN is demonstrated. The L_1 distances between voxel pairs before and after rolling are computed. If the distance falls within the range of mean ± standard deviation, the two voxels are considered to share information, and an edge is constructed between them. As shown in Fig. 3, the application of DA-GCN enables a more precise extraction of the necrotic tumor region compared to CNN-based methods, demonstrating its effectiveness in capturing tumor-specific features.

Finally, as shown in Fig. 2(b), a feed-forward network (FFN) refines $\mathbf{Y}$ via a channel-expansion and reduction process, combined with residual connections and layer scaling. This process produces the output $\mathbf{Z}$, which represents the output of Stage 1. The MFE-GAFG consists of four stages, as shown in Fig. 1 each generating a $\mathbf{Z}$ with progressively smaller spatial dimensions and increasing feature dimensions. These four $\mathbf{Z}$ representations are then stacked along the channel dimension and passed through an MGMT classification head followed by a softmax layer, producing the final probability values for MGMT methylated and unmethylated status.

3 Experiment and Result

3.1 Dataset and Implementation Details

This study uses the UCSF-PDGM dataset [17], which includes 501 glioma patients with standardized 3T multi-sequence MRI scans (T1, T1c, T2, FLAIR,

DWI, SWI), tumor segmentation labels, and genetic data (IDH, MGMT). We focused on WHO grade IV cases [18], selecting 367 patients with enhancing tumor, necrotic tumor, and edema regions based on segmentation and MGMT status (positive/negative). Based on experimental results, we identified T1c, FLAIR, SWI, and DWI as the most informative MRI sequence for MGMT status classification. The dataset was split 0.85:0.15 for training and testing, with four-fold cross-validation. Tumor regions were extracted and cropped to 128×128×128. PyRadiomics [19] was then used to extract radiomic features, and ANOVA selected the four most relevant features per region for MGMT prediction.

Our MGG-Net was trained on an NVIDIA RTX A6000 GPU. We employed weighted focal loss [20] for the MGMT classification. Four-fold cross-validation was conducted, with each fold trained for 80 epochs. The learning rate used cosine annealing, decaying from 0.0002 to 0.0001. The batch size was set to 4.

3.2 Comparison with Existing Methods

We evaluate model performance using accuracy (Acc), precision (Prec), recall (Rec), F1-score (F1), and AUC [21], reported as mean ± standard deviation (SD). For comparison, we selected ResNet+SVM [11] and vViT [13], representing CNN and Transformer architectures, respectively, to evaluate MGG-Net's advantages in local feature extraction and global modeling. As shown in Table 1, MGG-Net (T1c, DWI, SWI, FLAIR) achieved the highest performance in Acc, Prec, F1, and AUC, while maintaining competitive Rec. Interestingly, traditional methods showed decreased performance when using DWI and SWI instead of T1 and T2, while MGG-Net achieved better results with DWI and SWI, highlighting its superior ability to handle low-contrast inputs.

Table 1. Performance of Different Methods Incorporating Radiomic Features (Mean ± SD)

Method	Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC (%)
ResNet+SVM (T1, T1c, T2, FLAIR) [11]	$66.07_{\pm 4.2}$	$73.67_{\pm 0.9}$	$\mathbf{81.88}_{\pm 10.0}$	$77.20_{\pm 4.6}$	$53.56_{\pm 3.8}$
ResNet+SVM (T1c, DWI, SWI, FLAIR) [11]	$61.61_{\pm 0.9}$	$70.41_{\pm 1.1}$	$80.00_{\pm 4.7}$	$74.80_{\pm 1.4}$	$45.35_{\pm 3.0}$
vViT (T1c, T2) [13]	$52.68_{\pm 2.9}$	$70.71_{\pm 3.9}$	$59.38_{\pm 11.0}$	$63.63_{\pm 5.7}$	$49.22_{\pm 6.6}$
MGG-Net (T1, T1c, T2, FLAIR)	$66.97_{\pm 2.0}$	$75.20_{\pm 2.7}$	$80.63_{\pm 3.7}$	$77.70_{\pm 12.0}$	$63.28_{\pm 3.3}$
MGG-Net (T1c, DWI, SWI, FLAIR)	$\mathbf{72.33}_{\pm 5.1}$	$\mathbf{81.50}_{\pm 2.9}$	$79.38_{\pm 8.4}$	$\mathbf{80.18}_{\pm 4.5}$	$\mathbf{69.02}_{\pm 2.2}$

Previous studies commonly used T1, T1c, T2, FLAIR, and radiomics for glioma analysis. We hypothesized that replacing T1 and T2 with DWI and SWI could better capture tumor cellularity, necrosis, and vascular features [22], enhancing classification. To validate this, we compared both sequence sets across methods. Results confirmed our hypothesis: MGG-Net performed more robustly than the baselines, especially when using DWI and SWI. The contrast between ResNet+SVM's performance drop and MGG-Net's improvement suggests that

the MFE module extracts more informative features from low-contrast images, and the GAFG module effectively fuses global multi-sequence information. While the improvement of MGG-Net across sequence sets was modest, its relative gain over traditional methods was substantial.

3.3 Ablation Study on MGG-Net

To gain deeper insights into the impact of different MRI sequence combinations on MGMT status classification, we conducted an ablation study, as shown in Table 2. Previous study [11] has shown that using a single T1 sequence can achieve higher classification accuracy compared to certain multi-modal combinations. However, our ablation results demonstrate that integrating multiple sequences, particularly T1c, FLAIR, SWI, and DWI, enhances classification performance. The best performance was achieved using these four MRI sequences along with radiomics features, yielding 72.3 ± 5.1% accuracy, 81.5 ± 2.9% precision, 80.2 ± 4.5% F1-score, and the highest AUC of 69.0 ± 2.2%. Notably, the inclusion of SWI and DWI significantly improved the results, as SWI provides insights into tumor angiogenesis and necrosis, while DWI reflects tumor cellularity and microstructural changes, both of which are critical for glioblastoma characterization. These findings suggest that while a single T1 sequence may perform well in certain cases, leveraging complementary information from multiple modalities leads to a more robust MGMT classification model.

Table 2. Performance of Different MRI Sequence Combinations (Mean ± SD)

T1c	FLAIR	SWI	DWI	T1	T2	Radiomics	Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC (%)
✓	✓			✓	✓	✓	$66.97_{\pm 2.0}$	$75.20_{\pm 2.7}$	$80.63_{\pm 3.7}$	$77.70_{\pm 12.0}$	$63.3_{\pm 3.3}$
✓	✓	✓	✓				$68.75_{\pm 3.2}$	$74.58_{\pm 2.4}$	$\mathbf{85.63}_{\pm 5.4}$	$79.59_{\pm 2.5}$	$63.01_{\pm 4.2}$
✓						✓	$60.27_{\pm 2.9}$	$72.01_{\pm 1.2}$	$72.50_{\pm 4.3}$	$72.2_{\pm 2.7}$	$44.49_{\pm 7.1}$
✓	✓	✓	✓			✓	$\mathbf{72.33}_{\pm 5.1}$	$\mathbf{81.50}_{\pm 2.9}$	$79.38_{\pm 8.4}$	$\mathbf{80.18}_{\pm 4.5}$	$\mathbf{69.02}_{\pm 2.2}$

Additionally, we compared our MGG-Net with MGC-Net, which replaces the graph module with a CNN [23] (Table 3). MGG-Net achieved higher accuracy and AUC, demonstrating its superiority in capturing tumor spatial relationships on multi-modal feature maps. These results confirm the advantage of integrating graph-based learning for MGMT status classification.

Table 3. Replacing Graph Module with CNN (Mean ± SD)

Model	Acc (%)	Prec (%)	Rec (%)	F1 (%)	AUC (%)
MGG-Net	$\mathbf{72.33}_{\pm 5.1}$	$\mathbf{81.50}_{\pm 2.9}$	$79.38_{\pm 8.4}$	$\mathbf{80.18}_{\pm 4.5}$	$\mathbf{69.02}_{\pm 2.2}$
MGC-Net	$70.54_{\pm 4.5}$	$78.42_{\pm 4.4}$	$\mathbf{81.88}_{\pm 8.5}$	$79.71_{\pm 3.7}$	$68.26_{\pm 4.9}$

Because some classification models [24,25] yield correct results without focusing on key regions, we visualized feature maps to analyze the mechanisms behind performance gains. Figure 3 presents feature maps of two test cases. It is evident that the most discriminative features are concentrated in the necrotic core of the tumor, which aligns with the superior classification performance achieved when incorporating DWI and SWI. These sequences capture essential tumor characteristics, such as cellular density and vascular abnormalities, which are critical to predicting the status of MGMT. Furthermore, the GAFG-based representation exhibits a more focused activation compared to the CNN-based approach, demonstrating that the graph-based method more effectively captures key tumor features. This further supports the advantage of our multi-sequence MRI fusion strategy in improving classification accuracy.

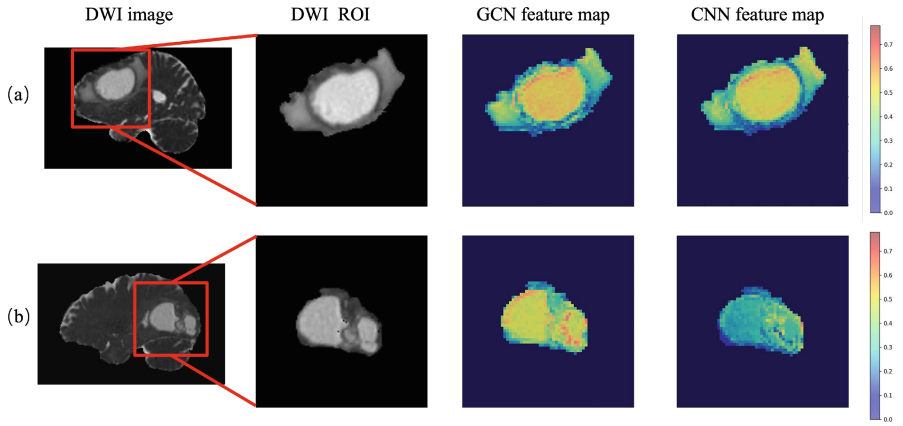


Fig. 3. Visualization results of classification feature maps using Graph and CNN-based methods. Rows (a) and (b) correspond to two patients from the test set. The "DWI image" column presents the tumor structure of each patient, while the "DWI ROI" column highlights the tumor region. The "GCN feature map" and "CNN feature map" columns show the extracted feature maps of the tumor region using the Graph and CNN-based methods, respectively. The red-highlighted areas indicate the key regions of interest for the models. (Color figure online)

4 Conclusion

In this study, we propose MGG-Net, a novel network for multi-sequence MRI and radiomic features processing that leverages GCNs to enhance global feature extraction. By integrating multi-scale features, the model effectively utilizes diverse imaging sequences, particularly DWI and SWI, which provide crucial information on tumor cellularity and necrosis, improving MGMT status

classification. Although the limited radiomic features used may restrict significant performance improvements, future work will explore enhanced fusion methods. Feature visualization confirms that MGG-Net prioritizes key tumor subregions, strengthening its effectiveness in glioblastoma analysis and highlighting the potential of graph-based learning for medical image classification.

Acknowledgments. This study was supported by the AMED No. 24ym0126812j0003 and the JSPS KAKENHI Grant Number JP25K21579.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Weller, M., Wick, W., Aldape, K.: Brada: glioma. Nat. Rev. Dis. Primers. **1**(1), 1–18 (2015)
2. Louis, D.N., Perry, A., Wesseling, P.: The 2021 who classification of tumors of the central nervous system: a summary. Neuro Oncol. **23**(8), 1231–1251 (2021)
3. Gibson, D., Vo, A.H., Lambing, H., Bhattacharya, P., Tahir, P., Chehab, F.F.: A systematic review of high impact CpG sites and regions for MGMT methylation in glioblastoma [a systematic review of MGMT methylation in GBM]. BMC Neurol. **24**(1), 103 (2024)
4. Wick, W., Weller, M., Van Den Bent, M.: MGMT testing-the challenges for biomarker-based glioma treatment. Nat. Rev. Neurol. **10**(7), 372–385 (2014)
5. Mabray, M.C., Barajas Jr., R.F., Cha, S.: Modern brain tumor imaging. Brain Tumor Res. Treat. **3**(1), 8 (2015)
6. Chaddad, A., Kucharczyk, M.J., Daniel, P., Sabri, S., Jean-Claude, B.J., Niazi, T.: Radiomics in glioblastoma: current status and challenges facing clinical implementation. Front. Oncol. **9**, 374 (2019)
7. Pálsson, S., Cerri, S., Van Leemput, K.: Prediction of MGMT methylation status of glioblastoma using radiomics and latent space shape features. In: Crimi, A., Bakas, S. (eds.) BrainLes 2021. LNCS, vol. 12963, pp. 222–231. Springer, Cham (2021). https://doi.org/10.1007/978-3-031-09002-8_20
8. Kawaguchi, R.K., Takahashi, M., Miyake, M.: Assessing versatile machine learning models for glioma radiogenomic studies across hospitals. Cancers **13**(14), 3611 (2021)
9. Lu, Y., Patel, M., Natarajan, K.: Machine learning-based radiomic, clinical and semantic feature analysis for predicting overall survival and MGMT promoter methylation status in patients with glioblastoma. Magn. Reson. Imaging **74**, 161–170 (2020)
10. Saeed, N., Hardan, S., Abutalip, K., Yaqub, M.: Is it possible to predict MGMT promoter methylation from brain tumor MRI scans using deep learning models? In: International Conference on Medical Imaging with Deep Learning, pp. 1005–1018. PMLR (2022)
11. Saxena, S., Jena, B., Mohapatra, B., Gupta, N.: Fused deep learning paradigm for the prediction of o6-methylguanine-DNA methyltransferase genotype in glioblastoma patients: a neuro-oncological investigation. Comput. Biol. Med. **153**, 106492 (2023)

12. Tang, Z., Xu, Y., Jin, L., Aibaidula, A.: Deep learning of imaging phenotype and genotype for predicting overall survival time of glioblastoma patients. IEEE Trans. Med. Imaging **39**(6), 2100–2109 (2020)
13. Usuzaki, T., Takahashi, K., Inamori, R., Morishita, Y.: Identifying key factors for predicting o6-methylguanine-DNA methyltransferase status in adult patients with diffuse glioma: a multimodal analysis of demographics, radiomics, and MRI by variable vision transformer. Neuroradiology **66**(5), 761–773 (2024)
14. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. Adv. Neural. Inf. Process. Syst. **32** (2019)
15. Han, K., Wang, Y., Guo, J., Tang, Y., Wu, E.: Vision GNN: an image is worth graph of nodes. Adv. Neural. Inf. Process. Syst. **35**, 8291–8303 (2022)
16. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T.: MobileNets: efficient convolutional neural networks for mobile vision applications (2017). https://arxiv.org/abs/1704.04861
17. Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D.: The university of California San Francisco preoperative diffuse glioma MRI dataset. Radiology: Artif. Intell. **4**(6), e220058 (2022)
18. Whitfield, B.T., Huse, J.T.: Classification of adult-type diffuse gliomas: impact of the world health organization 2021 update. Brain Pathol. **32**(4), e13062 (2022)
19. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V.: Computational radiomics system to decode the radiographic phenotype. Can. Res. **77**(21), e104–e107 (2017)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
21. Powers, D.M.: Evaluation: from precision, recall and F-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020)
22. Yang, X., Hu, C., Xing, Z., Lin, Y., Su, Y., Wang, X.: Prediction of Ki-67 labeling index, ATRX mutation, and MGMT promoter methylation status in IDH-mutant astrocytoma by morphological MRI, SWI, DWI, and DSC-PWI. Eur. Radiol. **33**(10), 7003–7014 (2023)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25** (2012)
24. Zeng, Y., Zhang, X., Wang, J., Usui, A., Ichiji, K., Bukovsky, I.: Inconsistency between human observation and deep learning models: assessing validity of postmortem computed tomography diagnosis of drowning. J. imaging inform. med. **37**(3), 1–10 (2024)
25. Zhang, Z., Zhang, X., Ichiji, K., Bukovskỳ, I., Homma, N.: How intra-source imbalanced datasets impact the performance of deep learning for Covid-19 diagnosis using chest x-ray images. Sci. Rep. **13**(1), 19049 (2023)

Minuscule Cell Detection in AS-OCT Images with Progressive Field-of-View Focusing

Boyu Chen[1], Ameenat Solebo[2], Daqian Shi[1(✉)], Jinge Wu[1], and Paul Taylor[1(✉)]

[1] Institute of Health Informatics, University College London, London, UK
{daqian.shi,p.taylor}@ucl.ac.uk
[2] Great Ormond Street Institute of Child Health, University College London, London, UK

Abstract. Anterior Segment Optical Coherence Tomography (AS-OCT) is an emerging imaging technique with great potential for diagnosing anterior uveitis, a vision-threatening condition. This condition is characterized by the presence of inflammatory cells in the eye's anterior chamber (AC). Automatic detection of these cells on AS-OCT images has attracted great attention. However, this task is challenging since each cell is minuscule (extremely small), representing less than 0.005% of the high-resolution image. Moreover, pixel-level noise introduced by OCT can be misclassified as cells, leading to false positive detections. These challenges make both traditional image processing algorithms and state-of-the-art (SOTA) deep learning object detection methods ineffective for this task. To that end, we propose a minuscule cell detection framework that progressively refines the field-of-view from the whole image to the AC region, and further to minuscule regions potentially containing individual cells. Our framework consists of: (1) a Field-of-Focus module that uses a vision foundation model to zero-shot segment the AC region, and (2) a Fine-grained Object Detection module that introduces Minuscule Region Proposal followed by our Cell Mamba to distinguish individual cells from noise. Experimental results demonstrate that our framework outperforms SOTA methods, improving F1 by around 7% over the best baseline and offering a more reliable alternative for cell detection. Our code is available at: https://github.com/joeybyc/MCD.

Keywords: AS-OCT · Anterior Uveitis · Cell Detection

1 Introduction

Anterior Segment Optical Coherence Tomography (AS-OCT) is an imaging technique for visualizing the eye's anterior chamber (AC), with particular potential in diagnosing a vision-threatening condition, anterior uveitis. This condition is characterized by inflammatory cells in the AC [27]. Automated detection of these cells using AS-OCT images has attracted substantial research interest [10].

However, this task is challenging due to the minuscule size of the cells. As shown in Fig. 1a (1), on x160 magnification of two regions of an AS-OCT image

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 365–375, 2026.
https://doi.org/10.1007/978-3-032-04947-6_35

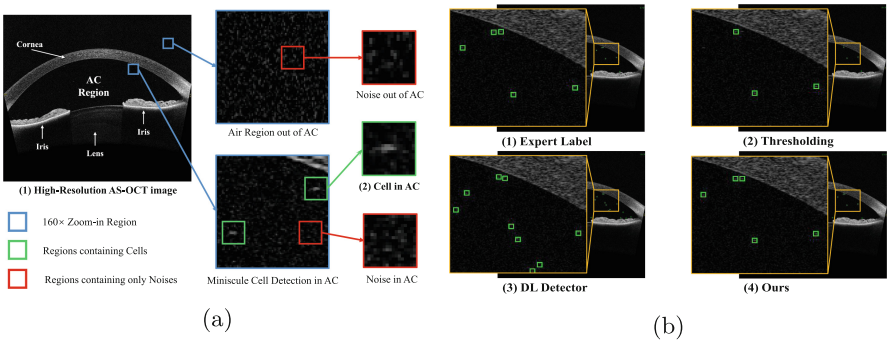


Fig. 1. (a) Cells and noise in the AS-OCT image. (b) Qualitative comparisons of cell detection results using different methods, with green boxes indicating the detected cells. (Color figure online)

(blue boxes), the hyperreflective particles (green boxes in Fig. 1a (2)) represent target cells. Each occupies less than 0.005% of the image. Moreover, pixel-level noise (highlighted in red boxes) in OCT images closely resembles the cells in texture, leading to false positive detections. Traditional denoising methods are ineffective as they indiscriminately remove both cells and noise.

Previous automated detection methods primarily rely on thresholding, which determines a decision boundary (threshold) for classifying pixels based on their intensity values (ranging from 0 for black to 255 for white). Pixels exceeding this threshold are classified as objects, while others are considered background. In particular, the Isodata algorithm [22], which is the default thresholding method in ImageJ[1], has been widely used [1,3,9,19,28].

However, these methods still require manual delineation of the AC region to exclude non-cell particles detected outside this area, making the process time-consuming and error-prone. Moreover, these methods could miss the cells with intensity values below their calculated decision boundary (see Fig. 1b (2)).

Deep learning (DL) object detection methods present another potential solution [21]. However, they struggle to capture the fine-grained details of cells during feature extraction and down-scaling operations, as each cell occupies an extremely small portion of the image. Consequently, DL detectors often misclassify pixel-level noise as cells due to their textural similarity (see Fig. 1b (3)).

To tackle these challenges, we introduce a progressive field-of-view focusing strategy. The core idea is to gradually refine the focus of analysis: first, by concentrating on the AC region where such cells are typically found; then, by zooming in on extremely small, fine-grained subregions that may contain individual cells; and finally, by examining the spatial patterns embedded within the latent representations derived from these minuscule subregions to accurately distinguish cells from background noise. To implement this approach, we present a Minuscule Cell Detection (MCD) framework, consisting of two modules:

[1] https://imagej.net/ij/.

The Field-of-Focus (FoF) module leverages a foundation vision model to zero-shot segment the AC region, eliminating the need for manual intervention. Within this region, our Fine-grained Object Detection (FOD) module detects individual cells through two key components. We design the Minuscule Region Proposal to identify extremely small regions where cells might be present, capturing potential cells that thresholding methods would miss. Subsequently, our Cell Mamba learns to process fine-grained details within these minuscule regions to distinguish cells from noise, overcoming the limitation of conventional DL detectors in capturing features of such tiny objects. As a result, our framework not only detects more cells than thresholding methods but also achieves higher detection accuracy compared to existing DL detectors (see Fig. 1b (4)).

Our contributions are:

1) We propose MCD to detect minuscule inflammatory cells in AS-OCT images, where each cell represents less than 0.005% of the image. MCD achieves SOTA cell detection performance, improving F1 by around 7% over the best baseline and offering a more reliable alternative for cell detection.
2) We reveal a critical limitation in anterior uveitis studies using thresholding methods: these studies directly use the results detected by these methods for subsequent analysis without verification, assuming detected objects are cells. Our experiments demonstrate these methods are actually inaccurate, which has crucial clinical implications for future research.

2 Method

Figure 2 illustrates our MCD framework, which comprises a Field-of-Focus (FoF) module and a Fine-grained Object Detection (FOD) module.

The FoF (bottom-left of Fig. 2) first segments the anterior chamber (AC) region. This resolves the limitation of threshold-based methods that cannot eliminate false positives outides the AC region without manual effort. Unlike conventional DL segmentation models (e.g., U-Net [23]) that require extensive annotated data, FoF utilizes a vision foundation model for zero-shot segmentation. These types of models can segment regions given an image and points (prompts) inside the object of interest. We design an Image-to-AC-Prompt (I2ACP) algorithm (Sect. 2.1) to automatically generate these prompt points. In this work, we adopt Segment Anything Model (SAM) [16] as our foundation model to produce the AC segmentation mask.

Within the AC region, the Fine-grained Object Detection (FOD) module identifies cells through two components. First, a Minuscule Region Proposal (MiRP) algorithm (Sect. 2.2) locates candidate boxes that potentially contain cells (middle of Fig. 2), aiming to identify cells that thresholding methods may miss. These regions preserve fine-grained details, which are processed by our Cell Mamba (Sect. 2.3) to distinguish between actual cells and background noise (bottom-right of Fig. 2). This learns the fine-grained details that traditional DL detectors lose.

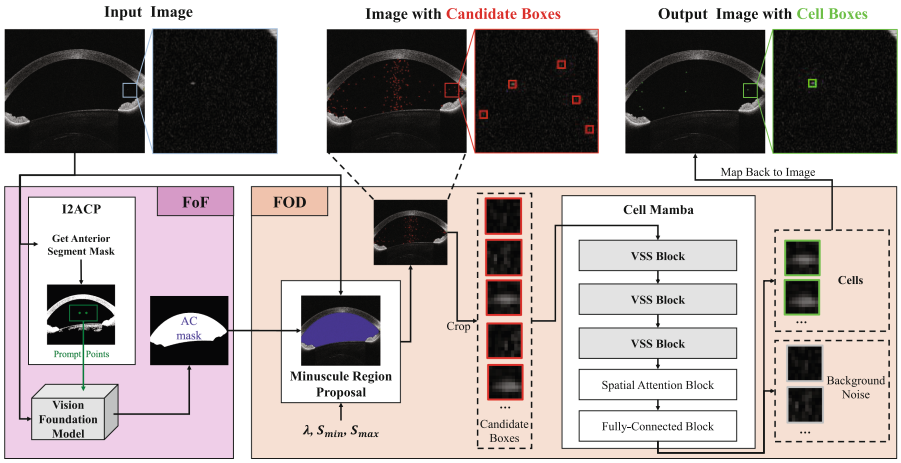


Fig. 2. The overall architecture of our minuscule cell detection framework.

2.1 Image-to-AC-Prompt (I2ACP)

The I2ACP automatically generates prompts for vision foundation models to segment the AC region by exploiting two key anatomical features: (1) the anterior segment (AS) structure (cornea, iris, and lens, see Fig. 1a(1)) appears as high-intensity regions surrounding the AC, and (2) the AS centroid and its neighboring points naturally fall within the AC.

To generate the prompts, the I2ACP first creates a binary mask by thresholding the AS-OCT image at its mean intensity, capturing the bright AS structure (the mask inside the I2ACP block of Fig. 2). The largest connected component typically represents the AS. In cases where AS appears as two disconnected components, we compute the area ratio $r = A_2/A_1$ between the second-largest (A_2) and largest (A_1) components, merging them if $r > \mathcal{R}$. Based on the mask's centroid (x^{cent}, y^{cent}), we generate a set of prompts p_i using controlled offsets:

$$p_i = (x^{cent} + \delta_i^w, y^{cent} + \delta_i^h)$$

These prompts (green points in the I2ACP block of Fig. 2) guide the vision foundation model to accurately segment the AC region.

2.2 Minuscule Region Proposal (MiRP) for Finding More Cells

MiRP identifies minuscule regions potentially containing cells within the segmented AC region. It exploits the observation that cells appear as bright, minuscule particles against darker backgrounds. Given a high-resolution AS-OCT image $\mathcal{I}$, an AC mask $\mathcal{M}^{AC}$ from FoF, and cell size bounds $[\mathcal{S}_{min}, \mathcal{S}_{max}]$, MiRP first converts $\mathcal{I}$ to grayscale $\mathcal{G}$.

We use Otsu's method [20], a widely-used particle detection method in medical imaging, to calculate an initial thresholding decision boundary $\mathcal{T}^{init}$ from $\mathcal{G}$. To capture lower-intensity cells that this boundary might miss, we introduce an adjustment factor λ:

$$\mathcal{T}^{opt} = \lambda \times \mathcal{T}^{init}$$

where $\lambda < 1.0$ lowers the decision boundary, enabling detection of dimmer particles overlooked by conventional thresholding methods (Fig. 1b(2)).

Using $\mathcal{T}^{opt}$, we generate a binary mask $\mathcal{M}^{bright} := \mathcal{G} > \mathcal{T}^{opt}$ and identify candidate cells by extracting connected components within the AC ($\mathcal{M}^{bright} \wedge \mathcal{M}^{AC}$). We then apply size constraints $[\mathcal{S}_{min}, \mathcal{S}_{max}]$ to filter out candidates too large or too small to be cells [9,19]. For each remaining particle, we generate $w \times h$ bounding boxes centered at its centroid as candidate boxes (shown in the middle of Fig. 2).

While lowering the decision boundary helps detect more potential cells, it inevitably introduces noise particles that conventional thresholding would exclude. We propose Cell Mamba to tackle this by learning the fine-grained details that distinguish actual cells from background noise in these minuscule regions.

2.3 Cell Mamba for Fine-Grained Feature Learning

Cell Mamba captures the fine-grained details essential for distinguishing minuscule cells from noise. Given a candidate box $\mathcal{X} \in \mathcal{R}^{w \times h \times 3}$ (where w, h, and 3 represent width, height, and number of channels), we extract features by using three stacked Visual State-Space blocks [18], denoted as $VSS_i(\cdot)$:

$$\mathcal{F}_{vss} = VSS_3(VSS_2(VSS_1(\mathcal{X})))$$

The resulting feature map $\mathcal{F}_{vss} \in \mathcal{R}^{W^v \times H^v \times C^v}$ (where W^v, H^v, and C^v denote the width, height, and channels of the feature map) captures essential fine-grained details. We then apply a spatial attention block inspired by [29] to model spatial relationships, generating an attention map $\mathcal{A} \in \mathcal{R}^{W^v \times H^v \times 1}$:

$$\mathcal{A} = \sigma(\text{Conv}([\text{AvgPool}(\mathcal{F}_{vss}); \text{MaxPool}(\mathcal{F}_{vss})]))$$

where $\sigma(\cdot)$ is the sigmoid activation, $Conv(\cdot)$ denotes a convolutional layer with kernel size 3 and padding 1, and $AvgPool(\cdot)$ and $MaxPool(\cdot)$ represent channel-wise average and maximum pooling operations. The attended feature map is obtained through element-wise multiplication:

$$\mathcal{F}_{sa} = \mathcal{A} \odot \mathcal{F}_{vss}$$

where $\mathcal{A}$ is broadcast across all channels of $\mathcal{F}_{vss}$. Then, $\mathcal{F}_{sa}$ is flattened and passed through fully connected layers $FC(\cdot)$ to predict if the box contains a cell:

$$\hat{Y} = FC(Flatten(\mathcal{F}_{sa}))$$

Table 1. Comparison of AC segmentation performance.

Methods	Swin-Unet [5]	UNet++ [32]	UNet [23]	nnUNet [11]	FoF
IoU (%)	95.88	95.63	95.98	96.23	**96.53**
Dice (%)	97.10	97.54	97.86	97.92	**98.07**

During training, we construct samples by cropping bounding boxes from annotated cell regions (positive samples) and non-cell areas (negative samples), maintaining an imbalanced sampling ratio $\mathcal{N}^{pos} : \mathcal{N}^{neg}$. The network is optimized using cross-entropy loss:

$$\mathcal{L} = -\frac{1}{\mathcal{N}^b} \sum_{i=1}^{\mathcal{N}^b} \sum_{c=1}^{2} Y_{i,c} \log(\hat{Y}_{i,c})$$

where $\mathcal{N}^b$ is the batch size, i represents the index of each sample in a batch, $Y_{i,c}$ represents the ground truth label for class c, and $\hat{Y}_{i,c}$ is the predicted probability. Training terminates if validation loss shows no improvement for $\mathcal{N}^{patient}$ consecutive epochs.

3 Experiment

3.1 Datasets and Evaluation Metrics

We evaluated MCD using 1,376 high-resolution AS-OCT images (1598 × 1465 pixels) from the Imaging in Childhood Uveitis studies [2,9,26]. Six ophthalmic clinicians annotated the data via the Citizen Science project [13] on the Zooniverse[2], with final verification by a senior ophthalmologist. The dataset was split into 630 images for AC segmentation and 746 for cell detection evaluation.

We annotated 630 images for AC segmentation, using Intersection over Union (IoU) and the Dice coefficient as the metrics. We annotated 746 images for cell detection. Clinicians annotated by clicking cells, generating 10×10 pixel ground truth boxes centered at clicked points, each occupying <0.005% of image area.

We use two matching criteria for a correct cell detection: (1) bounding box containing a ground truth point (denoted as "point") and (2) IoU > 30% with ground truth box (denoted as "30"). The low IoU threshold accounts for significant IoU decrease in small object detection [8]. Each ground truth cell is matched only once to prevent double-counting.

To evaluate cell counting accuracy, we calculate two Mean Absolute Error (MAE) metrics across all images: MAE^{all} measures the average absolute difference between predicted and ground truth cell counts across all images, while MAE^{c} considers only correctly detected cells. To evaluate spatial localization accuracy of detection boxes, we calculate precision, recall, and F1-score based on the two matching criteria. All experiments used 40/10/50% training/validation/testing splits, repeated 5 times with different random splits.

[2] https://www.zooniverse.org/.

Table 2. Comparison Between Predicted and Ground Truth Cell Counts. † indicates methods that process only 300 × 300 image patches instead of the whole image.

Methods	MAE^{all}	MAE^{c}_{point}	MAE^{c}_{30}
† Faster-RCNN [21]	2.49	1.03	1.15
† Cascade-RCNN [4]	2.45	0.86	0.97
† RetinaNet [17]	1.74	1.12	1.24
† RepPoint [30]	2.88	1.18	1.29
Otsu [20]	1.59	1.59	1.69
Isodata [22]	1.57	1.56	1.66
MCD (Ours)	**0.85**	**0.72**	**0.84**

3.2 Baselines and Implementation Details

In the I2ACP, we generated two prompt points at offsets $(\delta_1^w, \delta_1^h) = (0, \delta' \times W)$ and $(\delta_2^w, \delta_2^h) = (0, -\delta' \times W)$ from the centroid of the AS. We used grid search over δ' [0–0.5] and $\mathcal{R}$ [0.4–0.95] in 0.05 steps on the validation set, yielding optimal values of $\delta' = 0.1$ and $\mathcal{R} = 0.65$. In the MiRP, λ was searched from 0.70 to 1.00 in 0.01 steps on the validation set. As λ grows, MCD's precision rises and recall declines, with F1 exhibiting a rise-peak-decline pattern. The optimal λ was selected based on the F1 peak and empirically set as 0.83. We set $\mathcal{S}_{max} = 25$ based on clinical observations that larger objects are unlikely to be cells [9,19], and $\mathcal{S}_{min} = 1$ to include all potential cell candidates. Both w and h were set as 10. For our Cell Mamba, $\mathcal{N}^{pos}$, $\mathcal{N}^{neg}$, $\mathcal{N}^{patient}$, and $\mathcal{N}^{b}$ were set to 1, 5, 30 and 128, respectively.

We compared FoF against SOTA segmentation baselines (UNet [23], UNet++ [32], nnUNet [11], Swin-Unet [5]) and evaluated cell detection against both threshold-based and DL-based approaches, considering only cells within segmented AC regions for a fair comparison.

Threshold-based methods include Otsu [20] and Isodata [22] (widely used in AS-OCT [1,3,9,19,28]). Both methods identify connected components (removing >25 and <2 pixels [9,19]), with results converted to 10×10 pixel boxes at centroids. For DL-based methods, we evaluated Faster-RCNN [21], Cascade-RCNN [4], RetinaNet [17], RepPoint [30], DETR [6], Deformable-DETR [33], DINO [31], and YOLO11 [12]. Our initial attempts to train these models on full-resolution images failed to detect any cells as cells are miniscule, occupying less than 0.005% of the image area. Then, we divided images into 300×300 patches for training/inference, with results mapped back and standardized to 10×10 boxes. For brevity, we report only models achieving F1 scores >55%.

4 Results and Discussion

Table 1 shows FoF, with zero-shot training, achieves a high IoU of 96.53% and a Dice coefficient of 98.07%, results comparable with the baseline models that require training.

Table 3. Cell detection results on Precision (%), Recall (%) and F1 (%). † indicates methods that process only 300 × 300 image patches instead of the whole image.

Methods	Presicion_{point}	Recall_{point}	F1_{point}	Presicion_{30}	Recall_{30}	F1_{30}
† Faster-RCNN [21]	52.27	78.56	62.77	50.56	75.99	60.72
† Cascade-RCNN [4]	54.93	82.11	65.82	53.33	79.71	63.9
† RetinaNet [17]	59.13	76.66	66.76	57.16	74.11	64.54
† RepPoint [30]	47.41	75.47	58.24	45.95	73.14	56.44
Otsu [20]	**91.22**	67.35	77.48	**88.37**	65.25	75.06
Isodata [22]	90.43	68.03	77.65	87.72	65.99	75.32
MCD	85.10	**84.79**	**85.01**	82.48	**82.02**	**82.32**

Table 2 shows that our MCD achieves the lowest cell counting error across all metrics, with MAE^{all}, MAE^{c}_{point}, MAE^{c}_{30} of 0.85, 0.72, 0.84, respectively. This means that, on average, our MCD's cell count differs from the ground truth by less than one cell per image, indicating highly accurate quantification of inflammatory cells.

Table 3 shows that Our MCD achieves the highest F1 scores across all criteria: 85.01% for F1_{point}, and 82.32% for F1_{30}. Our MCD improves F1 by around 7% over the best baseline. This significant improvements demonstrates our MCD's ability to effectively detect cells while maintaining a low false detection rate.

Previous thresholding detection methods used in anterior uveitis studies [1,3,9,14,15,19,24,25] reported strong correlations between their detected cell counts and the clinical severity score of anterior uveitis. However, these studies proceeded without verifying the correctness of individual cell detections. Our evaluation reveals two major limitations in these methods: cell count inaccuracy (as shown in Table 2) and imprecise spatial localization of detected cells (as shown in Table 3). While these methods demonstrated correlations with clinical scores, these correlations were likely based on incomplete and spatially inaccurate cell detection. This imprecision could negatively impact future research that requires accurate cell counting or studies investigating the relationship between spatial cell distribution patterns and disease progression.

In contrast, our MCD achieves good performance in detecting the cells. It opens up new possibilities for investigating the spatial distribution of inflammatory cells within the AC, which could enhance our understanding of disease pathogenesis and progression. Such insights could contribute to the development of more personalized treatment strategies, potentially improving patient outcomes. Furthermore, our framework could serve as a valuable reference for detecting extremely small objects in other medical imaging scenarios.

Despite our MCD's superior performance in detecting minuscule cells in AS-OCT images, several limitations must be acknowledged. A challenge is the limited availability of public datasets for these tasks, making the development and release of comprehensive datasets essential for future research advances. Additionally, AS-OCT image quality can vary - artifacts and quality degradation may occur [7], potentially affecting MCD's effectiveness.

5 Conclusion

We tackle the challenge of automatically detecting cells in AS-OCT images for anterior uveitis study by proposing a detection framework named MCD. We also reveals critical limitations in previous anterior uveitis studies that relied on threshold-based methods for cell detection, suggesting potential underestimation of inflammatory cell populations. MCD not only improves the technical aspects of cell detection but also holds promise for advancing clinical care and research in ocular inflammation.

Acknowledgments. This work was supported in part by the AWS Doctoral Scholarship in Digital Innovation (awarded through the UCL Centre for Digital Innovation) and an NIHR Clinician Scientist grant (CS-2018-18-ST2-005; ALS). We are grateful to the participating children and their families and the clinicians within the Great Ormond Street and Moorfields Eye Hospital Uveitis clinics.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agarwal, A., Ashokkumar, D., Jacob, S., Agarwal, A., Saravanan, Y.: High-speed optical coherence tomography for imaging anterior chamber inflammatory reaction in uveitis: clinical correlation and grading. Am. J. Ophthalmol. **147**(3), 413–416 (2009)
2. Akbarali, S., et al.: Imaging-based uveitis surveillance in juvenile idiopathic arthritis: feasibility, acceptability, and diagnostic performance. Arthritis Rheumatol. **73**(2), 330–335 (2021)
3. Baghdasaryan, E., et al.: Analysis of ocular inflammation in anterior chamber-involving uveitis using swept-source anterior segment oct. Int. Ophthalmol. **39**, 1793–1801 (2019)
4. Cai, Z., Vasconcelos, N.: Cascade R-CNN: delving into high quality object detection. In: CVPR 2018, pp. 6154–6162 (2018)
5. Cao, H., et al.: Swin-UNet: UNet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. LNCS, vol. 13803, pp. 205–218. Springer, Cham(2022). https://doi.org/10.1007/978-3-031-25066-8_9
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
7. Chen, B., Solebo, A.L., Taylor, P.: Automated image quality assessment for anterior segment optical coherence tomograph. In: ISBI 2023, pp. 1–4. IEEE (2023)
8. Cheng, G., et al.: Towards large-scale small object detection: survey and benchmarks. IEEE Trans. Pattern Anal. Mach. Intell. (2023)
9. Etherton, K., et al.: Quantitative and qualitative assessment of anterior segment optical coherence tomography capture of disease state in childhood anterior uveitis. Br. J. Ophthalmol. **107**(7), 966–972 (2023)

10. Invernizzi, A., et al.: Objective quantification of anterior chamber inflammation: measuring cells and flare by anterior segment optical coherence tomography. Ophthalmology **124**(11), 1670–1677 (2017)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
12. Jocher, G., Qiu, J., Chaurasia, A.: Ultralytics YOLO (2023). https://github.com/ultralytics/ultralytics. Accessed June 2025
13. Jones, M.L., Spiers, H.: The crowd storms the ivory tower. Nat. Methods **15**(8), 579–580 (2018)
14. Kang, T.S., et al.: Development of fully automated anterior chamber cell analysis based on image software. Sci. Rep. **11**(1), 10670 (2021)
15. Keino, H., et al.: Automated quantitative analysis of anterior segment inflammation using swept-source anterior segment optical coherence tomography: a pilot study. Diagnostics **12**(11), 2703 (2022)
16. Kirillov, A., et al.: Segment anything. In: ICCV 2023, pp. 4015–4026 (2023)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV 2017, pp. 2999–3007 (2017)
18. Liu, Y., et al.: VMamba: visual state space model. In: NeurIPS 2024, pp. 103031–103063 (2024)
19. Lu, M., et al.: Quantitative analysis of anterior chamber inflammation using the novel CASIA2 optical coherence tomography. Am. J. Ophthalmol. **216**, 59–68 (2020)
20. Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica **11**(285–296), 23–27 (1975)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)
22. Ridler, T., Calvard, S.: Picture thresholding using an iterative selection method. IEEE Trans. Syst. Man Cybern. **8**(8), 630–632 (1978)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Rose-Nussbaumer, J., Li, Y., Lin, P., Suhler, E., Asquith, M., Rosenbaum, J.T., Huang, D.: Aqueous cell differentiation in anterior uveitis using Fourier-domain optical coherence tomography. Invest. Ophthalmol. Vis. Sci. **56**(3), 1430–1436 (2015)
25. Sharma, S., Lowder, C.Y., Vasanji, A., Baynes, K., Kaiser, P.K., Srivastava, S.K.: Automated analysis of anterior chamber inflammation by spectral-domain optical coherence tomography. Ophthalmology **122**(7), 1464–1470 (2015)
26. Solebo, A.L., Bellchambers, A., Kellett, S., Rahi, J.S., Dick, A.D.: Establishing the normative data set necessary for imaging-based childhood uveitis surveillance: a cross-sectional study. Invest. Ophthalmol. Visual Sci. **65**(1), 9 (2024)
27. Standardization of Uveitis Nomenclature (SUN) Working Group, et al.: Standardization of uveitis nomenclature for reporting clinical data. Results of the First International Workshop. Am. J. Ophthalmol. **140**(3), 509–516 (2005)
28. Uthayananthan, P., Tanwar, N., Rahi, J.S., Dick, A.D., Solebo, A.L.: Imaging-based detection of anterior chamber inflammation: a comparative diagnostic accuracy study. Am. J. Ophthalmol. **270**, 131–139 (2025)

29. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: convolutional block attention module. In: ECCV 2018, pp. 3–19 (2018)
30. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: RepPoints: point set representation for object detection. In: ICCV 2019, pp. 9657–9666 (2019)
31. Zhang, H., et al.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
32. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
33. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)



MixStyleFlow: Domain Generalization in Medical Image Segmentation Using Normalizing Flows

Reza Safdari, Mohammad-Ali Nikouei Mahani, Mohamad Koohi-Moghadam(✉), and Kyongtae Tyler Bae(✉)

The University of Hong Kong, Pokfulam, Hong Kong SAR, China
{safdari,koohi,baekt}@hku.hk

Abstract. Despite the success of deep learning in medical image segmentation, domain shifts caused by variations in scanners and imaging protocols often degrade performance, limiting real-world clinical deployment. Domain generalization (DG) aims to address this issue by learning robust models that generalize well across different domains. While existing DG methods based on feature-space domain randomization have shown promise, they suffer from a limited and unordered search space of feature styles. In this work, we propose MixStyleFlow, a novel DG approach that utilizes normalizing flows to explicitly model the distribution of domain feature styles. By sampling domain feature styles from the learned normalizing flows and mixing them with original feature statistics along the feature channel dimension, our method effectively expands and diversifies domain features in a controllable manner. We evaluate MixStyleFlow on two medical segmentation tasks—prostate MRI and fundus imaging—demonstrating superior generalization performance on unseen target-domain data. Our results highlight the potential of normalizing flows for improving domain generalization in medical image segmentation, paving the way for more robust deep learning models capable of handling diverse clinical scenarios. The code is available at https://github.com/Reza-Safdari/MixStyleFlow.

Keywords: Domain Generalization · Medical Image Segmentation · Normalizing Flows · Feature Perturbation · Feature Style Augmentation

1 Introduction

Deep learning has revolutionized medical image segmentation, enabling accurate delineation of anatomical structures and pathological regions in different imaging modalities like MRI, CT, and fundus photography [1]. However, a major challenge is the sensitivity of these models to domain shifts variations in image distributions due to different scanners, imaging protocols, patient populations, and acquisition settings [2, 3]. Models trained on specific domain data often perform poorly on unseen domains, limiting their generalizability and robustness in real-world clinical scenarios.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 376–385, 2026.
https://doi.org/10.1007/978-3-032-04947-6_36

Domain generalization (DG) addresses this challenge by learning models that can generalize well to unseen domains without requiring access to target-domain data during training [4, 5]. In medical imaging, achieving robust DG is crucial due to the heterogeneous nature of data arising from multi-institutional resources and diverse clinical practices [6]. Traditional approaches to tackle domain shifts include domain adaptation techniques, which require access to target-domain data during training [2]. However, collecting and annotating new data for each target domain is impractical and costly. Therefore, there is a pressing need for methodologies that enhance the generalization capability of segmentation models across diverse domains using only source-domain data.

Recent DG methods have explored feature-space domain randomization techniques aimed at learning domain-invariant representations by perturbing feature statistics during training [6]. Approaches such as MixStyle [7], MaxStyle [8] and their variants manipulate the style information in feature maps to simulate domain shifts, thereby encouraging the model to learn robust features. These methods operate on the assumption that content-preserving style transformations can be achieved by altering the statistics (e.g., mean and standard deviation) of the features. While these methods have shown promise results, using the statistics of source-domain data for feature perturbation may limit the search space and operate within a limited and unordered domain. Consequently, they may not fully capture the complex variations present in real-world domain shifts, potentially hindering the model's ability to generalize to unseen domains. To address these limitations, we propose MixStyleFlow, a novel domain generalization framework that utilizes the power of normalizing flows to explicitly model the distribution of domain feature styles.

2 Related Work

Feature-space domain randomization has emerged as a powerful approach for DG. Random Convolutions (RandConv) [9] introduces a domain randomization strategy by applying convolutional layers with randomly sampled weights to input images. This effectively alters local textures while preserving global shape information, encouraging models to rely more on structural cues rather than superficial textures. Dynamic Style Augmentation (DSU) [10] enhances DG by injecting Gaussian noise into feature statistics, generating diverse style variations beyond those present in the original data. This technique has shown particular effectiveness in medical image segmentation. Similarly, Treasure in Distribution (TriD) [11] expands feature perturbations by sampling statistics from a uniform distribution rather than relying solely on source-domain data. TriD further improves adaptability to domain shifts by employing a style-mixing strategy, blending original and augmented statistics along feature channels to promote domain-invariant representations.

Similarly, MixStyle [7] perturbs feature statistics by mixing instance-level mean and standard deviation from different training samples, implicitly simulating style variations across domains. MaxStyle [8] extends this idea by introducing adversarial style compositions, maximizing style diversity and generating more challenging training examples. Beyond these methods, Exact Feature Distribution Matching (EFDM) [12] introduces a more precise feature perturbation approach by explicitly aligning the empirical cumulative distribution functions (eCDFs) of image features. Unlike approaches that rely only

on first- and second-order statistics, EFDM performs exact histogram matching in feature space, ensuring a more faithful simulation of domain shifts. By offering a richer set of perturbations beyond traditional Gaussian-based transformations, EFDM significantly improves the generalization capability of segmentation models.

3 Method

3.1 Background on Normalizing Flows

Real Non-Volume Preserving (RealNVP) is a type of normalizing flow model used for density estimation and generative modeling [13]. It transforms a simple base distribution, such as Gaussian, into a complex target distribution through a sequence of invertible transformations [14]. The core idea behind RealNVP is the use of coupling layers, which enable efficient computation of both the forward transformation (sampling) and the inverse transformation (density evaluation). Each coupling layer splits the input x into two parts, x_1 and x_2, and applies an affine transformation to one part while keeping the other unchanged. The transformation is defined as:

$$\begin{aligned} x_1\prime &= x_1 \\ x_2\prime &= x_2 \odot e\mathrm{xp}(s(x_1)) + t(x_1) \end{aligned} \tag{1}$$

where $s(x_1)$ and $t(x_1)$ are scale and translation functions, typically parameterized by neural networks. The symbol $\odot$ represents the elementwise (Hadamard) product. The inverse transformation, crucial for likelihood estimation, is easily computable as:

$$x_2 = (x_2\prime - t(x_1)) \odot e\mathrm{xp}(-s(x_1)) \tag{2}$$

Since the transformation is invertible with a tractable Jacobian determinant, the log-likelihood of a sample x under the model is computed using the change of variables formula:

$$\log p_X(x) = \log p_Z(f(x)) + \sum_i \log \left| \det \frac{\partial f_i}{\partial x} \right| \tag{3}$$

where $f(x)$ maps the input to the latent space, $p_Z(z)$ is the probability density of the transformed sample under a simple prior distribution (e.g., Gaussian), and $\det \frac{\partial f_i}{\partial x}$ is the determinant of the Jacobian of the transformation. The loss function, derived from the negative log-likelihood (NLL) over N input samples $x_1, x_2, \ldots, x_N$, is given by:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log p_X(x_i) \tag{4}$$

Minimizing this loss ensures the model effectively learns an expressive transformation that maps the data distribution to the Gaussian latent space while maintaining invertibility and efficient density estimation.

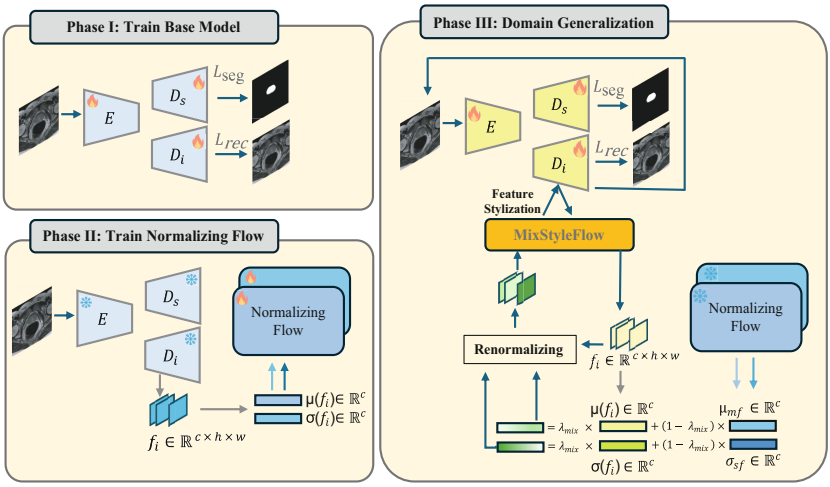


Fig. 1. The proposed MixStyleFlow framework. **Phase I** involves training the segmentation model on source dataset without domain generalization techniques. Subsequently, in **Phase II**, feature statistics (feature means and standard deviations) are extracted, and two normalizing flow models are trained to learn their distributions. In **Phase III**, the segmentation model is trained on source data and stylized images created through style feature augmentation at specific decoder layers. This augmentation is achieved by interpolating original feature statistics with in-domain and out-of-domain statistics sampled from the trained normalizing flow models.

3.2 Encoder with Dual-Decoder Architecture for Segmentation

We employed a dual-branch network presented in [8], consisting of one encoder (E) and two decoders. One decoder, called the segmentation decoder (D_s), is trained using segmentation loss to generate the semantic mask, while the other decoder, called the image decoder (D_i), is trained using reconstruction loss, allowing it to exploit both complementary image content features and task-specific shape features for the segmentation task Fig. 1. We also utilize the image decoder to apply style augmentation by randomizing the statistics (i.e., mean and standard deviation) of the feature maps in specific layers (the exact randomization method will be discussed in the next section). The reconstructed image from this decoder enables us to interpret and visualize the impact of the augmented features, ensuring that the randomization preserves their semantic and anatomical structures. For the segmentation loss L_{seg}, we use a combination of Binary Cross-Entropy (BCE) Loss L_{BCE} and Dice Loss L_{Dice}:

$$\begin{aligned} L_{\text{seg}} &= L_{\text{BCE}} + L_{\text{Dice}} \\ L_{\text{BCE}} &= -\sum_i \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right] \\ L_{\text{Dice}} &= 1 - \frac{2 \sum_i y_i \hat{y}_i + \epsilon}{\sum_i y_i + \sum_i \hat{y}_i + \epsilon} \end{aligned} \tag{5}$$

here yi and $\hat{y}_i$ represent the ground truth and predicted values, respectively, and ϵ is a small constant added for numerical stability. For the reconstruction loss, we use $l2$ norm as follows:

$$L_{\text{rec}} = \sum\nolimits_i |I_i - \hat{I}_i|^2 \tag{6}$$

here I_i is the original input image and $\widehat{I}_i$ is the reconstructed image produced by the image decoder.

3.3 Proposed Method: MixStyleFlow for Domain Generalization

Our proposed MixStyleFlow framework consists of three phases that are presented in Fig. 1. In **Phase I**, we train the model on source datasets for segmentation using the architecture in Sect. 3.2, without domain generalization techniques. After training, in **Phase II**, we freeze its parameters and extract channel-wise feature statistics from the decoder. We then train two normalizing flow models (MuFlow for means, STDFlow for standard deviations) as described in Sect. 4.2, to learn their distributions. This enables (1) in-domain sampling to enrich the style space and (2) out-of-distribution sampling to expand it beyond the original dataset.

In **Phase III**, we train the segmentation architecture on the source data and stylized images generated by applying style feature augmentation at specific image decoder layers. Formally, let $f_i \in \mathbb{R}^{c\times h\times w}$ be c-dimensional feature maps extracted at a certain layer in the image decoder (D_i) for image x_i, where h and w represent the height and width of the feature maps, respectively. MixStyleFlow performs style augmentation for f_i by first normalizing it with its channel-wise means and standard deviations $\mu(f_i), \sigma(f_i) \in \mathbb{R}^c$: $\bar{f}_i = (f_i - \mu(f_i))/\sigma(f_i)$ and then transforming the normalized feature with a linear combination of its channel-wise statistics $\{\mu(f_i), \sigma(f_i)\}$ and sampled statistics from learned distributions $\{\mu_{mf}, \sigma_{sf}\}$ as follows:

$$\begin{gathered} \text{MixStyleFlow}(f_i) = \gamma_{\text{mix}} \odot \bar{f}_i + \beta_{\text{mix}} \\ \gamma_{\text{mix}} = \lambda_{\text{mix}}\sigma(f_i) + (1-\lambda_{\text{mix}})\sigma_{sf},\ \beta_{\text{mix}} = \lambda_{\text{mix}}\mu(f_i) + (1-\lambda_{\text{mix}})\mu_{mf} \\ \sigma_{sf} \in \mathbb{R}^c \sim \textit{STDFlow},\ \mu_{mf} \in \mathbb{R}^c \sim \textit{MuFlow} \end{gathered} \tag{7}$$

here λ_{mix} is a coefficient used for applying interpolation between original statistics and sampled statistics, randomly sampled from [0, 1].

4 Experiments and Results

4.1 Datasets and Evaluation Metrics

To evaluate the effectiveness of MixStyleFlow in domain generalization for medical image segmentation, we conducted experiments on two segmentation tasks: prostate MRI segmentation and optic disc/optic cup (OD/OC) segmentation in fundus images. These datasets include multiple domains representing different imaging centers, scanners, and acquisition protocols, ensuring a diverse and challenging evaluation settings. For prostate MRI segmentation, We use the Medical Decathlon [15] for training and intra-domain testing, while the remaining six sites from [16] are used for testing. We followed [17] to preprocess MRI volumes, retaining only prostate slices, resizing them to 384 × 384, and using them as 2D training data. For OD/OC segmentation, we used color fundus images from five different domains [18–21], covering a total of 1,441 images, training on one source domain and evaluating the model on the remaining domains (Table 1). We utilized the Dice Similarity Coefficient (DSC) for evaluating segmentation performance.

Table 1. Overview of the datasets used for training and evaluating MixStyleFlow

Task	Modality	Domain Names	Cases in each Domain
Prostate Segmentation	MRI	Medical Decathlon, Site A–F	32 30; 30; 19; 13; 12; 12
OD/OC Segmentation	Fundus Image	Domain1–5	195; 95; 400; 650; 101

4.2 Implementation Details

We trained three pairs of normalizing flows (NFs) to model the distributions of feature means and standard deviations from image decoder layers 2, 3, and 4, using temperatures of 3.5, 2.5, and 1.5, respectively, for 100 epochs. These decoder layers were selected experimentally based on preliminary results. Each NF model comprises four masked affine coupling flows with ActNorm layers [22]. The coupling network included two MLPs for scale and translation, with layer dimensions [32, 64, 32], [16, 32, 16], and [16, 32, 16] for image decoder layers 2, 3, and 4, respectively, capturing complex feature relationships. The segmentation model was trained with a mini-batch size of 20 using the AdamW optimizer. The learning rate was 0.0001 for prostate segmentation and 0.001 for OD/OC segmentation. The training steps 1,600 epochs for the prostate and 100 epochs for the OD/OC segmentation. We implemented our approach in PyTorch on a single NVIDIA RTX 6000 GPU.

4.3 Results

Comparing with Other DG Methods. We evaluated MixStyleFlow on both intra-domain (IID) and cross-site prostate segmentation tasks (Table 2). The results demonstrate the effectiveness of our approach in improving generalization across unseen domains. For the IID, MaxStyle (87.27) and MixStyleFlow (86.47) showed competitive performance, indicating their ability to maintain segmentation accuracy within the source domain. For cross-site generalization, our proposed MixStyleFlow consistently outperformed other methods across multiple unseen domains (Sites A–F), demonstrating its robustness to domain shifts. Compared to MaxStyle, which also performed well on some sites, MixStyleFlow exhibited superior generalization, particularly in more challenging domains such as Site E (68.5 vs. 58.43) and Site F (74.38 vs. 57.72).

To further assess the generalization capability of MixStyleFlow, we evaluated it on the joint segmentation of the optic disc (OD) and optic cup (OC) in fundus images across multiple domains (Table 3). MixStyleFlow consistently outperformed other methods in most unseen domains. Especially, in comparison with the second-best performing method, TriD, our model gets the best score in all most domains with comparable results on Domain 5.

Table 2. Evaluation results (Dice scores) on the prostate intra-domain and unseen cross-site test sets. The 'Avg' column represents the average Dice score across the unseen domains (Site A–F). The best results are highlighted with **bold**.

Methods	IID	A	B	C	D	E	F	Avg
Baseline	84.94	87.94	71.38	79.74	79.61	50.18	54.17	70.5
MixStyle	84.03	88.98	54.38	81.54	84.05	63.69	62.95	72.6
DSU	83.11	88.21	59.24	82.24	83.04	58.56	64.03	72.55
MaxStyle	**87.27**	90.20	**83.94**	83.28	85.8	58.43	57.72	76.56
MixStyleFlow	86.47	**90.25**	83.75	**84.06**	**85.85**	**68.5**	**74.38**	**81.13**

Table 3. Evaluation results for joint OD and OC segmentation in fundus images. Each cell shows the Dice scores for OD (first) and OC (second). The model is trained on one domain and evaluated on the others. The best results are in **bold**.

Methods	Domain1	Domain2	Domain3	Domain4	Domain5	Avg
SAN-SAW	76.42, 59.01	83.79, 73.23	84.17, 65.51	81.83, 62.36	87.00, 64.42	82.64, 64.91
RandConv	79.63, 64.14	85.00, 72.40	87.77, 69.57	83.08, 64.38	86.31, 60.37	84.36, 66.17
MixStyle	75.67, 60.84	86.35, 73.77	85.86, 66.60	84.86, 66.44	86.54, 65.99	83.36, 66.73
EFDM	78.79, 57.73	84.83, 72.30	85.25, 65.94	82.13, 61.62	85.43, 63.02	83.29, 64.12
DSU	76.88, 61.26	84.17, 74.10	89.12, 70.16	83.53, 63.19	87.09, 59.65	84.16, 65.67
MaxStyle	77.40, 65.44	86.95, 74.52	87.95, 67.62	84.69, 66.05	87.95, 64.84	84.99, 67.69
TriD	81.86, 66.67	**88.19**, 75.43	89.62, 70.85	84.81, 67.53	**87.88, 66.96**	86.47, 69.49
MixStyleFlow	**86.16, 66.99**	88.16, **75.64**	**90.58, 75.74**	**87.17, 74.91**	86.09, 62.97	**87.63, 71.25**

Low Data Regime. To evaluate performance in low-data settings, we trained our model with 10% and 30% of the prostate dataset [15] and assessed results across different domains [16]. As shown in Fig. 2, MixStyleFlow consistently outperformed MaxStyle across most domains. With only 10% training data (Fig. 2.A), MixStyleFlow achieved notable gains in domain B (38.80% vs. 24.93%) and E (20.31% vs. 6.96%), demonstrating robustness in extremely low-data scenarios. Increasing the training set to 30% improved both methods (Fig. 2.B), but MixStyleFlow remained superior, particularly in domain B (72.40% vs. 64.18%) and E (34.32% vs. 15.46%). These results highlight MixStyleFlow's strong generalization even with limited supervision.

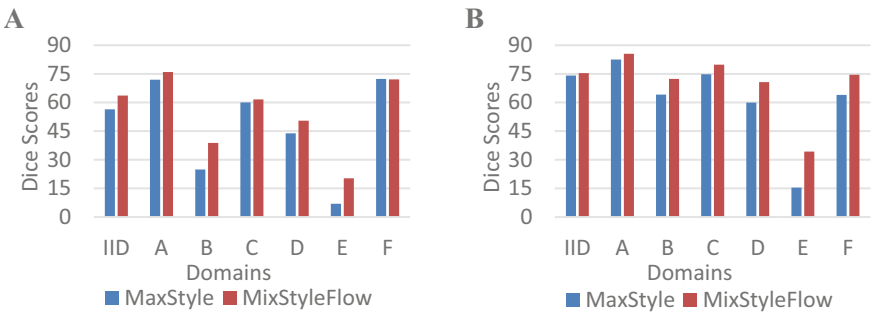


Fig. 2. Dice score evaluation on prostate intra-domain and cross-site test sets: (A) results with 10% training data; (B) results with 30% training data.

5 Conclusion

In this work, we introduced MixStyleFlow, a novel domain generalization framework that utilizes normalizing flows to model and perturb feature styles in a structured and expressive way. Unlike existing feature-space randomization techniques with limited and unordered perturbations, MixStyleFlow explicitly captures the distribution of domain feature styles, enabling controllable and diverse style augmentations. Our experiments on prostate MRI and optic disc/cup segmentation in fundus images demonstrated its superior generalizability to unseen domains. While we expect our proposed MixStyleFlow method will perform well with a high degree of generalizability in segmenting anatomical structures in various medical imaging applications, the computational overhead of normalizing flows may affect training efficiency compared with simpler perturbation methods. Future work could improve efficiency with optimized flow architectures and extend the approach to more imaging modalities and tasks.

Acknowledgments. This work was conducted in the JC STEM Lab of Innovative Medical Imaging Research funded by The Hong Kong Jockey Club Charities Trust.

Disclosure of Interests. The authors declare no competing financial or non-financial interests.

References

1. Azad, R., et al.: Medical image segmentation review: the success of u-net. IEEE Trans. Pattern Anal. Mach. Intell. **46**, 10076–10095 (2024)
2. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Trans. Biomed. Eng. **69**(3), 1173–1185 (2021)
3. Xie, X., et al.: A survey on incorporating domain knowledge into deep learning for medical image analysis. Med. Image Anal. **69**, 101985 (2021)
4. Wang, J., et al.: Generalizing to unseen domains: a survey on domain generalization. IEEE Trans. Knowl. Data Eng. **35**(8), 8052–8072 (2022)

5. Zhou, K., et al.: Domain generalization: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4396–4415 (2022)
6. Niu, Z., et al.: A survey on domain generalization for medical image analysis. arXiv preprint arXiv:2402.05035 (2024)
7. Zhou, K., et al., Domain Generalization with MixStyle. In: Proceedings of Ninth International Conference on Learning Representations, pp. 1–15 (2021)
8. Chen, C., et al.: Maxstyle: adversarial style composition for robust medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Heidelberg (2022)
9. Xu, Z., et al.: Robust and generalizable visual representation learning via random convolutions. arXiv preprint arXiv:2007.13003 (2020)
10. Li, X., et al.: Uncertainty modeling for out-of-distribution generalization. In: International Conference on Learning Representations (2022)
11. Chen, Z., et al.: Treasure in distribution: a domain randomization based multi-source domain generalization for 2d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Heidelberg (2023)
12. Zhang, Y., et al.: Exact feature distribution matching for arbitrary style transfer and domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
13. Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. arXiv preprint arXiv:1605.08803 (2016)
14. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International Conference on Machine Learning. PMLR (2015)
15. Antonelli, M., et al.: The medical segmentation decathlon. Nat. Commun. **13**(1), 4128 (2022)
16. Liu, Q., et al.: Feddg: federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
17. Hu, S., et al.: Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation. IEEE Trans. Med. Imaging **42**(1), 233–244 (2022)
18. Almazroa, A., et al.: Retinal fundus images for glaucoma analysis: the RIGA dataset. In: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications. SPIE (Antonelli, M., et al., *The medical segmentation decathlon.* Nature communications, 2022. **13**(1): p. 4128)
19. Orlando, J.I., et al.: Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image Anal. **59**, 101570 (2020)
20. Sivaswamy, J., et al.: Drishti-gs: retinal image dataset for optic nerve head (onh) segmentation. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI). IEEE (2014)

21. Zhang, Z., et al.: Origa-light: an online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. IEEE (2010)
22. Kingma, D.P., Dhariwal, P.: Glow: generative flow with invertible 1x1 convolutions. Adv. Neural Inf. Process. Syst. **31** (2018)

MoDiff: A Morphology-Emphasized Diffusion Model for Ambiguous Medical Image Segmentation

Jung Su Ahn[1], Ki Hoon Kwak[2], Jung Woo Seo[2], and Young-Rae Cho[3,4(✉)]

[1] Graduate School of Computer Science, Yonsei University Mirae Campus, Wonju-si, Gangwon-do 26493, Republic of Korea
2023312031@yonsei.ac.kr

[2] Department of Computer Science and Engineering, Yonsei University Mirae Campus, Wonju-si, Gangwon-do 26493, Republic of Korea
{kihoon090,tjwjddn15584}@yonsei.ac.kr

[3] Department of Software, Yonsei University Mirae Campus, Wonju-si, Gangwon-do 26493, Republic of Korea

[4] Department of Digital Healthcare, Yonsei University Mirae Campus, Wonju-si, Gangwon-do 26493, Republic of Korea
youngcho@yonsei.ac.kr

Abstract. MoDiff is a morphology-emphasized diffusion model designed for ambiguous medical image segmentation. It replaces traditional one-hot encoding with probability-based label maps to capture inherent uncertainties and ensure consistent segmentation results. By determining the presence of individual radiologist labels, MoDiff enables diverse sampling that provides richer insights into ambiguous areas. Its Learnable Discrete Frequency Filter (LDF) extracts high-frequency details for improved boundary precision, and when integrated with the Morphology-based Cross Attention Network (MCA), it enhances feature synthesis for more accurate anatomical segmentation. Evaluations on the LIDC-IDRI and MS-MRI datasets confirm its superior accuracy, boundary precision, and consistency.

Keywords: Attention Mechanism · Diffusion · Segmentation · Fast Fourier Transform

1 Introduction

In recent years, probabilistic segmentation models–such as the Probabilistic U-Net [14], Conditional Variational Autoencoders (cVAE) [1], and diffusion models [4,11] have shown promise in addressing uncertainties in medical image analysis by generating multiple plausible segmentation hypotheses. However, sampling repeatedly from learned probability distributions can lead to inconsistent or overly varied outputs, complicating precise delineation of critical structures like

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 386–395, 2026.
https://doi.org/10.1007/978-3-032-04947-6_37

organ boundaries. This inconsistency may result in conflicting clinical interpretations, especially in tasks such as tumor diagnosis or organ segmentation. Moreover, training by randomly selecting reference labels fails to utilize all available annotation information.

To overcome these limitations, we propose the Morphology-Emphasized Diffusion Model for Ambiguous Medical Image Segmentation (MoDiff). MoDiff processes ambiguous medical images using a diffusion-based probabilistic segmentation approach that emphasizes morphological consistency. Unlike traditional methods using one-hot encoded label maps, we train the model with probability distribution-based label maps, enabling it to capture the inherent variability in medical annotations and produce consistent segmentation results across multiple samples.

Furthermore, to effectively extract morphological features and enhance noise reduction during the reverse diffusion process, we introduce the Learnable Discrete Frequency Filter (LDF) and the Morphology-based Cross Attention Network (MCA). LDF detects subtle boundary details and filters high-frequency noise using learnable parameters, while MCA synthesizes the derived features into a robust condition for denoising. This combined approach facilitates faster and more accurate learning of morphological structures compared to traditional methods.

2 Related Work

2.1 Stochastic Segmentation for Medical Imaging

Medical imaging inherently presents uncertainties due to device noise, low resolution, and anatomical variations. Stochastic segmentation techniques address these challenges by modeling uncertainty in the latent space to generate diverse segmentation outputs [3]. Notably, the Probabilistic U-Net [14] combines the U-Net architecture with a cVAE to produce multiple plausible segmentation hypotheses, offering more consistent results than pixel-wise probability models. Furthermore, cFlow [5] utilizes normalizing flows to extend simple Gaussian distributions into complex latent representations, while MoSE [6] and CIMD [2] capture uncertainty through multi-modal Gaussian and diffusion-based approaches, respectively.

2.2 Conditional Diffusion Model

Conditional Diffusion Models (CDMs) guide the denoising process with additional conditions to generate high-quality, condition-specific outputs [10]. BerDiff [8] replaces continuous Gaussian noise with the Bernoulli distribution to better handle discrete segmentation tasks. MedSegDiff [9] further enhances performance by integrating Dynamic Conditional Encoding and FFT-based noise suppression into a U-Net framework. Additionally, CCDM [7] employs categorical distributions to maintain clear label boundaries, resulting in more precise segmentation outcomes.

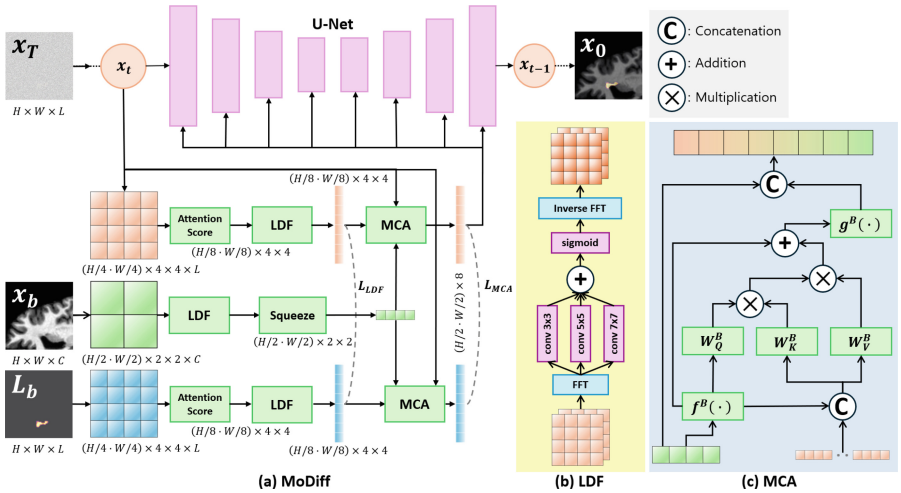


Fig. 1. a) This is a graphical model representing the overall reverse noise process in the training phase of MoDiff. During training, at each step t, the input image b is added to the noisy segmentation label x_t and used as the input image. Using the original label x'_t, b) LDF performs morphological structure extraction and high-frequency noise filtering, and c) the MCA, an attention-based network generating the final conditional image, calculates the loss.

3 Method

3.1 Morphology-Based Cross-Attention Network

The image $x_b \in \mathbb{R}^{H \times W \times C}$ is paired with segmentation labels $l_{b,i}$ (with i denoting multiple independent labels). Instead of randomly selecting a single label for training–which would require i times more iterations–we compute an average probability map L_b by averaging the i labels at each pixel. This improves training efficiency and result stability.

In conditional diffusion models, the condition directly affects performance, yet medical images often exhibit ambiguous boundaries and low contrast (e.g., in MRI or CT), making it difficult to preserve anatomical structure. To address this, we introduce MCA, which enhances morphological feature extraction. Traditional methods (like Canny or Sobel filters) fall short in capturing precise organ contours, and denoising can further degrade label maps.

Our approach learns a high-frequency filter in the frequency domain tailored to each image, enabling effective feature extraction with fewer denoising steps. Specifically, the original image b is divided into 2×2 patches, while the sampled label x_t is subdivided into 4×4 patches. This subdivision reduces high-frequency noise variance, as each patch in b maps to four patches in x_t through linear transformation and positional encoding. The linear mappings used in MCA are defined as follows:

$$Q = BW_Q, \quad K = SW_K, \quad V = SW_V, \tag{1}$$

where $B = \{x_b^1, x_b^2, \ldots, x_b^n\}$ is the set of patches of the original image, and$S$ is the set of patches of the sampled label with LDF applied. W_Q, W_K, and W_V are learnable projection matrices. Attention scores calculated for each small patch are concatenated to form $S = [x_t^1, x_t^2, \ldots, x_t^m]$. The queries and keys are linearly projected vectors, $Q \in \mathbb{R}^{H/2 \cdot W/2 \cdot 4}$, $K \in \mathbb{R}^{H/4 \cdot W/4 \cdot 16}$, and $V \in \mathbb{R}^{H/4 \cdot W/4 \cdot 16}$. Sinusoidal positional encoding is also added to represent the fixed position of each patch. The attention score $S' = [x'^1_t, x'^2_t, \ldots, x'^m_t]$ for each key and value pair is calculated as follows:

$$x'^m_t = \frac{(x_t^m W_q)(x_t^m W_k)^\top}{\sqrt{d}}, \tag{2}$$

where x'^m_t is the attention score embedded for the m-th label patch, W_q and W_k are learnable weight matrices, and d is the dimension of the key vector K. Softmax normalization is omitted to maximize the contrast between noise values and actual label values, enhancing noise removal performance. The MCA encoder E, utilizing the computed S and B, operates as follows:

$$y^B = [f(B) \,\|\, S'], \tag{3}$$

$$O^B = f^B(\text{Squeeze}(\text{LDF}(B))) + \text{MSA}\big(\text{LN}\big(\text{LDF}(y^B)\big)\big), \tag{4}$$

$$z^B = [g^B(O^B) \,\|\, B], \tag{5}$$

where f^B is an embedding function, g^B is a mapping function, and $\|$ denotes vector concatenation. LN represents layer normalization, and MSA is multi-head self-attention. MSA additionally applies softmax normalization to the formula for calculating the attention score of S, and the denominator includes multiplied by the number of heads h. In this study, the number of heads used is 4. The graphical model of proposed approach is illustrated in Fig. 1.

3.2 Learnable Discrete Frequency Filter

To further enhance morphological feature extraction and effectively remove noise, we introduce the LDF into MCA. LDF transforms MCAs output into the frequency domain, applies convolution and a sigmoid function to create a high-frequency filter, and then transforms it back to the spatial domain for noise removal. The LDF process is defined as follows:

$$\text{LDF}(y^B) = \mathcal{F}^{-1}\left(\sigma\left(\mathcal{F}(y^B) * H\right)\right), \tag{6}$$

where y^B is the input or intermediate output of the MCA encoder, representing the embedding vectors of the original image and sampled label patches. $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the FFT and its inverse, respectively. H is a learnable high-frequency filter, which undergoes convolution and is normalized with a sigmoid function σ. The convolution is applied hierarchically with filter sizes of 3×3, 5×5, and 7×7, and a stride value is used to preserve the shape of the input image. The calculated attention score is used as a condition for each reverse diffusion process and is concatenated with x_t for training. To prevent overfitting, the original image is also concatenated to the generated c_t, which is then used to infer the label for the next step.

3.3 Training

The diffusion model's primary loss, $\mathcal{L}_{\text{simple}}$, is defined as the mean squared error between the predicted noise ϵ_θ and the actual noise ϵ. However, this loss alone does not suffice for low-contrast, single-channel medical images. Therefore, auxiliary loss functions based on LDF and MCA are introduced.

For LDF, to ensure the generated feature map preserves the original image's morphological details, we add an auxiliary loss:

$$\mathcal{L}_{\text{LDF}} = \mathbb{E}\left[\left\|\text{LDF}(y^B) - y^B_{\text{target}}\right\|^2\right], \tag{7}$$

where y^B_{target} is the ideal morphological feature map. This term drives LDF to produce feature maps that closely match the desired structural details.

For MCA, an attention regularization loss, $\mathcal{L}_{\text{MCA}}$, is introduced to control the distribution of attention scores. This helps prevent bias toward specific patches and ensures the attention mechanism effectively learns the relationship between the original image and its sampled label.

The overall loss function is a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{simple}} + \lambda_1 \mathcal{L}_{\text{LDF}} + \lambda_2 \mathcal{L}_{\text{MCA}}, \tag{8}$$

with $\lambda_1 = \lambda_2 = 0.5$. This configuration improves noise removal while enhancing morphological feature extraction and balanced attention learning. During training, the model parameters are updated via backpropagation to minimize $\mathcal{L}_{\text{total}}$, thereby simultaneously boosting denoising capability and segmentation performance.

4 Experiments

4.1 Dataset and Experimental Setup

Lung Image Database Consortium Image Collection (LIDC-IDRI). We used the LIDC-IDRI dataset [12], comprising 1,018 low-dose chest CT scans from 1,010 subjects. Lung nodules were annotated by 12 radiologists, with each nodule labeled by four radiologists based on location, size, shape, and malignancy. The dataset provides 128×128 pixel 2D patches centered on each nodule. The training set includes 14,000 lesion images, while the test set has 1,096 images.

Multiple Sclerosis Lesion Segmentation (MS-MRI). This dataset [13] contains longitudinal MRI scans from five subjects focused on white matter lesions in multiple sclerosis. Two radiologists independently annotated the lesions, and each slice includes four modalities: proton density, FLAIR, MP-RAGE, and T2-weighted images. It is split into 5,661 training and 767 test lesion images, with preprocessed 128×128 pixel patches centered on the lesions.

Implementation Details. The proposed method was implemented using the PyTorch framework and a 4-way RTX 3090 setup. For CCDM and our model, we set the time step $T = 250$, while for other diffusion models, the time step was set to $T = 1000$ with a linear noise schedule. The optimizer used was Adam, with a learning rate of $1e^{-4}$ for MoDiff. Additionally, the scheduler was set to ReduceLROnPlateau with a factor of 0.2 and patience of 10. For the comparison models, we used the parameters provided in their respective published code or papers. For all models, the binary masks for each image were generated by thresholding the predicted probability maps at 0.9, retaining only the pixels with a probability of 0.9 or higher.

Evaluation Metrics. We evaluate performance using several metrics: GED [15–17] and HM-IoU [18,19] assess distribution differences between generated and ground-truth label maps, NCC [20] measures image similarity, CI [2] synthesizes various evaluation criteria. Each metric is computed over n samples (e.g., GED_n), with larger n yielding more precise estimates.

4.2 Comparison with the Baseline Methods

Quantitative Comparison. As shown in Table 1, 2, Our model consistently achieves lower GED_n and higher HM-IoU_n scores compared to other models, indicating improved performance with lower GED_n and higher HM-IoU_n values. Additionally, MoDiff demonstrates competitive performance in terms of NCC_n and CI_n scores against state-of-the-art models, where higher NCC_n and CI_n values reflect better performance. Overall, these results demonstrate MoDiff not only aligns more closely with the ground truth label distribution but also maintains high structural similarity and diversity in the generated samples, outperforming baseline methods across multiple evaluation metrics. Furthermore, performance improves for all models when using 32 sampled labels with different noise levels compared to 16 sampled labels. This suggests the model's diversity and ability to generate labels over a wider area contribute to enhanced performance.

Qualitative Comparison. Figure 2 illustrates visual comparisons between the proposed MoDiff model and other baseline methods. As depicted, MoDiff captures finer details and exhibits smoother boundaries, particularly when dealing with complex structures, outperforming other models in terms of visual fidelity. This qualitative improvement complements the quantitative results and underscores the effectiveness of MoDiff in generating high-quality samples closely resembling ground truth images.

Table 1. Quantitative comparison of results with state-of-the-art ambiguous segmentation networks in terms of GED, HM-IoU, NCC and CI on the LIDC-IDRI dataset. The best results are shown in bold, achieving state-of-the-art performance across all evaluation scores. Additionally, cases with a higher number of samples show relatively better performance.

Method	GED_{16}	$HM\text{-}IoU_{16}$	NCC_{16}	CI_{16}	GED_{32}	$HM\text{-}IoU_{32}$	NCC_{32}	CI_{32}
Prob.Unet [14]	0.3050	0.5572	0.4795	0.7312	0.2998	0.5501	0.4810	0.7366
cFlow [5]	0.2252	0.5802	0.5234	0.6846	0.2258	0.5857	0.5289	0.6952
MoSE [6]	0.2176	0.6202	0.3900	0.7258	0.2096	0.6225	0.3987	0.7280
CIMD [2]	0.2198	0.6125	0.4541	0.7482	0.2109	0.6184	0.4602	0.7543
CCDM [7]	0.1690	0.7922	0.5553	0.8362	0.1678	0.8107	0.5570	0.8590
MoDiff (*ours*)	**0.1159**	**0.8328**	**0.5929**	**0.8892**	**0.1163**	**0.8588**	**0.5953**	**0.8912**

Table 2. Quantitative comparison of results with state-of-the-art ambiguous segmentation networks in terms of GED, HM-IoU, NCC and CI on the MS-MRI dataset. The best results are shown in bold, achieving state-of-the-art performance across all evaluation scores. Additionally, cases with a higher number of samples show relatively better performance.

Method	GED_{16}	$HM\text{-}IoU_{16}$	NCC_{16}	CI_{16}	GED_{32}	$HM\text{-}IoU_{32}$	NCC_{32}	CI_{32}
Prob.Unet [14]	0.2956	0.7430	0.3734	0.6635	0.2573	0.7464	0.3778	0.6597
cFlow [5]	0.2674	0.7513	0.3341	0.7148	0.2458	0.7611	0.3411	0.7322
MoSE [6]	0.3278	0.7054	0.3858	0.6223	0.3594	0.7154	0.3870	0.6332
CIMD [2]	0.3984	0.6864	0.3247	0.6623	0.3566	0.6401	0.3276	0.6720
CCDM [7]	0.2306	0.8405	0.4005	0.7443	0.2289	0.8488	0.4013	0.7527
MoDiff (*ours*)	**0.1671**	**0.8485**	**0.4109**	**0.7932**	**0.1659**	**0.8562**	**0.4128**	**0.8023**

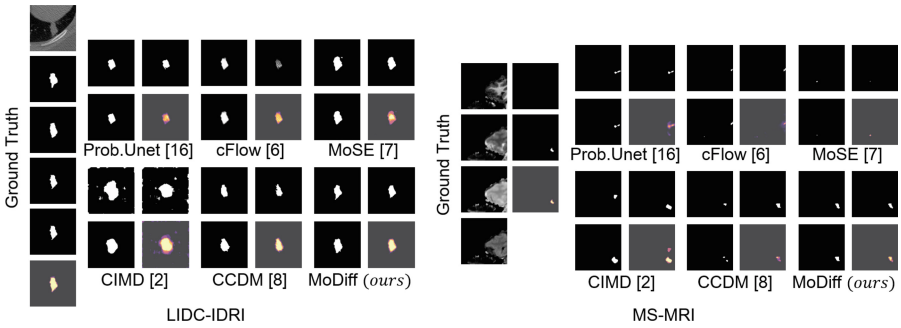


Fig. 2. The lesion segmentation results of the selected MoDiff and previous models for experiments are visualized. The left side shows the ground truth and the average label.

Table 3. Ablation Study results of hyperparameters on all the datasets.

Module		LIDC-IDRI				MS-MRI			
MCA	LDF	GED	HM-IoU	NCC	CI	GED	HM-IoU	NCC	CI
✓		0.1432	0.7954	0.4933	0.8281	0.2335	0.7649	0.3394	0.7428
	✓	0.1339	0.8191	0.5231	0.8377	0.2116	0.7748	0.3561	0.7502
✓	✓	**0.1163**	**0.8588**	**0.5953**	**0.8912**	**0.1659**	**0.8562**	**0.4128**	**0.8023**

Table 4. Case Study results of Edge Detection Methods on all the datasets.

Method	LIDC-IDRI				MS-MRI			
	GED	HM-IoU	NCC	CI	GED	HM-IoU	NCC	CI
Canny	0.2544	0.7588	0.4855	0.7590	0.2929	0.5883	0.3552	0.6313
Gaussian	0.2465	0.7642	0.5444	0.7469	0.2814	0.6102	0.3893	0.6954
Roberts	0.2208	0.8034	0.4902	0.6820	0.2738	0.6659	0.3617	0.7512
Prewitt	0.2265	0.8061	0.5047	0.7469	0.2549	0.7091	0.3859	0.7427
Sobel	0.2044	0.8222	0.5331	0.7590	0.2126	0.7218	0.3996	0.7843
LDF (*ours*)	**0.1163**	**0.8588**	**0.5953**	**0.8912**	**0.1659**	**0.8562**	**0.4128**	**0.8023**

4.3 Ablation and Case Study

Tables 3 present the ablation study results on the LIDC-IDRI and MS-MRI datasets, assessing the individual and combined contributions of the MCA and LDF modules. When using MCA alone (first row), the performance is the lowest, indicating its limited ability to capture detailed structural information. Employing only LDF (second row) improves several metrics, such as HM-IoU and NCC, yet does not reach optimal performance. Notably, the joint use of MCA and LDF (third row) yields the best results across all evaluation metrics, demonstrating that these modules complement each other to enhance both segmentation accuracy and consistency.

Tables 4 provide a comparative case study of various edge detection methods on the same datasets, including traditional approaches and the proposed LDF method. The results clearly show that while conventional edge detectors achieve moderate performance, they tend to concentrate weights on boundary pixels around organ contours or other edge regions, thereby failing to fully utilize the information from adjacent pixels. These findings underscore the superiority of the LDF method in accurately capturing edge details and maintaining segmentation consistency on both LIDC-IDRI and MS-MRI datasets.

5 Conclusion

MoDiff is a morphology-emphasized diffusion model that tackles the challenges of ambiguous medical image segmentation by integrating probability-based label

maps with a learnable morphological frequency space. This design enables the model to capture inherent uncertainties and fine structural details across multiple annotations, resulting in consistent and precise segmentation. Extensive evaluations on the LIDC-IDRI and MS-MRI datasets demonstrate that MoDiff outperforms conventional probabilistic models in segmentation accuracy, boundary delineation, and sample consistency. While increased sampling steps have led to improved performance metrics, this comes at the expense of computational efficiency. To address this limitation, future work will focus on refining the model architecture–exploring approaches such as noise scheduling and multimodal structures to reduce the number of required sampling steps.

Acknowledgments. This work was supported by an IITP grant funded by the Korean government MSIT (No. RS-2025-02304331, Digital Columbus Project).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc. (2015)
2. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: CVPR, pp. 11536–11546 (2023)
3. Monteiro, M., et al.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. Advances in Neural Information Processing Systems **33**, 12756–12767 (2020)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
5. Selvan, R., Faye, F., Middleton, J., Pai, A.: Uncertainty quantification in medical image segmentation with normalizing flows. In: MLMI 2020 (held with MICCAI 2020), pp. 80–90. Springer (2020)
6. Gao, Z., Chen, Y., Zhang, C., He, X.: Modeling multimodal aleatoric uncertainty in segmentation with mixture of stochastic experts. In: ICLR (2023)
7. Zbinden, L., Doorenbos, L., Pissas, T., Huber, A.T., Sznitman, R., Márquez-Neila, P.: Stochastic segmentation with conditional categorical diffusion models. In: ICCV, pp. 1119–1129 (2023)
8. Chen, T., Wang, C., Shan, H.: Berdiff: Conditional bernoulli diffusion model for medical image segmentation. In: MICCAI, pp. 491–501. Springer (2023)
9. Wu, J., et al.: Medsegdiff: medical image segmentation with diffusion probabilistic model. In: Medical Imaging with Deep Learning, pp. 1623–1639. PMLR (2024)
10. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
11. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10684–10695 (2022)
12. Armato, S.G., et al.: Lung image database consortium: developing a resource for the medical imaging research community. Radiology **232**(3), 739–748 (2004)

13. Carass, A., et al.: Longitudinal multiple sclerosis lesion segmentation: resource and challenge. Neuroimage **148**, 77–102 (2017)
14. Kohl, S., et al.: A probabilistic U-net for segmentation of ambiguous images. Advances in Neural Information Processing Systems **31** (2018)
15. Bellemare, M.G., et al.: The Cramer distance as a solution to biased Wasserstein gradients. arXiv preprint arXiv:1705.10743 (2017)
16. Salimans, T., Zhang, H., Radford, A., Metaxas, D.: Improving GANs using optimal transport. arXiv preprint arXiv:1803.05573 (2018)
17. Székely, G.J., Rizzo, M.L.: Energy statistics: a class of statistics based on distances. J. Stat. Plan. Inference **143**(8), 1249–1272 (2013)
18. Kassapis, E., Dikov, G., Gupta, D.K., Nugteren, C.: Calibrated adversarial refinement for stochastic semantic segmentation. In: ICCV, pp. 7057–7067 (2021)
19. Kohl, S.A.A., et al.: A hierarchical probabilistic U-net for modeling multi-scale ambiguities. arXiv preprint arXiv:1905.13077 (2019)
20. Baumgartner, C.F., et al.: PHiSeg: capturing uncertainty in medical image segmentation. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 119–127. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_14



MOST: MR Reconstruction Optimization for Multiple DownStream Tasks via Continual Learning

Hwihun Jeong[1,2], Se Young Chun[1,3], and Jongho Lee[1(✉)]

[1] Department of ECE, Seoul National University, Seoul, Korea
{hwihuni,sychun,jonghoyi}@snu.ac.kr
[2] Brigham and Women's Hospital, Harvard Medical School, Boston, USA
[3] INMC and IPAI, Seoul National University, Seoul, Korea

Abstract. Deep learning-based Magnetic Resonance (MR) reconstruction methods have focused on generating high-quality images but often overlook the impact on downstream tasks (e.g., segmentation) that utilize the reconstructed images. Cascading separately trained reconstruction network and downstream task network has been shown to introduce performance degradation due to error propagation and the domain gaps between training datasets. To mitigate this issue, downstream task-oriented reconstruction optimization has been proposed for a single downstream task. In this work, we extend the optimization to handle multiple downstream tasks that are introduced sequentially via continual learning. The proposed method integrates techniques from replay-based continual learning and image-guided loss to overcome catastrophic forgetting. Comparative experiments demonstrated that our method outperformed a reconstruction network without finetuning, a reconstruction network with naïve finetuning, and conventional continual learning methods. The source code is available at: https://github.com/SNU-LIST/MOST.

Keywords: Continual learning · MRI reconstruction · Downstream task

1 Introduction

Magnetic Resonance Imaging (MRI) has gained considerable prominence in clinical practice. The advent of deep learning has led to innovations in MRI such as disease classification [8,17,24,29], tumor or lesion segmentation [7,14], acquisition [28], and image processing [33]. In particular, the scan time reduction has been a focus of innovation, and various reconstruction networks have been proposed to produce high-quality images from highly undersampled k-space data [2,13,34]. When evaluating reconstructed images, the gold standard is the assessment by radiologists, which is often impractical due to high costs. Consequently, many studies rely on quantitative metrics such as the Structural Similarity Index

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 396–406, 2026.
https://doi.org/10.1007/978-3-032-04947-6_38

(SSIM) and Peak Signal-to-Noise Ratio (PSNR) [23,34]. Despite the utility, such metrics have limitations in reflecting the evaluation of radiologists. Moreover, when the reconstruction network is connected to a downstream task network (e.g., segmentation network), optimizing these metrics has been shown to deliver suboptimal outcomes in the downstream task [9,10]. As a result, efforts have been made to design a "downstream task-oriented" reconstruction network, optimizing it based on the performance of downstream tasks [11,30–32]. So far, however, these methods are designed for a single downstream task (Fig. 1a). Hence, when multiple downstream tasks exist and are progressively developed, the current solution is to develop multiple reconstruction networks each optimized for individual tasks, which may not be practical.

In this study, we propose an approach for the MR reconstruction Optimization for multiple downStream Tasks (MOST). Simple expansion from single-task to multi-task optimization is not straightforward [27]. While joint multi-task learning would be ideal when a new task arrives, it is often impractical due to the need to store and retrain all previous data, which is costly in terms of computation and storage. This issue is more critical in the medical domain, where long-term data storage is often restricted by privacy and regulatory constraints. Moreover, MRI data is large, often reaching hundreds of megabytes per scan, and general hospitals may perform over 100,000 scans annually. Given the need for multi-scanner and multi-task datasets, storage demands can quickly reach hundreds of terabytes, making data management challenging. Therefore, we consider a realistic scenario where the reconstruction network is sequentially finetuned with pretrained downstream task networks, assuming limited access to the training dataset of previously trained tasks (Fig. 1b).

When sequentially training a reconstruction network for multiple downstream tasks, one may consider naïve finetuning for the tasks. However, this approach can lead to a well-known issue of catastrophic forgetting, where performance on previously optimized tasks degrades substantially [3,16,18–20]. To overcome this issue and sustain the performances for the multiple downstream tasks, we propose to adopt the continual learning [4,6] and tailor it to our approach.

2 Method

2.1 Finetuning of Downstream Task-Oriented MR Reconstruction Network

Accelerated MRI reconstruction aims to produce high-quality images from undersampled k-space data, making them comparable to the images from fully sampled data. Traditionally, an MR reconstruction model denoted as f_R with parameters θ is trained using the fidelity loss function $\mathcal{L}_R$ such as SSIM loss:

$$\theta^* = \arg\min_{\theta} \mathcal{L}_R \left(f_R(x;\theta), y \right), \tag{1}$$

where x represents an undersampled k-space data, and y represents an aliasing-free label image.

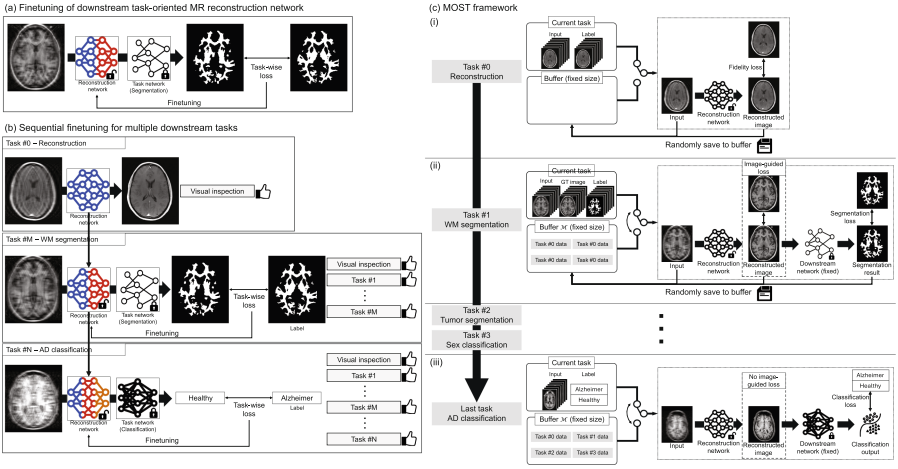


Fig. 1. (a) Combining individually trained reconstruction and downstream networks may result in performance degradation. Downstream task-oriented finetuning can mitigate these challenges. (b) In scenarios with progressively developed downstream tasks, sequential finetuning of the reconstruction network can be an effective solution. However, naïve finetuning can lead to catastrophic forgetting. (c) Overview framework of the MOST

However, when this reconstruction network is used for a downstream task such as segmentation or classification, a simple cascade of independently trained reconstruction network and downstream network may yield suboptimal results. For example, the reconstructed images may have different characteristics from the images trained for the downstream network due to the imperfection of the reconstruction network or domain gaps between the two training datasets. To address this issue, finetuning of the reconstruction network with end-to-end data (i.e., data pairs of undersampled k-space data and downstream task label) using a corresponding loss function is a feasible approach (Fig. 1a) [11,30–32]:

$$\theta^*_{task} = \arg\min_{\theta} \mathcal{L}_d \left(f_d(f_R(x;\theta)), z \right), \tag{2}$$

where f_d is the downstream task network, $\mathcal{L}_d$ is the corresponding loss function (e.g., cross-entropy for segmentation or classification), and z is the label (e.g., segmentation mask or classification label).

2.2 Sequential Finetuning for Multiple Downstream Tasks

When multiple downstream tasks exist, the task-oriented optimization for multiple downstream tasks ($t = 1, 2, 3, \cdots, T$) can be rewritten as follows:

$$\theta^*_{multi-task} = \arg\min_{\theta} \sum_{t} \mathcal{L}_{dt} \left(f_{dt}(f_R(x_t;\theta)), z_t \right), \tag{3}$$

where $\mathcal{L}_{dt}$ is the loss function of the t-th downstream task, and f_{dt} is the t-th downstream task network. The end-to-end dataset of finetuning for the t-th downstream task $D_t = \{x_t^i, z_t^i\}_{i=1}^{n_t}$ includes an undersampled k-space data x and a downstream task label z with n_t samples. However, such multi-task learning is difficult to be optimized [27] and does not adapt well to real-world applications. In this work, we propose that a single reconstruction network can be sequentially finetuned for multiple downstream tasks (Fig. 1b). Our scenario assumes limited access to previously trained datasets due to privacy regulations and computational costs. Initially, the reconstruction network f_R is trained exclusively for image reconstruction using a loss function ($\mathcal{L}_R$) and a reconstruction dataset (D_R) as written in Eq. 1. Then, we sequentially finetune the reconstruction network for multiple downstream tasks:

$$\theta_t^* = \arg\min_{\theta} \mathcal{L}_{dt}\left(f_{dt}(f_R(x_t;\theta)), z_t\right). \tag{4}$$

Finetuning process exclusively targets the reconstruction network while freezing downstream task networks. When finetuning for the t-th downstream task, we assume that only a small subset of the previous downstream task dataset ($D_p, p < t$) is available.

2.3 MOST

In our proposed approach, MOST, replay-based continual learning is combined with an image-guided (IG) loss to prevent catastrophic forgetting during sequential finetuning for multiple downstream tasks (Fig. 1c). For the replay, a fixed-size buffer ($\mathcal{M}$) is maintained to store a subset of input-output pairs from previous tasks. After finishing each task, we randomly select input-output pairs from that task and add them to the buffer. The buffer size remains constant throughout the finetuning process, with newly added data pairs replacing older ones. Hence, the number of data in the buffer is kept the same across the tasks.

Unlike conventional replay-based continual learning methods [4–6, 25], which generally combine the data of the current and previous tasks in the same mini-batch, our method presents a challenge in creating a combined mini-batch because the input data structure and label type can be different among tasks (e.g., segmentation: 2D input and 2D label vs. classification: 3D input and binary label). To address this challenge, data from the past task in the buffer are used in every K iterations during the finetuning process (Fig. 1c-ii and 1c-iii). The stored data are used in a round-robin fashion across the tasks.

Segmentation often uses high-quality aliasing-free images for label generation, so leveraging that information is reasonable. Therefore, for the downstream task of segmentation, we introduce an additional image-guided loss $\mathcal{L}_{IG}$, to further enhance the finetuning process (Fig. 1c-ii). This loss, which is calculated using a reconstruction loss $\mathcal{L}_R$, computes the difference between the intermediate reconstructed image $f_R(x_t;\theta)$ and the aliasing-free image y_t to enhance the performance of the reconstruction network:

$$\mathcal{L}_{IG} = \mathcal{L}_R(f_R(x_t; \theta), y_t). \tag{5}$$

The reconstructed images are also saved to the buffer to compute the image-guided loss for previous tasks.

3 Experiments and Results

3.1 Dataset and Implementation Details

We used T_1-weighted images from multiple datasets, including FastMRI [34], OASIS 1 [21], BraTS [22], IXI [1], and ADNI [15] for image reconstruction, white matter (WM) segmentation, tumor segmentation, sex classification, and Alzheimer's disease (AD) classification, respectively. To generate undersampled k-space data, we applied a forward model to the fully sampled image, assuming single-coil acquisition at an acceleration factor of 4. For the segmentation and classification datasets, we divided the dataset into two parts: one for the initial downstream task network training and the other for downstream task-oriented finetuning.

For the reconstruction network, a single coil version end-to-end variational network [13,34] is deployed. The U-net architecture [26] was used for the network part of the model. We also employed the same U-net architecture for the downstream task of segmentation tasks, and we utilized a CNN for classification tasks. The cross-entropy loss served as a loss function for segmentation and classification tasks. The finetuning process for each downstream task was conducted for a maximum of 5 epochs, and the network with the minimum validation loss was chosen. We utilized buffer data every three iterations ($K = 3$).

We assessed the quality of the reconstructed images using the structural similarity index (SSIM), segmentation results with the Dice similarity coefficient (DICE), and classification performance with the area under the curve (AUC) after completing finetuning for the last downstream task (Last Metric; LM). Additionally, forgetting measures (FM) were computed to quantify how much a model forgets previously learned tasks as it learns new ones. The forgetting measures were calculated by the difference between the best metric and the worst metric of a task obtained during sequential finetuning.

3.2 Performance Evaluation

We compared the performance of our MOST approach with a reconstruction network without downstream task-oriented finetuning (Without Finetuning) and a reconstruction network with naïve sequential finetuning (Naïve Finetuning). The buffer size was fixed to 10 subjects. The order of the tasks was WM segmentation, tumor segmentation, sex classification, and finally AD classification.

The qualitative results of the reconstruction and WM segmentation tasks are illustrated in Fig. 2. Without downstream task-oriented finetuning (Fig. 2; third column), the reconstructed images demonstrate good quality for the reconstruction dataset, but exhibit inferior results for the WM segmentation dataset due

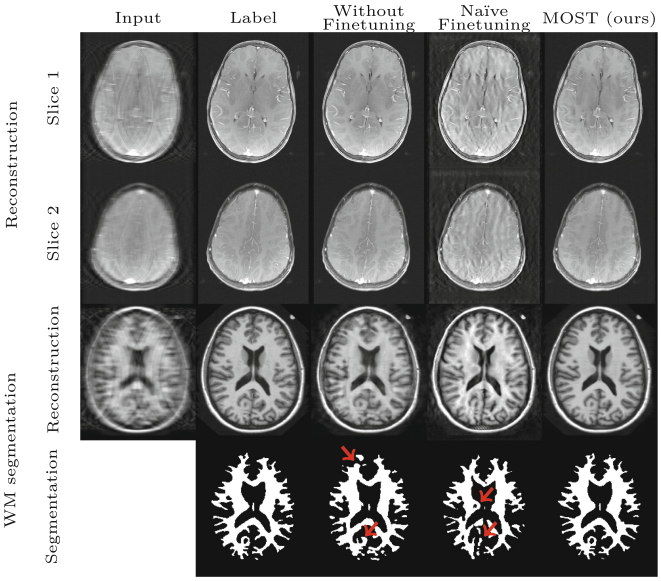


Fig. 2. The results of the reconstruction and WM segmentation before and after the downstream task-oriented finetuning are illustrated.

Table 1. The last metric (LM) and forgetting measure (FM) of SSIM for the reconstruction, DICE for the segmentation, and AUC for the classification are reported.

	Reconstruction SSIM		WM seg DICE		Tumor seg DICE		Sex class AUC		AD class AUC	
	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)
Without Finetuning	0.977	–	0.861	–	0.564	–	0.983	–	0.800	–
Naïve Finetuning	0.866	0.110	0.835	0.098	0.445	0.179	0.976	0.007	0.812	–
MOST (ours)	0.971	**0.006**	**0.938**	**0.002**	**0.642**	**0.003**	0.983	**0.000**	**0.813**	–

to domain gaps or error propagation. Naïve Finetuning results in poor quality in both reconstruction and segmentation, indicating catastrophic forgetting (Fig. 2; fourth column). The MOST approach effectively mitigates this issue, producing high-quality results in both tasks (Fig. 2; last column).

Table 1 provides the quantitative metrics. Compared to Without Finetuning, MOST successfully improved performance in all four downstream tasks with a minor decrease in reconstruction quality, demonstrating the effectiveness of finetuning. Compared to Naïve Finetuning, MOST exhibited superior LM and FM, demonstrating its effectiveness in preventing catastrophic forgetting.

3.3 Additional Evaluation

Task order experiments were conducted with three different task orders (Table 2a). Compared to Without Finetuning and Naïve Finetuning, MOST

Table 2. Result of last metric (LM) and forgetting measures (FM) with three different downstream task orders (a) and comparison result to the other continual learning methods (b) are illustrated.

(a)	Reconstruction SSIM		WM seg DICE		Tumor seg DICE		Sex class AUC		AD class AUC	
	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)
Without Finetuning	0.977	–	0.861	–	0.564	–	0.983	–	0.800	–
	Order 1: Recon → WM seg → Tumor seg → Sex class→ AD class									
Naïve Finetuning	0.866	0.110	0.835	0.098	0.445	0.179	0.976	0.007	0.812	–
MOST (ours)	0.971	**0.006**	**0.938**	**0.002**	**0.642**	**0.003**	**0.983**	**0.000**	**0.813**	–
	Order 2: Recon → Tumor seg → AD class → WM seg → Sex class									
Naïve Finetuning	0.843	0.133	0.930	0.002	0.453	0.181	0.977	-	0.761	0.051
MOST (ours)	0.970	**0.007**	**0.936**	**0.001**	**0.641**	**0.010**	**0.983**	–	**0.804**	**0.012**
	Order 3: Recon → Sex class → AD class → Tumor seg → WM seg									
Naïve Finetuning	0.791	0.186	**0.934**	–	0.421	0.207	0.956	0.031	0.761	0.053
MOST (ours)	0.959	**0.018**	0.920	–	**0.598**	**0.037**	**0.983**	**0.000**	**0.802**	**0.007**
(b)	Reconstruction SSIM		WM seg DICE		Tumor seg DICE		Sex class AUC		AD class AUC	
	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)
Naïve Finetuning	0.866	0.110	0.835	0.098	0.445	0.179	0.976	0.007	0.812	–
EWC [16]	0.910	0.066	0.884	0.048	0.536	0.085	0.976	0.007	0.808	–
LWF [19]	0.937	0.040	0.892	0.026	0.488	0.057	0.893	0.041	0.785	–
ER [6]	0.953	0.024	0.859	0.073	0.479	0.154	0.965	0.018	0.810	–
DER [4]	0.971	0.006	0.929	0.004	0.622	**0.000**	0.981	0.002	0.809	–
MOST (ours)	0.971	0.006	**0.938**	**0.002**	**0.642**	0.003	**0.983**	**0.000**	**0.813**	–

exhibited superior LM and FM in most of the tasks and task orders, showing minimal dependency on the task order. When comparing across the orders, starting with segmentation yielded marginal performance benefits, as indicated by a inferior LM and FM in Order 3 when compared to Orders 1 and 2.

We also **compared the results to conventional continual learning methods** including Elastic Weight Consolidation (EWC) [16], Learning Without Forgetting (LWF) [19], a memory-based method (ER) [6], and another replay-based method (DER) [4]. The results (Table 2b) clearly demonstrated that the MOST approach achieved overall superior performance compared to the other continual learning methods. In contrast, EWC, LWF, ER, and DER showed varying degrees of forgetting and performance degradation. These findings highlight the advantages of MOST over the conventional continual learning methods, mostly due to the fact that the image-guided loss can provide more comprehensive information.

When three **different buffer sizes** (4, 10, and 50 subjects) were evaluated, the results showed consistent performances as shown in Table 3a, indicating that MOST performed well even with the buffer size of 4.

To assess the independent contributions of replay-based continual learning and image-guided loss to the performance of MOST, an **ablation study** was conducted. Three networks were compared: original MOST, MOST without replay-based continual learning, and MOST without image-guided loss. The

Table 3. (a) MOST is tested for three different buffer sizes (4, 10, and 50 subjects). (b) Ablation study without replay-based continual learning or image-guided (IG) loss is illustrated.

(a)		Reconstruction SSIM		WM seg DICE		Tumor seg DICE		Sex class AUC		AD class AUC	
buffer size		LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)
4		0.970	0.007	0.933	0.007	0.640	0.006	0.982	0.001	0.810	–
10		0.971	0.006	0.938	0.002	0.642	0.003	0.983	0.000	0.813	–
50		0.970	0.007	0.937	0.003	0.640	0.004	0.981	0.003	0.810	–
(b)		Reconstruction SSIM		WM seg DICE		Tumor seg DICE		Sex class AUC		AD class AUC	
Replay	IG loss	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)	LM(↑)	FM(↓)
X	O	0.950	0.026	0.883	0.056	0.604	0.046	0.976	0.007	**0.817**	–
O	X	0.971	0.006	0.929	0.004	0.622	**0.000**	0.981	0.002	0.809	–
O	O	0.971	0.006	**0.938**	**0.002**	**0.642**	0.003	**0.983**	**0.000**	0.813	–

ablation study results (Table 3b) demonstrated that original MOST outperformed MOST without either replay-based continual learning or image-guided loss. Furthermore, the performance of MOST without image-guided loss confirms that replay-based continual learning alone substantially contributes to performance improvement and catastrophic forgetting reduction.

4 Discussion

In this work, we demonstrated that a single reconstruction network can be sequentially finetuned for multiple downstream tasks. This method successfully generalized recently proposed single downstream task-oriented reconstruction optimization to multiple downstream tasks. In particular, the application of replay-based continual learning along with the image-guided loss successfully prevented catastrophic forgetting, which was observed in naïve finetuning.

MOST demonstrated improved performance in all task orders when compared to that of Naïve Finetuning, but it showed little dependency on task order (Table 2a). In particular, MOST underperformed in Order 3, which may be attributed to the introduction of classification tasks earlier than segmentation tasks. Classification labels contained less detailed information compared to segmentation labels, leading to less effective finetuning.

Our proposed framework is based on the single-coil setting, where the reconstruction model takes single-coil k-space data as the input and produces an alias-free image. The reason for this setting is the limited available end-to-end datasets that include multi-coil k-space data and downstream task labels.

Current continual learning methods mostly target high-level computer vision tasks like classification or segmentation [12], because catastrophic forgetting is rare in regression tasks. In our scenario, although we employ sequential training for the regression network, the cascaded network ultimately performs classifications or segmentations. Hence, catastrophic forgetting occurs in the downstream

tasks. If the reconstruction network is merely finetuned with the original reconstruction task for different domains (e.g., images with different hardware or parameters), the effect of catastrophic forgetting may decrease, and continual learning may not be required.

In our approach, we manually trained the downstream task networks instead of adopting off-the-shelf state-of-the-art (SOTA) models, resulting in inferior metrics. This approach was necessary because the SOTA model is trained on the entire training dataset, whereas we required the dataset to be split for downstream network training and finetuning. Moreover, the BraTS dataset contained multimodal MRI data, but we only used the T_1-weighted images for consistency across downstream tasks. Additionally, most SOTA models included preprocessing steps on the input images (e.g., skull stripping), which are challenging to be integrated into MOST.

5 Conclusion

We introduced that a single reconstruction network can be sequentially finetuned for multiple downstream tasks and demonstrated that replay-based continual learning and image-guided loss can prevent catastrophic forgetting.

Acknowledgments. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2022R1A4A1 030579), Korea Health Industry Development Institute (RS-2024-00439677), Korea Basic Science Institute (RS-2024-00435727), Samsung Electronics Co., Ltd (IO201216-08215-01), INMC and IOER at Seoul National University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ixi - information extraction from images. https://brain-development.org/ixi-dataset
2. Aggarwal, H.K., Mani, M.P., Jacob, M.: Modl: model-based deep learning architecture for inverse problems. IEEE TMI **38**(2), 394–405 (2018)
3. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: learning what (not) to forget. In: ECCV (2018)
4. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: Dark experience for general continual learning: a strong, simple baseline. In: NeurIPS (2020)
5. Cha, H., Lee, J., Shin, J.: Co2l: contrastive continual learning. In: ICCV (2021)
6. Chaudhry, A., et al.: On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486 (2019)
7. Chen, L., Bentley, P., Rueckert, D.: Fully automatic acute ischemic lesion segmentation in dwi using convolutional neural networks. Neuroimage Clin. **15**, 633–643 (2017)

8. Cuingnet, R., et al.: Automatic classification of patients with alzheimer's disease from structural mri: a comparison of ten methods using the adni database. Neuroimage **56**(2), 766–781 (2011)
9. Desai, A.D., et al.: The international workshop on osteoarthritis imaging knee mri segmentation challenge: a multi-institute evaluation and analysis framework on a standardized dataset. Radiol. Artif. Intell. **3**(3), e200078 (2021)
10. Desai, A.D., et al.: Skm-tea: a dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation. arXiv preprint arXiv:2203.06823 (2022)
11. Fan, Z., Sun, L., Ding, X., Huang, Y., Cai, C., Paisley, J.: A segmentation-aware deep fusion network for compressed sensing mri. In: ECCV (2018)
12. Hadsell, R., Rao, D., Rusu, A.A., Pascanu, R.: Embracing change: continual learning in deep neural networks. Trends Cogn. Sci. **24**(12), 1028–1040 (2020)
13. Hammernik, K., et al.: Learning a variational network for reconstruction of accelerated mri data. Magn. Reson. Med. **79**(6), 3055–3071 (2018)
14. Havaei, M., et al.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017)
15. Jack Jr, C.R., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. J. Magn. Reson. Imaging **27**(4), 685–691 (2008)
16. Kirkpatrick, J., et al.: Overcoming catastrophic forgetting in neural networks. PNAS **114**(13), 3521–3526 (2017)
17. Li, F., Tran, L., Thung, K.H., Ji, S., Shen, D., Li, J.: A robust deep model for improved classification of ad/mci patients. IEEE J. Biomed. Health Inform. **19**(5), 1610–1616 (2015)
18. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: a continual structure learning framework for overcoming catastrophic forgetting. In: ICML (2019)
19. Li, Z., Hoiem, D.: Learning without forgetting. IEEE TPAMI **40**(12), 2935–2947 (2017)
20. Loo, N., Swaroop, S., Turner, R.E.: Generalized variational continual learning. In: ICLR (2021)
21. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. **19**(9), 1498–1507 (2007)
22. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE TMI **34**(10), 1993–2024 (2014)
23. Muckley, M.J., et al.: Results of the 2020 fastmri challenge for machine learning mr image reconstruction. IEEE TMI **40**(9), 2306–2317 (2021)
24. Péran, P., et al.: Mri supervised and unsupervised classification of parkinson's disease and multiple system atrophy. Mov. Disord. **33**(4), 600–608 (2018)
25. Prabhu, A., Torr, P.H., Dokania, P.K.: Gdumb: a simple approach that questions our progress in continual learning. In: ECCV (2020)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241 (2015)
27. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. In: NeurIPS (2018)
28. Shin, D., et al.: Deep reinforcement learning-designed radiofrequency waveform in mri. Nat. Mach. Intell. **3**(11), 985–994 (2021)
29. Suk, H.I., Lee, S.W., Shen, D., Initiative, A.D.N., et al.: Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. Neuroimage **101**, 569–582 (2014)

30. Sun, L., Fan, Z., Ding, X., Huang, Y., Paisley, J.: Joint cs-mri reconstruction and segmentation with a unified deep network. In: IPMI (2019)
31. Wang, Z., et al.: One network to solve them all: a sequential multi-task joint learning network framework for mr imaging pipeline. In: MICCAI Workshop (2021)
32. Wu, Z., et al.: Learning task-specific strategies for accelerated mri. IEEE Trans. Comput. Imaging (2024)
33. Yoon, J., et al.: Quantitative susceptibility mapping using deep neural network: Qsmnet. Neuroimage **179**, 199–206 (2018)
34. Zbontar, J., et al.: fastmri: an open dataset and benchmarks for accelerated mri. arXiv preprint arXiv:1811.08839 (2018)

MTCNet: Motion and Topology Consistency Guided Learning for Mitral Valve Segmentation in 4D Ultrasound

Rusi Chen[1], Yuanting Yang[2], Jiezhi Yao[1], Hongning Song[2], Ji Zhang[2], Yongsong Zhou[1], Yuhao Huang[1], Ronghao Yang[1], Dan Jia[2], Yuhan Zhang[1], Xing Tao[1], Haoran Dou[3], Qing Zhou[2(✉)], Xin Yang[1(✉)], and Dong Ni[1,4,5,6]

[1] Medical Ultrasound Image Computing (MUSIC) Lab, School of Biomedical Engineering, Medical School, Shenzhen University, Shenzhen, China
xinyang@szu.edu.cn
[2] Renmin Hospital of Wuhan University, Wuhan, China
qingzhou.wh.edu@hotmail.com
[3] Centre for CIMIM, Manchester University, Manchester, UK
[4] School of Artificial Intelligence, Shenzhen University, Shenzhen, China
[5] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China
[6] School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China

Abstract. Mitral regurgitation is one of the most prevalent cardiac disorders. Four-dimensional (4D) ultrasound has emerged as the primary imaging modality for assessing dynamic valvular morphology. However, 4D mitral valve (MV) analysis remains challenging due to limited phase annotations, severe motion artifacts, and poor imaging quality. Yet, the absence of inter-phase dependency in existing methods hinders 4D MV analysis. To bridge this gap, we propose a Motion-Topology guided consistency network (**MTCNet**) for accurate 4D MV ultrasound segmentation in semi-supervised learning (SSL). MTCNet requires only sparse end-diastolic and end-systolic annotations. First, we design a cross-phase motion-guided consistency learning strategy, utilizing a bi-directional attention memory bank to propagate spatio-temporal features. This enables MTCNet to achieve excellent performance both per- and inter-phase. Second, we devise a novel topology-guided correlation regularization that explores physical prior knowledge to maintain anatomically plausible. Therefore, MTCNet can effectively leverage structural correspondence between labeled and unlabeled phases. Extensive evaluations on the first largest 4D MV dataset, with 1408 phases from 160 patients, show that MTCNet performs superior cross-phase consistency compared to other advanced methods (Dice: 87.30%, HD: 1.75 mm). Both the code and the dataset are available at https://github.com/crs524/MTCNet.

Keywords: Mitral Valve Segmentation · 4D Ultrasound · Semi-supervised Learning · Consistency Guided Learning

R. Chen, Y. Yang, J. Yao and H. Song—Contributed equally.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 407–417, 2026.
https://doi.org/10.1007/978-3-032-04947-6_39

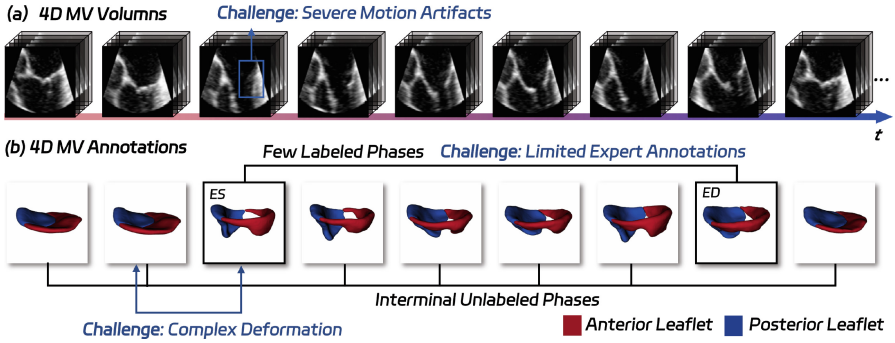


Fig. 1. Illustration of MV volumes and annotations in 4D ultrasound.

1 Introduction

Mitral regurgitation (MR) is a common cardiovascular disease with high morbidity and mortality [6,16]. Transesophageal Echocardiography (TEE) is the gold standard for diagnosing and quantifying MR. It offers a real-time view of the mitral valve (MV), providing both temporal and spatial perspectives [21]. Accurate 4D MV segmentation enables precise measurement of MV structure and functional analysis, as well as patient-specific 3D printing for surgical planning.

However, automatic MV segmentation in 4D ultrasound faces challenges caused by a few phase annotations, severe motion artifacts, and complex deformations (See Fig. 1). To tackle these challenges, we aim to develop a 4D MV segmentation method with only end-diastolic (ED) and end-systolic (ES) phase annotations, while ensuring high accuracy and temporal coherence.

Despite the promising progress in deep learning [8,9,13,25], research on 4D MV segmentation remains limited, with most studies focusing on single-volume segmentation. Some studies [1,3] employed U-Net [19] for 3D MV segmentation, while others [5,17] utilized the nnU-Net [10] framework. Although these methods achieve promising accuracy in static 3D volumes, they struggle with MV motion due to a lack of information extraction from unannotated phases. Semi-supervised learning (SSL) [20,26,27] has emerged as a promising paradigm to enhance model performance with limited labeled data by utilizing the abundant information in unlabeled data. Ivantsits et al. [11] simplify MV as an idea tubular sheet to reconstruct 4D MV surface, which limits its use in simulating the common organic mitral regurgitation disease with valve abnormalities. Munaf'o et al. [18] proposed an SSL method for 4D MV segmentation, where ED and ES phases of all patients are treated as labeled, while intermediate phases are considered unlabeled ones. However, they treat each phase independently, neglecting patient-level phase consistency, leading to suboptimal segmentation.

To address these challenges, we leverage motion and topology-guided consistency learning to account for the relationships between patient-level labeled and unlabeled phases while preserving anatomical constraints. In this paper, we pro-

pose a motion-topology guided consistency method termed MTCNet to regard all phases of a patient as input samples to mine cross-phase relevance. Firstly, to structurally enhance the motion coherence, we introduce motion-guided consistency learning (MCL). Through a well-designed bi-directional memory bank (BMB), MCL can encourage MTCNet to effectively learn from key labeled phases and then propagate the semantic features to the unlabeled ones. Secondly, considering the anatomical prior of the MV's surface-volume invariance, we embed this property into a topology consistency regularization (TCR), optimizing inter-phase dissimilarity through an extra surface and volume continuity constraint. Extensive evaluations on the largest 4D MV dataset demonstrate that MTCNet outperforms existing SSL approaches. Our contributions are three-fold:

- We introduce a novel motion and topology-guided learning framework for 4D MV segmentation, which achieves superior per-phase and inter-phase performance with sparse ES and ED annotations.
- We propose a MCL strategy to effectively propagate semantic information across labeled and unlabeled phases, thereby enhancing motion consistency.
- We design a simple yet effective TCR with surface and volume variance that boosts segmentation accuracy and preserves topological coherence.

2 Methods

Problem Setting. Given a patient-level sequence $\mathbf{D}$ comprising a labeled subset $\mathbf{D}_l^i = \{(X_l, Y, t)\}$ and an unlabeled subset $\mathbf{D}_u^i = \{(X_u, t')\}$ for patient i, where t and t' represent the indices of the ES and ED phases, and the intermediate phases, respectively. Our goal is to segment all phases within a cycle using a limited number of labeled phases. Specifically, $X_l \in \mathbb{R}^{D\times H\times W}$ and $X_u \in \mathbb{R}^{D\times H\times W}$ denote labeled phases and unlabeled phases, respectively. $Y \in \{0,1\}^{D\times H\times W}$ represents the MV labels of X_l.

2.1 Patient-Level Cross-Phase Learning Framework

The overall framework of our method is illustrated in Fig. 2. Inspired by the Mean Teacher architecture [20], a powerful SSL method, MTCNet is designed to enhance learning from both labeled and unlabeled data. For each training iteration, MTCNet takes triplet phases $\mathcal{T}$ as input from the sequence $\mathbf{D}_l^i$. Specifically, one of the triplet phases is a labeled phase, while the other two phases are unlabeled volumes from intermediate phases. Thus, the input can be defined as $\mathcal{T} = \{X_l, X_u^1, X_u^2\}$, where $X_l \in \{X_{ES}, X_{ED}\}$. Simultaneously, $\mathcal{T}$ is fed into the student and teacher models, which share the same architecture.

During training, the teacher network supervises the student network by generating high-confidence pseudo-labels. The teacher network's parameters θ_t are updated via exponential moving average (EMA) [20] from the student's parameters θ_s. It can be formulated as $\theta_t = \alpha\theta_t + (1-\alpha)\theta_s$, where $\alpha \in (0,1)$ is the momentum mitigating the overfitting of the teacher network on limited labeled

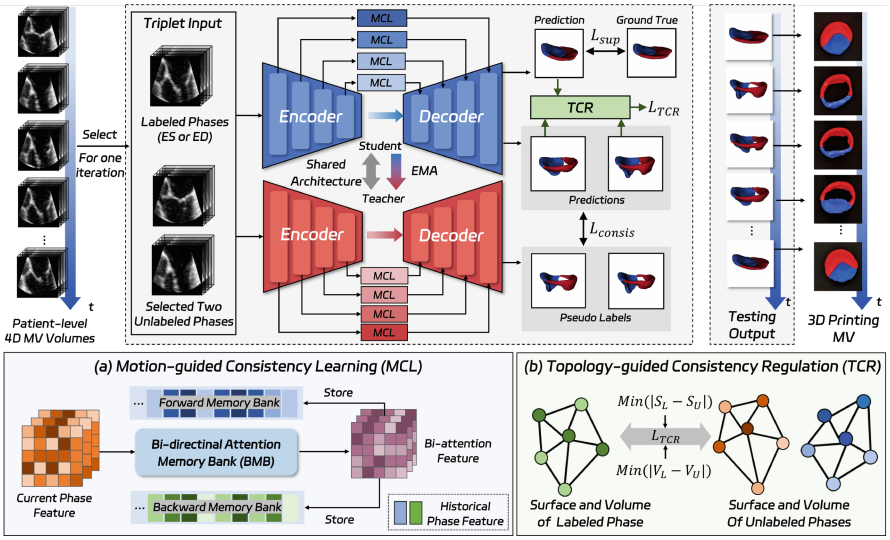


Fig. 2. Overall framework of our proposed MTCNet.

data. Ultimately, MTCNet generates the predicted segmentation masks for both the labeled and unlabeled volumes. The total training objective is:

$$\mathcal{L}_{seg} = \mathcal{L}_{sup}(f_s(\theta), Y) + \beta \cdot \mathcal{L}_{consis}(f_s(\theta), f_t(\theta)), \tag{1}$$

where β is the loss weight, L_{sup} and L_{consis} indicate the supervised loss and consistency loss, respectively. $f_s(\theta)$ represents the segmentation model's output of labeled volume, and $f_t(\theta)$ for unlabeled volumes. Both losses combine dice and binary cross-entropy loss with a 0.8 to 0.2 weight ratio. During the testing stage, MTCNet is able to predict masks for all phases in an end-to-end manner.

2.2 Motion-Guided Consistency Learning

To effectively enhance the inter-phase motion coherence, we propose an MCL strategy through the BMB block. As shown in Fig. 2 (a), we propose forward memory bank M_f and backward memory bank M_b to store the temporal information for all phases of a patient. M_f captures systolic-diastolic deformation, while M_b encodes reverse motion, jointly modeling bidirectional patterns to reduce phase misalignment in cardiac cycles.

Figure 3 illustrates the detailed design of the BMB block. Here, we take a memory bank to illustrate the mechanism, as both follow the same process. Given a memory bank M containing T memory phases, the memory encoder would generate memory key $k^M \in \mathbb{R}^{C^k \times TDHW}$ and memory value $v^M \in \mathbb{R}^{C^v \times TDHW}$. Similarly, the query encoder produces a query key $k^Q \in \mathbb{R}^{C^k \times DHW}$ and a query value $k^Q \in \mathbb{R}^{C^k \times DHW}$, where D, H, and W are the multi-scale feature dimensions. Current phase feature is the output feature of encoders both in the student

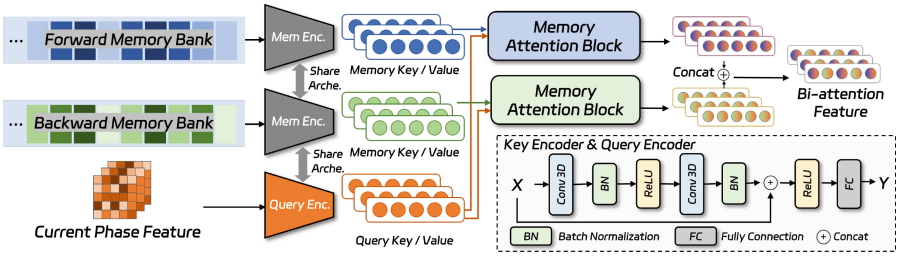


Fig. 3. Detailed design for bi-directional memory bank.

and teacher models. Therefore, multi-scale memory aggregation minimizes information loss, especially for subtle boundary changes in low signal-to-noise ratio ultrasound phases. Then, we compute the normalized affinity matrix $\mathbf{W}$ to to effectively weigh the dependencies across phases:

$$\mathbf{W}_{ij} = \frac{\exp(c(\mathbf{k}_i^M, \mathbf{k}_j^Q))}{\sum_n \exp(c(\mathbf{k}_n^M, \mathbf{k}_j^Q))}, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{TDHW \times DHW}$, $\mathbf{k}_i$ denotes the feature vector at the i-th position and c represents the dot product.

With the normalized affinity matrix $\mathbf{W}$, the aggregated readout feature $v^Q \in \mathbb{R}^{C^v \times DHW}$ for the query phase is computed as a weighted sum of the memory features using a top-k operation. To be specific, the weighted sum of the top-k memory features is calculated as: $v_r^Q = v_r^M \mathbf{W}_r$, where r represents two forward and backward memory banks. Finally, the v_1^Q and v_2^Q are concatenated to obtain the bi-directional attention feature. This feature is stored in the memory banks M_f and M_b and passed to the decoder to produce the segmentation mask.

2.3 Topology-Guided Consistency Regulation

To ensure structural integrity during complex deformations across phases, we propose the TCR mechanism. This approach is grounded in prior knowledge that surface area and volume should be stable throughout such deformations [15] (see Fig. 2 (b)). Specifically, for a predicted probability map P_t for t-th phase, we first obtain its binarized form $B_t = \mathbb{I}(P_t > 0.5)$, which represents the region of interest. Then, the normalized surface area $S(P_t)$ of the t-th phase can be defined as:

$$S(P_t) = \frac{\int_{\mathcal{S}} \nabla B_t(\mathbf{x}), d\mathcal{A}(\mathbf{x})}{\int_{\mathcal{S}} \mathbb{I}(\nabla B_t(\mathbf{x}) > 0), d\mathcal{A}(\mathbf{x}) + \epsilon} \approx \frac{\sum_{v \in V} \nabla B_t(v), \Delta A(v)}{\sum_{v \in V} \mathbb{I}(\nabla B_t(v) > 0), \Delta A(v) + \epsilon}, \tag{3}$$

where $\nabla = (\partial_x, \partial_y, \partial_z)$ denotes 3D Sobel operators so that $\nabla B_t(\mathbf{x})$ represents the spatial gradient of the volumetric mask B_t at location $\mathbf{x}$. $\Delta A(v)$ is the discrete

voxel-level form and ϵ prevents division by zero. Sobel operators ensure the computation is differentiable, enabling seamless integration into gradient-based optimization frameworks. Thus, the surface-area-based loss can be defined as:

$$\mathcal{L}_{\text{surf}} = \sum_{t=2}^{3} \left(\left| 1 - \frac{S(P_t)}{S(P_1)} \right| + \lambda \left| S(P_t) - S(P_1) \right| \right). \tag{4}$$

The dual-term design overcomes key limitations of static shape constraints. The relative term adapts to patient-specific anatomy via the annotated $S(P_1)$, while the absolute term prevents error propagation through physical consistency. Follow a similar strategy to surface area consistency, the volume consistency loss can be formulated based on the voxel-wise summation of B_t:

$$\mathcal{L}_{\text{vol}} = \sum_{t=2}^{3} \left(\left| 1 - \frac{V(P_t)}{V(P_1)} \right| + \lambda \left| V(P_t) - V(P_1) \right| \right), \tag{5}$$

where $\sum_i B_t(i)$ is the voxel-wise summation of B_t. Finally, the total topological consistency regularization loss $\mathcal{L}_{\text{tcp}}$ can be written as: $\mathcal{L}_{\text{tcp}} = \mathcal{L}_{\text{surf}} + \mathcal{L}_{\text{vol}}$. The $\mathcal{L}_{\text{tcp}}$ acts as a physics-informed regularizer that enforces temporal plausibility. Combined with Eq. 1, the total loss for triplet input phases integrates both segmentation and physiological regularization $\mathcal{L}_{total} = \mathcal{L}_{\text{seg}} + \sigma \cdot \mathcal{L}_{\text{tcp}}$, where σ is a constant set to 0.1 by default based on empirical observations.

3 Experiments and Results

Implementation Details. For a fair comparison, all SSL experiments were conducted in the same setting, with only ES and ED annotated labels for training. Our model was implemented with PyTorch 1.11.0 on two NVIDIA GeForce RTX A40 GPUs. All input volumes were all resampled to $128 \times 128 \times 128$ based on isotropic spacing resizing. During training, the Adam optimizer was used with an initial learning rate of 10^{-4}, then reduced by 0.1 every 20 epochs.

Datasets and Evaluation Metrics. We collected 4D TEE MV data from 160 patients with 1408 phases from cooperating hospitals. The in-house dataset comprised 147 cases with functional mitral regurgitation, and 13 normal cases. The dataset was randomly split into training (112 cases), validation (16 cases), and testing (32 cases) subsets. The number of phases varied per case due to heart rate differences, with an average of nine phases per cycle. All imaging data were acquired using the Philips X5-1 transthoracic volume probe. Two sonographers with more than 10 years of experience manually annotated the MV volumes, consisting anterior leaflets (AL) and posterior leaflets (PL). We also validated on the mid-diastolic (MD) phase and its adjacent transitional phase (MD-1) between MD and ED. For evaluation, Dice coefficient, 95th-percentile Hausdorff Distance (HD) and Conformity (Conf) [4], were adopted for quantitative comparison.

Comparison Study. We evaluated MTCNet against multiple SSL methods [7,12,14,20,22,23,26], the registration-based approach VoxelMorph (VM) [2], and the two-shot video object segmentation method (T-VOS) [24]. VM and T-VOS both require one first phase as reference. As depicted in Table 1, MTCNet achieves state-of-the-art (SOTA) performance, surpassing compared methods in Dice (87.30%) and Conf (66.71%) in all phases segmentation. Specifically, our method surpasses the best baseline ICT (86.23%) by 1.07% ($p<0.05$). Notably, although MD is the most challenging phase due to the leaflets are maximally separated and often near the left ventricular wall, MTCNet surpasses the sub-optimal T-VOS method by 2.4% (82.69% vs. 80.29%) in Dice, highlighting MTCNet 's ability in handling complex deformation. Additionally, compared to reference-phase-driven methods VM and T-VOS, MTCNet shows more potential results in capturing continuous motion more effectively. Moreover, MTCNet achieves robust performance for both labeled phases (ES and ED) and unlabeled phases (MD and MD-1). This highlights MTCNet 's superior ability in per-phase and inter-phase performance.

Table 1. Comparison results for different methods. All Phases stands for total phases in a cardiac cycle. * denotes statistically significant differences (t-test) between MTCNet and compared SOTA methods.

Method	ES			MD			MD-1			ED			All Phases		
	Dice	HD	Conf	Dice	HD	Conf	Dice	HD	Conf	Dice	HD	Conf	Dice	HD	Conf
MT [20]	86.13	1.74	66.96	79.25	2.94	32.62	85.63	1.89	64.42	87.50	1.41	70.58	85.61*	1.78	63.27*
UA-MT [26]	86.44	**1.68**	67.85	79.76	2.75	37.50	86.22	1.85	66.53	87.85	1.35	71.62	86.08*	**1.69**	65.11*
SASS [12]	85.90	1.85	66.45	79.67	2.74	35.21	86.64	**1.69**	68.03	87.76	1.45	71.29	86.07*	1.71	64.75*
DTC [14]	84.71	2.07	61.04	79.14	3.15	25.81	86.26	1.99	66.39	87.50	1.54	70.10	85.31*	1.92*	60.69*
ICT [22]	86.57	1.79	68.01	80.17	2.83	40.29	85.79	2.03	58.71	88.17	1.33	72.42	86.23*	1.77	64.81*
MCF [23]	85.64	2.09	64.77	78.53	3.29	16.00	86.33	1.97	65.92	88.20	1.52	72.35	85.54*	2.06*	60.01*
CC-Net [7]	85.23	1.88	63.38	78.63	2.89	33.67	86.14	1.75	66.53	87.17	1.39	69.61	85.16*	1.79	64.81*
VM [2]	–	–	–	73.99	4.03	6.65	83.89	2.37	57.24	87.79	1.69	70.35	83.01*	2.46*	42.33*
T-VOS [24]	–	–	–	80.29	2.81	**44.54**	83.97	2.66	60.63	85.15	2.05	63.98	84.56*	2.12*	61.36*
MTCNet	**87.42**	1.83	**70.55**	**82.69**	**2.45**	34.96	**87.14**	2.03	**69.14**	**89.20**	**1.33**	**75.10**	**87.30**	1.75	**66.71**

Table 2. Ablation study of MTCNet. M and T represent MCL and TCR, respectively. Four key phases denote the four key phases (ES, MD, MD-1, ED). * denotes statistically significant differences (t-test) between best model and other configurations.

Method	Four Key Phases									All Phases								
	AL			PL			Mean			AL			PL			Mean		
	Dice	HD	Conf	Dice	HD	Conf	Dice	HD	Conf	Dice	HD	Conf	Dice	HD	Conf	Dice	HD	Conf
Based	85.81	1.93	**62.16**	83.21	2.29	55.55	84.51	2.11	58.85	86.95	**1.74**	**66.98**	84.08	1.98	59.53	85.51*	1.86	63.25*
Based+M	86.90	1.92	61.63	85.15	2.23	60.42	86.02	2.08	61.03	87.75	1.78	66.71	85.57	2.00	62.12	86.66*	1.89	64.41*
Based+M+T	**87.64**	**1.90**	60.35	**85.58**	**1.92**	**64.52**	**86.61**	**1.91**	**62.43**	**88.41**	1.75	66.82	**86.19**	**1.76**	**66.61**	**87.30**	**1.75**	**66.71**

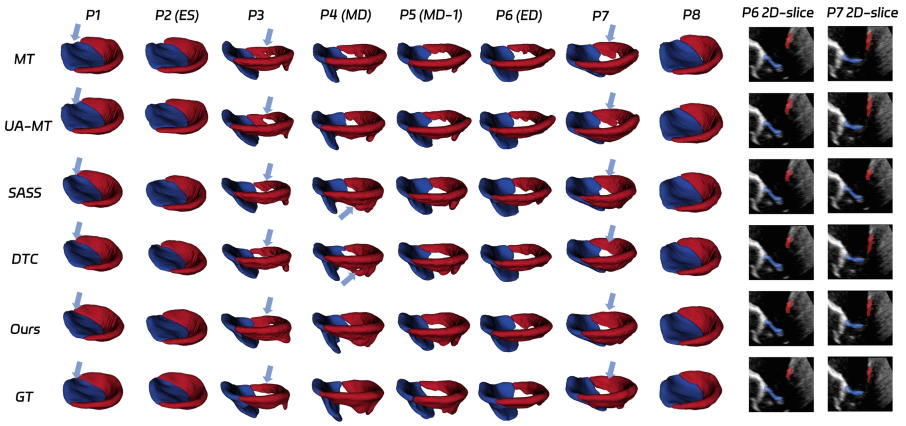


Fig. 4. Examples in consecutive phases with 3D volumes (Left seven columns) and 2D slices (Right two columns). Blue and red represent the AL and PL, respectively. (Color figure online)

Ablation Study. To demonstrate the impact of different components, we performed an ablation study in Table 2. Based-MTCNet means directly setting labeled and unlabeled data from the same patient based on Mean Teacher learning strategy. When integrating MCL, the mean Dice improves by 1.51% (84.51% → 86.02%) for all four key phases segmentation. Specifically, for AL and PL Dice, notably increasing +1.09% and +1.94%, respectively. This validates MCL's ability to propagate motion cues across phases. TCR further refines anatomical plausibility, reducing PL HD by 0.31 mm (2.23 → 1.92) and achieving a mean HD of 1.91 mm. Under full-phase evaluation, applying TCR improves PL Conf by 2.3% (64.41% → 66.71%), which significantly improved the consistency among different phases. The best model, Based+MCL+TCR, achieves a 1.79% improvement in Dice, a 0.11 mm reduction in HD and a 3.46% improvement in Conf compared to the baseline. The progressive improvements highlight their complementary roles: MCL captures temporal dependencies through BMB, while TCR enforces surface-volume continuity via physical regularization.

Qualitative Results. We further perform a qualitative comparison. Figure 4 clearly shows that MTCNet produces smoother and more complete segmentations. Particularly in the MD phase where severe motion and deformation often cause holes (bule arrows). Artifacts in this region are most prominent due to the valve's motion, indicating MTCNet 's superior ability to handle complex motion and structure changes. To further validate the clinical applicability of our approach, we compared the 3D printing results (See Fig. 5). While other methods produce incomplete shapes and missing details, our model achieves greater realism and completeness, showcasing its potential for patient-specific planning.

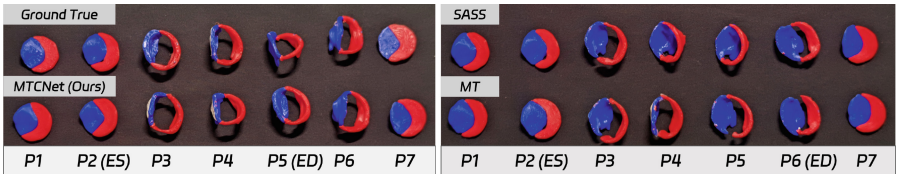


Fig. 5. Examples of 3D printing models of MV among different methods.

4 Conclusion

In this study, we present MTCNet, a SSL framework that leverages motion and topology-guided consistency learning for 4D MV segmentation. MTCNet performs better than the SOTA with sparse annotation, especially in motion and structure coherence. We contribute this gain in performance in two key contributions. First, the proposed MCL strategy can learn motion information across phases through BMB mechanism, improving segmentation performance in per-phase and inter-phase. Additionally, a robust TCR strategy with surface and volume variance as prior knowledge boosts segmentation accuracy while preserving topological coherence. MTCNet can also facilitate MV dynamics analysis and support personalized hemodynamic assessment and treatment planning.

Acknowledgments. This work was supported by the grant from National Natural Science Foundation of China (82201851, 82102046, 12326619, 62171290); Science and Technology Planning Project of Guangdong Province (2023A0505020002); Frontier Technology Development Program of Jiangsu Province (BF2024078); Guangxi Province Science Program (2024AB17023); Yunnan Major Science and Technology Special Project Program (202402AA310052); Yunnan Key Research and Development Program(202503AP140014).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aly, A.H., et al.: Fully automated 3d segmentation and diffeomorphic medial modeling of the left ventricle mitral valve complex in ischemic mitral regurgitation. Med. Image Anal. **80**, 102513 (2022)
2. Balakrishnan, G., Zhao, A., Sabuncu, M., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE TMI: Trans. Med. Imaging **38**, 1788–1800 (2019)
3. Carnahan, P., Moore, J., Bainbridge, D., Eskandari, M., Chen, E.C.S., Peters, T.M.: DeepMitral: fully automatic 3D echocardiography segmentation for patient specific mitral valve modelling. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 459–468. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_44

4. Chang, H.H., Zhuang, A.H., Valentino, D.J., Chu, W.C.: Performance measure characterization for evaluating neuroimage segmentation algorithms. Neuroimage **47**(1), 122–135 (2009)
5. Chen, J., et al.: Automatic 3d mitral valve leaflet segmentation and validation of quantitative measurement. Biomed. Signal Process. Control **79**, 104166 (2023)
6. El Sabbagh, A., Reddy, Y.N., Nishimura, R.A.: Mitral valve regurgitation in the contemporary era: insights into diagnosis, management, and future directions. JACC: Cardiovasc. Imaging **11**(4), 628–643 (2018)
7. Huang, H., Chen, Z., Chen, C., Lu, M., Zou, Y.: Complementary consistency semi-supervised learning for 3d left atrial image segmentation. Comput. Biol. Med. **165**, 107368 (2023)
8. Huang, Y., et al.: Flip learning: weakly supervised erase to segment nodules in breast ultrasound. Med. Image Anal. **102**, 103552 (2025)
9. Huang, Y., et al.: Segment anything model for medical images? Med. Image Anal. **92**, 103061 (2024)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
11. Ivantsits, M., et al.: Mv-gnn: generation of continuous geometric representations of mitral valve motion from 3d+ t echocardiography. Comput. Biol. Med. **182**, 109154 (2024)
12. Li, S., Zhang, C., He, X.: Shape-Aware Semi-supervised 3D Semantic Segmentation for Medical Images, pp. 552–561. Springer, Heidelberg (2020)
13. Luo, G., et al.: Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound: the tdsc-abus challenge. arXiv preprint arXiv:2501.15588 (2025)
14. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 8801–8809 (2021)
15. May-Newman, K., Yin, F.: A constitutive law for mitral valve tissue (1998)
16. McDonagh, T.A., et al.: 2021 esc guidelines for the diagnosis and treatment of acute and chronic heart failure: developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the european society of cardiology (esc) with the special contribution of the heart failure association (hfa) of the esc. Eur. Heart J. **42**(36), 3599–3726 (2021)
17. Munafò, R., et al.: A deep learning-based fully automated pipeline for regurgitant mitral valve anatomy analysis from 3d echocardiography. IEEE Access (2024)
18. Munafò, R., et al.: Automatic 4d mitral valve segmentation from transesophageal echocardiography: a semi-supervised learning approach. In: Medical & Biological Engineering & Computing, pp. 1–16 (2025)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. Adv. Neural Inf. Process. Syst. **30** (2017)
21. Vahanian, A., et al.: 2021 esc/eacts guidelines for the management of valvular heart disease: developed by the task force for the management of valvular heart disease of the european society of cardiology (esc) and the european association for cardio-thoracic surgery (eacts). Eur. Heart J. **43**(7), 561–632 (2022)

22. Verma, V., et al.: Interpolation consistency training for semi-supervised learning. Neural Netw. **145**, 90–106 (2022)
23. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Mcf: mutual correction framework for semi-supervised medical image segmentation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15651–15660 (2023)
24. Yan, K., et al.: Two-shot video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2257–2267 (2023)
25. Yan, Z., et al.: A foundation model for general moving object segmentation in medical images. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2024)
26. Yu, L., Wang, S., Li, X., Fu, C.-W., Heng, P.-A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 605–613. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_67
27. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated Images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 408–416. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47

Multifrequency Neural Network-Based Wave Inversion in MR Elastography

Héloïse Bustin[1,2(✉)], Tom Meyer[2], Jakob Jordan[2], Lars Walczak[1,2,3], Heiko Tzschätzsch[2], Ingolf Sack[2], and Anja Hennemuth[1,2,3,4]

[1] Deutsches Herzzentrum der Charité, Institute of Computer-assisted Cardiovascular Medicine, Augustenburger Platz 1, 13353 Berlin, Germany
heloise.bustin@charite.de

[2] Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt Universität zu Berlin, Charitéplatz 1, 10117 Berlin, Germany

[3] Fraunhofer MEVIS, Berlin, Germany

[4] DZHK (German Centre for Cardiovascular Research), Berlin, Germany

Abstract. Magnetic Resonance Elastography (MRE) is a non-invasive imaging modality quantifying soft tissue stiffness. The reconstruction of stiffness maps is based on solutions of an inverse problem, which poses challenges in balancing accuracy, computational resources, and robustness. To stabilize the reconstruction, many inversion techniques, and most recently neural network-based inversion techniques, have explored multifrequency acquisition and reconstruction. However, these techniques typically perform separate single-frequency inversions followed by multifrequency aggregation. In this work, we propose a fully multifrequency neural network-based inversion trained on synthetically generated data that directly incorporates the relationship between multifrequency acquisitions, assuming a viscoelastic material model. Our proposed approach provides flexibility with respect to the acquisition frequencies, ensuring its practical applicability in the clinical and research setting. We evaluated our method using finite element simulations and *in vivo* abdominal MRE datasets, achieving increased accuracy and providing a more reliable and effective solution for MRE-based tissue characterization than standard reconstruction approaches.

Keywords: Magnetic resonance elastography · Multifrequency inversion · Deep learning

1 Introduction

Magnetic Resonance Elastography (MRE) is a non-invasive imaging modality quantifying soft tissue stiffness by imaging the propagation of mechanical waves through the body [17]. Tissue stiffness is a key biomarker in many pathological conditions and is increasingly used in applications such as staging liver fibrosis

I. Sack and A. Hennemuth—These authors contributed equally to this work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 418–427, 2026.
https://doi.org/10.1007/978-3-032-04947-6_40

[14]. The acquisition process involves placing an actuator on the patient producing mechanical vibrations that induce acoustic waves (Fig. 1). Motion-encoding gradients synchronized with the vibrations encode tissue displacements into the phase of the complex MR signal. This process is repeated to obtain 3D displacement wavefields at different vibration frequencies [11,12].

MRE involves solving an inverse problem to reconstruct either the shear modulus or the shear wave speed (SWS) as a surrogate for tissue stiffness from acquired wave displacement data (Fig. 1). Solutions to the inverse problem are ill-posed due to unknown boundary conditions as well as the presence of noise and compression waves in the acquired wavefields. While a variety of methods have been developed [4,5], balancing accuracy, computational cost, and robustness remains a challenge. Two prominent inversion methods used in research and clinical practice are local frequency estimation (LFE) [9] and the phase-gradient inversion method known as k-MDEV [21]. Both methods are publicly available for server-based MRE data processing at https://bioqic-apps.charite.de [13].

In recent years, neural networks have shown promising results in MRE wave inversion [3,10,15,16,19,20]. A key challenge is the lack of ground truth for training machine learning models that can translate to clinical practice. To address this, several studies have explored training on small image patches of synthetically generated data [3,10,15,19,20], including TWENN [10] and ElastoNet [3]. During inference, small image patches are extracted from the acquired image, iteratively reconstructing it, a strategy that has been shown to generalize well to *in vivo* data. Many of these machine learning-based methods rely on multifrequency acquisitions to stabilize the inversion process but are trained assuming single-frequency input data. The predicted stiffness maps for each frequency are reconstructed separately and then aggregated. Such an approach may not fully leverage the information available in multifrequency data. We hypothesize that training a neural network directly on multifrequency data would lead to a more stable inversion process and, therefore, to more accurate stiffness maps.

In this work, we introduce the first fully multifrequency neural network-based inversion method in MRE (MF-ElastoNet), which integrates multiple acquisitions at different frequencies as direct inputs to the model. The model is designed to be flexible in both the number and values of acquired frequencies to facilitate its use in research and clinical settings. We assume a viscoelastic material model for the tissue's frequency response to ensure its applicability to a wide variety of soft tissue types. We evaluate our method on finite element method (FEM) simulations and *in vivo* abdominal datasets, demonstrating the accuracy and anatomical detail of the SWS maps produced.

2 Method

2.1 Neural Network-Based Wave Inversion

Training a neural network to learn the inverse operation of reconstructing the SWS from the acquired displacement wavefields involved generating a synthetic dataset of wave displacement and shear wavelength pairs $(\boldsymbol{u}, \lambda_s)$. The shear

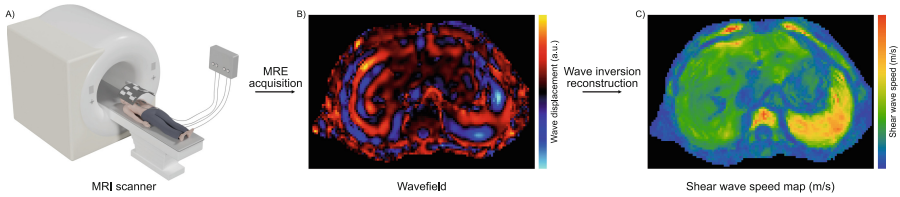


Fig. 1. MRE acquisition pipeline. A) MRI acquisition using actuators vibrating at a selected frequency. B) Wavefield acquired through encoding the wave displacements in the phase of the MR signal. C) SWS (shear wave speed) map obtained following the wave inversion reconstruction used as a proxy for the tissue stiffness.

wavelength, defined in pixel numbers, served as a proxy for the SWS to ensure that the data are independent of the image resolution Δr and vibration frequency f. The relationship is given by $c = \Delta r \cdot f \cdot \lambda_{\mathrm{s}}$, where c is the SWS. For a multifrequency acquisition with N frequencies, we have $\lambda_{\mathrm{s}} = (\lambda_1, ..., \lambda_N)$. The following minimization problem was then solved:

$$\hat{W} = \arg\min_{W} D(\lambda_{\mathrm{s}}, G(\boldsymbol{u}, W)) \tag{1}$$

where $D(\cdot, \cdot)$ is the cost function to be minimized by the neural network $G(\boldsymbol{u}, W)$ parameterized by weights W [8].

2.2 Training Data Generation

We used a linear isotropic viscoelastic material model to represent biological soft tissue [7,18], assuming local homogeneity [3,10,15]. Small patches of wave images were generated for each frequency and displacement component acquisition with the complex wave displacement $\boldsymbol{u}$ at a given location $\boldsymbol{r}$ expressed as a superposition of shear waves, compression waves, and noise:

$$\begin{aligned} u(\boldsymbol{r}) = &\sum_{j=1}^{N_{\mathrm{s}}} a_{\mathrm{s}j}\, e^{-\mathrm{i}\left(\left(\frac{2\pi}{\lambda_{\mathrm{s}j}} - \mathrm{i}\frac{1}{\delta_{\mathrm{s}j}}\right)\boldsymbol{n}(d,\theta,\varphi)\cdot\boldsymbol{r}_j + \phi_{\mathrm{s}j}\right)} \\ &+ \sum_{j=1}^{N_{\mathrm{c}}} a_{\mathrm{c}j}\, e^{-\mathrm{i}\left(\left(\frac{2\pi}{\lambda_{\mathrm{c}j}} - \mathrm{i}\frac{1}{\delta_{\mathrm{c}j}}\right)\boldsymbol{n}(d,\theta,\varphi)\cdot\boldsymbol{r}_j + \phi_{\mathrm{c}j}\right)} + \eta(\boldsymbol{r}) \end{aligned} \tag{2}$$

where a_{s}, a_{c}, ϕ_{s}, and ϕ_{c} are the amplitudes and phases of the shear and compression waves. N_{s} and N_{c} are the numbers of superimposed shear and compression wave sources in 3D space. $\boldsymbol{n}$ is the normal vector for each point source in polar coordinates (d, θ, φ). λ_{s}, λ_{c}, δ_{s}, and δ_{c} are the wavelengths and penetration lengths of the shear and compression waves in pixel numbers. $\eta(\boldsymbol{r})$ is zero-mean Gaussian noise applied to the real and imaginary parts of the wavefield. The parameter ranges were chosen as in [3].

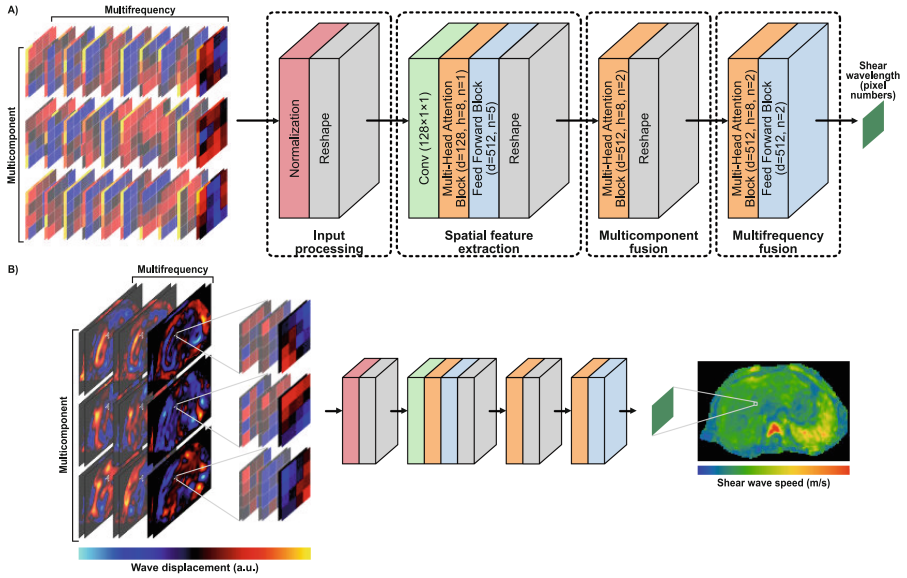


Fig. 2. A) Neural network training: Multifrequency and multicomponent complex-valued 4×4 wave displacement patches and associated shear wavelength pairs are synthetically generated for training. In the multi-head attention blocks, d is the embedded dimension, h the number of heads, and n the number of blocks. In the feed-forward blocks, d is the number of features and n the number of blocks. B) Inference: 4×4 patches are extracted from the real and imaginary parts of the acquired wavefield for every frequency and component, reconstructing the SWS (shear wave speed) maps.

2.3 Multifrequency Modeling

Several conditions were set to model the tissue's frequency response while ensuring the applicability of the method to a wide variety of acquisition protocols and tissue types.

Number of Acquired Frequencies. This number is unknown a priori and can vary significantly depending on the imaging protocol. To capture the vast majority of clinical and research use cases, the number of acquisition frequencies was uniformly sampled between one and ten for each sample.

Material Model Assumptions. We assumed that the target material follows a linear isotropic viscoelastic model, which captures the mechanical behavior of a wide variety of soft biological tissues [7]. The SWS in such tissues has been widely observed to exhibit a power-law dependence on frequency, given by $c \propto f^{\frac{\alpha}{2}}$ with $0 \leq \alpha \leq 1$, where the extremes correspond to ideal elastic and viscous behaviors, respectively [7,18]. For $\alpha = 0$, c is constant, and the wavelength follows $\lambda(f) = \frac{c}{f}$. For $\alpha = 1$, $c(f) \propto \sqrt{f}$, and the wavelength follows

$\lambda(f) \propto \frac{1}{\sqrt{f}}$. This formulation allowed us to constrain the generated wavelengths based only on the ratios between acquired frequencies and not specific values.

1. *Wavelength constraint*: We assumed N acquisition frequencies and that the associated wavelengths fall within a training range $[\lambda_{\min}, \lambda_{\max}]$. For a given acquisition n, we constrained the corresponding wavelength such that:

$$\forall n \in [1; N], \lambda_{\min} \leq \lambda_1 \leq \lambda_n \leq \lambda_N \leq \lambda_{\max} \tag{3}$$

2. *Power-law constraint*: From the material model, we derived constraints on the wavelengths and on the distance between the wavelengths associated to neighboring acquisition frequencies for the synthetic data generation:

$$\forall n \in [1, N], \lambda_{\min} \sqrt{\frac{f_1}{f_n}} \leq \lambda_n \leq \lambda_{\max} \sqrt{\frac{f_N}{f_n}} \tag{4}$$

$$\forall n \in [1, N-1], \lambda_{\min} \left(\sqrt{\frac{f_n}{f_{n+1}}} - 1 \right) \leq \lambda_{n+1} - \lambda_n \leq \lambda_{\max} \left(\frac{f_n}{f_{n+1}} - 1 \right) \tag{5}$$

3. *Penetration length constraint*: To ensure physically plausible data, the penetration length δ associated to a given wavelength λ must follow:

$$\delta \leq \frac{\lambda}{2\pi} \tag{6}$$

Acquisition Frequencies. The acquisition frequencies used in MRE vary depending on the application. In human applications, frequencies generally range from 5 to 100 Hz, with a step size of 5 or 10 Hz. In research protocols, particularly in small animals, acquisitions extend to frequencies in the range of 1000 Hz with step sizes of 100 Hz.

The multifrequency material model detailed above only sets constraints based on the relative ratios of the frequencies rather than their absolute values. Therefore, the training dataset was generated to cover a range of frequency ratios from 0.5 to 0.95 instead of specific frequencies, ensuring the applicability of the model across various acquisition protocols.

2.4 Neural Network Architecture

Each input to the neural network consisted of 4×4 complex-valued wave patches, with the real and imaginary parts stacked as channels for the three displacement components and each acquisition frequency. The complex-valued patches were normalized by the average of their modulus.

The neural network architecture, displayed in Fig. 2, followed a three-step hierarchical structure, each consisting of a vision transformer processing information at the spatial, displacement components, and frequency levels. In the spatial

feature extraction block, attention was applied along the stacked pixel dimension. In the multicomponent fusion block, it was applied along the displacement direction dimension. In the multifrequency fusion block, the frequency dimension was padded to the maximum size of ten acquisition frequencies, accommodating the variable input size. Attention was then applied along the frequency dimension, with masking corresponding to the padding.

A feed-forward block followed the multi-head attention layer in the spatial feature extraction and multifrequency fusion blocks. It consisted respectively of five and two feed-forward layers interspersed with layer normalization.

The GeLU activation function was used throughout the network, and Softplus was used as output activation to ensure physically realistic values.

2.5 Image Reconstruction

For each acquired frequency and displacement component, 4×4 patches were extracted from the acquired wavefield. The corresponding estimated SWS $\hat{c}$ was then reconstructed from the estimated shear wavelength $\hat{\lambda}_s$ in pixel numbers using the trained model:

$$\hat{c} = \Delta r \cdot f \cdot \hat{\lambda}_s = \Delta r \cdot f \cdot G(\boldsymbol{u}, W) \tag{7}$$

All reconstructed values were assembled to form the final SWS map. The training and inference reconstruction processes are illustrated in Fig. 2.

3 Experiments

3.1 Setup

Data. All datasets used are publicly available on the open-access MRE platform https://bioqic-apps.charite.de. To test our method, we used numerical FEM simulation datasets of a box with four stiff inclusions of decreasing sizes, at excitation frequencies 50, 60, 70, 80, 90, and 100 Hz with a 1 mm isotropic resolution [2] and of an abdomen at the excitation frequencies 30, 36, 42, and 48 Hz with a 2 mm isotropic resolution [1]. We also used an *in vivo* abdomen dataset acquired at 30, 40, and 50 Hz with a 2.7×2.7×5 mm^3 voxel size [6] and *in vivo* kidney datasets consisting of 23 subjects acquired at 40, 50, 60, and 70 Hz and 2.5 mm isotropic resolution [13].

Implementation Details. Our model was trained for 500,000 epochs on a single A-100 GPU, with training data generated on the fly using a batch size of 2048. The total number of training samples generated corresponds to the product of the batch size and the number of epochs. We used the Adam optimizer with a learning rate of 0.0001. We implemented the code using the PyTorch library.

Table 1. Comparison of the RMSE (root mean square error) in m/s for LFE, k-MDEV, TWENN, ElastoNet, and MF-ElastoNet in the FEM datasets in the ROIs (regions of interest).

		Classical inversions		Neural network inversions		
Dataset	ROI	LFE	k-MDEV	TWENN	ElastoNet	**MF-ElastoNet**
FEM box	inclusions	1.60	0.42	0.66	0.43	**0.38**
	matrix	**0.26**	0.27	0.85	0.43	**0.26**
FEM abdomen	liver	0.21	0.27	0.28	0.19	**0.13**
	pancreas	0.39	0.38	0.42	0.34	**0.27**
	spleen	**0.15**	0.35	0.38	0.26	0.19

Baseline. We compared MF-ElastoNet with several state-of-the-art MRE wave inversion methods: the established classical inversion methods LFE [9] and k-MDEV [21] currently used in research and clinical practice, and the recent neural network-based methods TWENN [10] and ElastoNet [3]. These methods all rely on different training protocols, inference pipelines, and algorithm choices, these design differences must be considered when comparing them.

Metrics. For the FEM datasets, where a ground truth was present, we quantified the performance of the compared methods using the root mean square error (RMSE). For the *in vivo* datasets, where no ground truth was present, we assessed the mean and standard deviation of the regions of interest (ROIs).

3.2 Results

FEM Datasets. The simulated wavefield, ground truth, and reconstructed SWS maps for our method compared to the baseline methods are displayed in Fig. 3 for the FEM box and FEM abdomen datasets. The RMSE in the ROIs are summarized in Table 1. MF-ElastoNet achieved improved accuracy, compared to the baseline methods and particularly with respect to the single-frequency trained neural network-based methods.

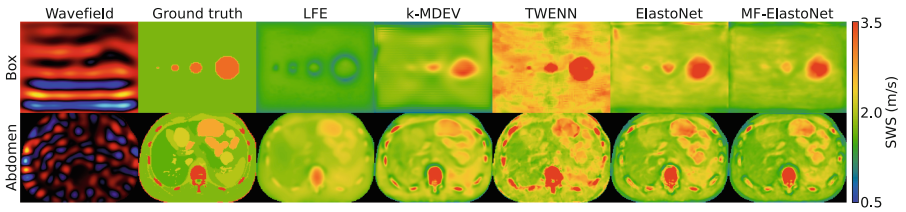


Fig. 3. Comparison of the simulated wavefield, ground truth, and predicted SWS (shear wave speed) maps produced by LFE, k-MDEV, TWENN, ElastoNet, and MF-ElastoNet for the FEM box and abdomen datasets.

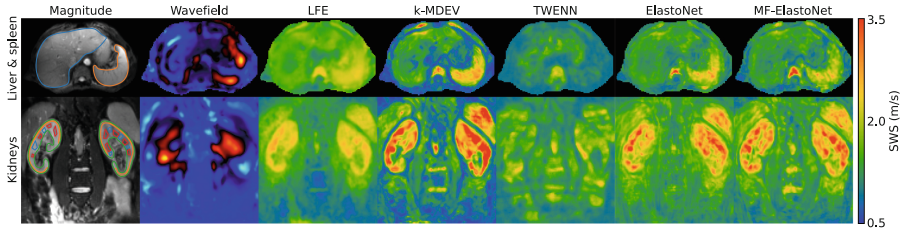


Fig. 4. Comparison of magnitude, wavefield, and predicted SWS (shear wave speed) maps produced by LFE, k-MDEV, TWENN, ElastoNet, and MF-ElastoNet for *in vivo* data: liver & spleen (liver: blue, spleen: orange) and kidneys (inner cortex: blue, medulla: red, outer cortex: orange, parenchyma: green).

Table 2. Comparison of the SWS (shear wave speed) mean ± standard deviation in m/s of LFE, k-MDEV, TWENN, ElastoNet, and MF-ElastoNet for the *in vivo* abdomen datasets in the ROIs (regions of interest).

		Classical inversions		Neural network inversions		
Dataset	ROI	LFE	k-MDEV	TWENN	ElastoNet	MF-ElastoNet
Liver & spleen	liver	1.54 ± 0.15	1.33 ± 0.30	1.25 ± 0.22	1.37 ± 0.18	1.36 ± 0.20
	spleen	2.02 ± 0.17	2.01 ± 0.49	1.10 ± 0.13	1.90 ± 0.36	1.87 ± 0.30
Kidneys	inner cortex	1.24 ± 0.07	1.91 ± 0.16	0.87 ± 0.10	1.70 ± 0.11	1.69 ± 0.16
	medulla	1.19 ± 0.07	1.52 ± 0.17	0.92 ± 0.11	1.49 ± 0.13	1.40 ± 0.10
	outer cortex	1.09 ± 0.07	1.49 ± 0.13	0.90 ± 0.11	1.17 ± 0.08	1.35 ± 0.08
	parenchyma	1.17 ± 0.06	1.56 ± 0.14	0.92 ± 0.12	1.40 ± 0.11	1.44 ± 0.11

***In vivo* datasets.** The acquired magnitude and wavefield images, as well as the reconstructed SWS maps for our method compared to the baseline methods, are displayed in Fig. 4 for the *in vivo* datasets. MF-ElastoNet showed qualitatively good anatomical detail in the reconstructed SWS maps comparable to established classical inversion methods LFE and k-MDEV. Table 2 reports the mean ± standard deviation of the liver and spleen across the ROI in the liver & spleen dataset and the inner cortex, outer cortex, medulla, and parenchyma of the kidneys across subjects in the kidneys dataset.

Ablation Study. We performed an ablation study to demonstrate the effectiveness of multifrequency training, maintaining all other training parameters unchanged. In Table 3, we report the RMSE for multifrequency training and single-frequency training with a posteriori combination for the FEM datasets. The multifrequency combination yielded a lower RMSE in the FEM datasets.

Table 3. Ablation study on multifrequency training. RMSE (root mean square error) in m/s in the FEM datasets in the ROIs (regions of interest).

Dataset	FEM box		FEM abdomen		
ROI	inclusions	matrix	liver	pancreas	spleen
Single-frequency	**0.38**	0.38	0.15	0.29	**0.19**
Multifrequency	**0.38**	**0.34**	**0.13**	**0.27**	**0.19**

4 Discussion and Conclusion

In this paper, we proposed a fully multifrequency neural network-based wave inversion method for MRE (MF-ElastoNet). The learned multifrequency combination stabilizes the inversion, showing increased accuracy and good anatomical detail compared to single-frequency neural network-based inversions [3,10] and established classical inversion methods [9,21]. Being flexible in terms of acquisition frequencies and resolution, it is designed to be suitable irrespective of the MRE protocol. While our method focuses on SWS reconstruction, further developments may include the calculation of other mechanical parameters, such as the penetration rate, the incorporation of heterogeneity within training patches, and uncertainty quantification. MF-ElastoNet expands on existing neural network-based inversion methods toward their use in research and clinical applications.

Acknowledgments. This work was supported by the German Research Foundation (CRC1340 Matrix in Vision, GRK2260 BIOQIC, FOR5628).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ariyurek, C., Tasdelen, B., Ider, Y.Z., Atalar, E.: SNR weighting for shear wave speed reconstruction in tomoelastography. NMR Biomed. **34** (2020). https://doi.org/10.1002/nbm.4413
2. Barnhill, E., Davies, P.J., Ariyurek, C., Fehlner, A., Braun, J., Sack, I.: Heterogeneous Multifrequency Direct Inversion (HMDI) for magnetic resonance elastography with application to a clinical brain exam. Med. Image Anal. **46**, 180–188 (2018). https://doi.org/10.1016/j.media.2018.03.003
3. Bustin, H., et al.: ElastoNet: neural network-based multicomponent MR elastography wave inversion with uncertainty quantification. Med. Image Anal. **105**, 103642 (2025). https://doi.org/10.1016/j.media.2025.103642
4. Doyley, M.M.: Model-based elastography: a survey of approaches to the inverse elasticity problem. Phys. Med. Biol. **57**, R35–R73 (2012). https://doi.org/10.1088/0031-9155/57/3/R35
5. Fovargue, D., Nordsletten, D., Sinkus, R.: Stiffness reconstruction methods for MR elastography. NMR Biomed. **31**, e3935 (2018). https://doi.org/10.1002/nbm.3935

6. Guo, J., et al.: Patient-activated three-dimensional multifrequency magnetic resonance elastography for high-resolution mechanical imaging of the liver and spleen. RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren. **186**(03), 260–266 (2013). https://doi.org/10.1055/s-0033-1350510
7. Hirsch, S., et al.: MR elastography of the liver and the spleen using a piezoelectric driver, single-shot wave-field acquisition, and multifrequency dual parameter reconstruction. Magn. Reson. Med. **71**, 267–277 (2013). https://doi.org/10.1002/mrm.24674
8. Kamyab, S., Azimifar, Z., Sabzi, R., Fieguth, P.: Deep learning methods for inverse problems. PeerJ Comput. Sci. **8**, e951 (2022). https://doi.org/10.7717/peerj-cs.951
9. Knutsson, H., Westin, C.F., Granlund, G.: Local multiscale frequency and bandwidth estimation. In: Proceedings of 1st International Conference on Image Processing, ICIP-94. IEEE Computer Society Press (1994). https://doi.org/10.1109/icip.1994.413270
10. Ma, S., et al.: MR elastography with optimization-based phase unwrapping and traveling wave expansion-based neural network (TWENN). IEEE Trans. Med. Imaging **42**, 2631–2642 (2023). https://doi.org/10.1109/tmi.2023.3261346
11. Mariappan, Y.K., Glaser, K.J., Ehman, R.L.: Magnetic resonance elastography: a review. Clin. Anat. **23**(5), 497–511 (2010). https://doi.org/10.1002/ca.21006
12. Meyer, T., et al.: Magnetic resonance elastography in a nutshell: Tomographic imaging of soft tissue viscoelasticity for detecting and staging disease with a focus on inflammation. Prog. Nucl. Magn. Reson. Spectrosc. **144–145**, 1–14 (2024). https://doi.org/10.1016/j.pnmrs.2024.05.002
13. Meyer, T., et al.: Comparison of inversion methods in MR elastography: an open-access pipeline for processing multifrequency shear-wave data and demonstration in a phantom, human kidneys, and brain. Magn. Reson. Med. **88**, 1840–1850 (2022). https://doi.org/10.1002/mrm.29320
14. Morisaka, H., et al.: Magnetic resonance elastography is as accurate as liver biopsy for liver fibrosis staging. J. Magn. Reson. Imaging **47**, 1268–1275 (2017). https://doi.org/10.1002/jmri.25868
15. Murphy, M.C., Manduca, A., Trzasko, J.D., Glaser, K.J., Huston, J., Ehman, R.L.: Artificial neural networks for stiffness estimation in magnetic resonance elastography. Magn. Reson. Med. **80**, 351–360 (2017). https://doi.org/10.1002/mrm.27019
16. Ragoza, M., Batmanghelich, K.: Physics-informed neural networks for tissue elasticity reconstruction in magnetic resonance elastography, pp. 333–343. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-43999-5_32
17. Sack, I.: Magnetic resonance elastography from fundamental soft-tissue mechanics to diagnostic imaging. Nat. Rev. Phys. **5**, 25–42 (2022). https://doi.org/10.1038/s42254-022-00543-2
18. Sack, I., Jöhrens, K., Würfel, J., Braun, J.: Structure-sensitive elastography: on the viscoelastic powerlaw behavior of in vivo human tissue in health and disease. Soft Matter **9**(24), 5672 (2013). https://doi.org/10.1039/c3sm50552a
19. Scott, J.M., et al.: Artificial neural networks for magnetic resonance elastography stiffness estimation in inhomogeneous materials. Med. Image Anal. **63**, 101710 (2020). https://doi.org/10.1016/j.media.2020.101710
20. Scott, J.M., et al.: Impact of material homogeneity assumption on cortical stiffness estimates by MR elastography. Magn. Reson. Med. **88**, 916–929 (2022). https://doi.org/10.1002/mrm.29226
21. Tzschätzsch, H., et al.: Tomoelastography by multifrequency wave number recovery from time-harmonic propagating shear waves. Med. Image Anal. **30**, 1–10 (2016). https://doi.org/10.1016/j.media.2016.01.001



Multi-level Gated U-Net for Denoising TMR Sensor-Based MCG Signals

Zeyu Xing[2], Hao Li[3], Hao Dou[4], Zhong Zheng[1], Jingguo Dai[5], Chen Wang[5], Jian Cui[5(✉)], Xin Zhang[1], and Tianzi Jiang[1]

[1] Beijing Key Laboratory of Brainnetome and Brain-Computer Interface, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Management, Zhejiang University, Hangzhou, China
[3] School of Biomedical Engineering and State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
[4] Shanghai University of Finance and Economics, Shanghai, China
[5] Zhejiang Lab, Hangzhou, China
cuijian@zhejianglab.com

Abstract. Tunnel magnetoresistance (TMR) sensors have been recognized as a cost-effective alternative for measuring magnetocardiography (MCG) signals. However, their relatively high noise levels and susceptibility to contamination limit their practical clinical applications. To address these challenges, we propose a novel Multi-Level Gated U-Net (MGU-Net) model specifically designed for denoising long sequential MCG signals obtained from TMR sensors. The MGU-Net leverages the U-Net architecture to learn hierarchical representations, integrated with a novel Gated Linear Unit (GLU) module to capture the periodic pattern of Q, R, and S wave complex (QRS complex) from MCG. This design enhances periodic cardiac signatures and suppresses irregular noise components through adaptive gating mechanisms. We have developed a TMR-based MCG system and collected both simulated and real MCG data in a magnetically shielded environment. The results show that our method improve signal-to-noise ratio (SNR) from −2.142 dB to 10.505 dB on the simulated MCG dataset and from 3.958 dB to 14.514 dB on the real dataset, surpassing other state-of-the-art methods. Our model successfully recovers subtle P-wave and T-wave features from the noisy signals, illustrating a promising direction of using TMR-based systems for potential practical clinical applications.

Keywords: Magnetocardiography (MCG) · Denoising · Tunnel Magnetoresistance · Gated Linear Unit

1 Introduction

Magnetocardiography (MCG) has emerged as a critical non-invasive diagnostic modality for analyzing cardiac electrical activity through its associated magnetic field patterns. Current MCG measurement technologies primarily rely on

Z. Xing and H. Li—Contributed equally.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 428–437, 2026.
https://doi.org/10.1007/978-3-032-04947-6_41

superconducting quantum interference devices (SQUIDs) [3] or optically pumped magnetometers (OPMs) [1], each presenting distinct advantages and limitations. SQUID systems demonstrate exceptional sensitivity with ultra-low noise levels ($\sim$2–5 fT/$\sqrt{\text{Hz}}$) [15,20] and superior low-frequency stability, making them the gold standard for clinical MCG applications. However, their operational dependence on liquid helium cooling systems and substantial equipment costs ($\sim$ \$1M) significantly limit widespread clinical adoption. OPM technology offers comparable magnetic field resolution to SQUIDs while eliminating cryogenic requirements, but introduces new challenges including complex optical system configurations, precise alkali vapor cell management, and mandatory magnetic shielding for optimal performance. These technical demands result in increased maintenance complexity and operational costs that hinder practical implementation. In comparison, tunnel magnetoresistance (TMR) sensors with miniaturized semiconductor designs have been considered as a cost-effective alternative technique for measuring MCG [6,14]. However, their relatively high noise ($\sim$2200 pT/$\sqrt{\text{Hz}}$) [10,14] and susceptibility to noise contamination restrict their practical clinical applications.

To obtain clean magnetocardiography (MCG) signals using TMR sensors, efforts have been made to use traditional methods such as digital filters [16], AC modulation [22], and empirical mode decomposition (EMD) [14]. However, their effectiveness remains limited due to inherent constraints in handling non-stationary noise and preserving subtle cardiac features. Recent advances in deep learning have shown promise for electrocardiogram (ECG) denoising through architectures like DeepFilter (a multipath convolutional network) [17], DeScoD-ECG (diffusion-based model) [12], and Transformer-based TCDAE [2]. However, critical differences exist between ECG and MCG noise profiles: ECG denoising primarily targets baseline drift (0.05–2 Hz), electrode motion artifacts (0.1–10 Hz), and muscle noise (5–500 Hz), whereas TMR-based MCG systems predominantly exhibit 1/f electrical noise spanning 0.1–100 Hz with non-uniform spectral decay. Current deep learning approaches achieve a significant improvement on ECG datasets, but demonstrate suboptimal performance when applied to MCG signals due to mismatched noise characteristics and sensor-specific artifacts.

In this paper, we proposed an Multi-Level Gated U-Net (MGU-Net) model specifically for denoising MCG signals acquired by TMR sensors. We leverage the U-Net architecture to perform multi-scale feature extraction, combined with a novel gated linear unit (GLU) to enhance cardiac feature extraction and attenuate non-periodic noise by capitalizing on the periodicity inherent in the MCG signal. The proposed model has an excellent performance on the task of denoising TMR-based MCG signals, which opens up a promising direction of using TMR-based systems for potential practical clinical applications.

2 Method

We start with an analysis of noise characteristics in MCG signals. Figure 3(a) displays an episode of single-channel raw MCG signal acquired using the TMR sensor. As shown, the noise exhibits random fluctuations with irregular amplitude

and frequency variations, indicative of white noise originating from electronic systems (e.g., thermal agitation in sensor components) and 1/f noise inherent to TMR sensors in low-frequency domains. Crucially, subtle cardiac features such as P-waves and T-waves are mostly obscured by noise, with only the R-peak component faintly observable in specific segments. This observation highlights the challenge of extracting clean signals from a single cardiac cycle. To address this, we construct datasets with samples of around 10-second length to provide sufficient periodic information of Q, R, and S wave complex (QRS complex) for the model to learn. For processing such long-sequence signals, our model incorporates two critical design principles:

Hierarchical Feature Extraction: Direct processing of full-length MCG sequences (typically containing ∼10 cardiac cycles) poses significant computational challenges due to their extended duration and morphological complexity. To address this, we implement a multi-scale feature extraction strategy that progressively learns both localized waveform details and global contextual patterns. Drawing inspiration from successful applications in high-resolution image processing [13,21], our framework employs a U-Net architecture to facilitate comprehensive feature learning across multiple temporal resolutions, ensuring robust representation of both transient cardiac events and sustained rhythm characteristics.

Periodic Pattern Recognition: The most distinctive characteristic of MCG signals lies in the periodically occurring QRS complexes within extended signal sequences. To effectively utilize this periodicity for enhanced cardiac feature extraction and non-periodic noise suppression, our model incorporates a GLU designed specifically to capture long-range dependencies and reinforce these recurring patterns. This architecture enables systematic amplification of periodic cardiac signatures while attenuating irregular noise components through adaptive gating mechanisms.

2.1 The U-Net Architecture

As it can be seen in Fig. 1(a) Our model employs a U-Net architecture integrated with GLUs at multiple hierarchical levels to capture periodic patterns for denoising. The network incorporates two GLU variants: Competitive Gating (CG) modules and Noise Gating (NG) modules. The input signal first traverses an NG module, expanding its channel dimension from 1 to 64. Subsequent processing involves four downsampling stages, each comprising a ResBlock, a CG module, and a downsampling layer. Each downsampling operation halves the signal length while doubling the channel dimension.

At the bottleneck stage, the compressed signal undergoes feature aggregation across hierarchical layers, enabling the learning of high-level representations. Following this, three upsampling stages progressively restore spatial resolution. During upsampling, features from corresponding encoder stages are concatenated with the decoder pathway, followed by processing through ResBlock and CG

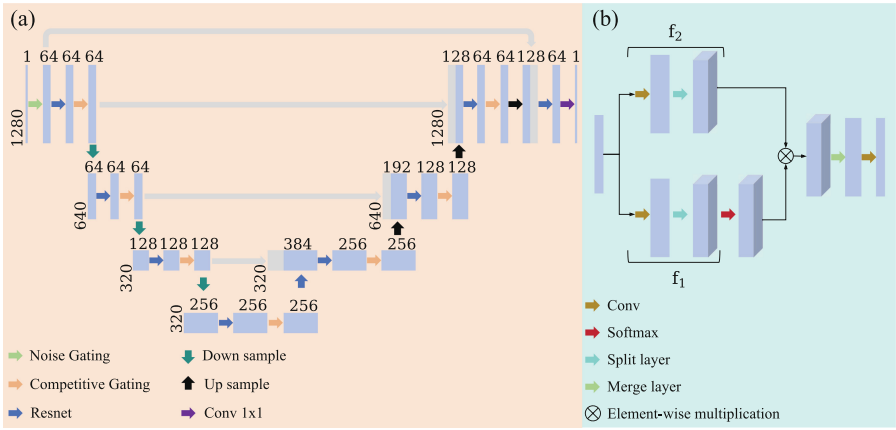


Fig. 1. Structure of the proposed Multi-Level Gated U-Net model. (a) The whole architecture (b) The GLU module.

modules . This multi-scale feature fusion mechanism preserves critical temporal relationships through skip connections while mitigating information loss.

After the final upsampling stage, the signal dimension reaches 64 × 1280. A 1×1 convolutional layer then reduces the channel dimension to 1, yielding the denoised output 1 × 1280. In this model, the convolutions were implemented with kernel sizes of 9 for the resnet block, 7 for the gating branch, and 4 for down/upsampling and followed by RMSNorm and SiLU activation. This hierarchical design achieves efficient information compression and reconstruction, balancing computational efficiency with robust feature extraction across transient cardiac events and sustained rhythm patterns.

2.2 The GLU Module

An intuitive approach for learning the periodic patterns in MCG signals is to employ the self-attention(SA) mechanism [18], which effectively models long-range dependencies to capture amplitude variations and inter-cycle interval regularity. Given an input MCG feature sequence $X_{\text{in}} \in \mathbb{R}^{T\times D}$, where T denotes the sample length and D the feature dimension, self-attention transforms X_{in} through learnable parameter matrices $W_Q, W_K, W_V \in \mathbb{R}^{D\times d_k}$ to obtain query (Q), key (K), and value (V) matrices $Q = X_{in}W_Q$, $K = X_{in}W_K$, $V = X_{in}W_V$. The output is computed as:

$$X_{\text{out}} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1}$$

We replace the $QK^\top$ computation with learnable linear projections, resulting a gated linear unit (GLU) [4]:

$$X_{\text{out}} = \sigma\left(f_1(X_{\text{in}};\theta_W)\right) \odot f_2(X_{\text{in}};\theta_V) \tag{2}$$

where f_1 and f_2 are linear mappings parameterized by θ_W and θ_V, respectively, and σ denotes the activation function. The replacement of SA with GLU is primarily driven by the periodic nature of MCG signals. As MCG exhibits strong self-correlation patterns and amplitude-dependent dependencies, GLU's gating mechanism (via element-wise multiplication of two linear projections) inherently captures these global periodic features. In contrast, SA requires separate computation of Query (Q) and Key (K) introducing redundant parameters (e.g., Q/K projection layers), which could lead to suboptimal convergence of the model.

As shown in Fig. 1(b), both f_1 and f_2 utilize the same pipeline to process X_{in}. Each pipeline comprises a convolutional layer followed by a multi-head split layer. The standard GLU with σ as the softmax activation function is called competitive gating module in our paper, as it calculates gating weights globally through Softmax to capture the global dependencies of the signal, allowing periodic features such as QRS complexes to receive higher weights throughout the sequence. In the first layer, the σ is implemented as the sigmoid function, as it is designed to perform preliminary processing of MCG signals with low signal-to-noise ratio (SNR) by suppressing random noise and enhancing key features. We call this variant of standard GLU as noise gating module. Our code is available at https://github.com/YorkXingZeyu/MCG-denoising-project.git.

3 Experiments

3.1 Data Acquisition

TMR-Based MCG System: We designed a TMR sensor consisting of thousands of MTJs (Fig. 2(a)) connected in series and parallel for MCG signal acquisition. Four TMR sensors were arranged in a Wheatstone bridge configuration (Fig. 2(b)) and encapsulated for protection to form a complete TMR sensor. To amplify the signals, we designed and fabricated an ultra-low noise amplification circuit board (Fig. 2(c)). The equivalent input noise at 1 Hz is approximately 10 nV/$\sqrt{\text{Hz}}$, which is lower than the noise of the TMR sensor itself. The TMR sensor is connected the amplification board to form a TMR probe. Additionally, we created a four-channel adapter module (Fig. 2(d)) to connect the TMR probe and output the signals to the DAQ for real-time data visualization and recording on a PC. The overall system noise level was reduced to 3–5 pt/$\sqrt{\text{Hz}}$ through optimized system design, enabling high-sensitivity detection of MCG signals.

Simulated MCG Dataset: We developed a simulated magnetic signal generation system (Fig. 2(e)) that uses a Helmholtz coil to convert electrical signals from an arbitrary waveform generator (AWG) into magnetic signals. Inside a magnetically shielded barrel, TMR probes capture the magnetic field generated by the Helmholtz coil, with the magnetic field strength serving as the ground truth. The input signals to the AWG are derived from preprocessed data in the Kiel Cardio database [7] from seven healthy male volunteers. The primary types of noise of clean signals from Kiel Cardio DB are thermal noise, optical system noise, and shot noise (<15 fT/$\sqrt{\text{Hz}}$), which are negligible after filtering,

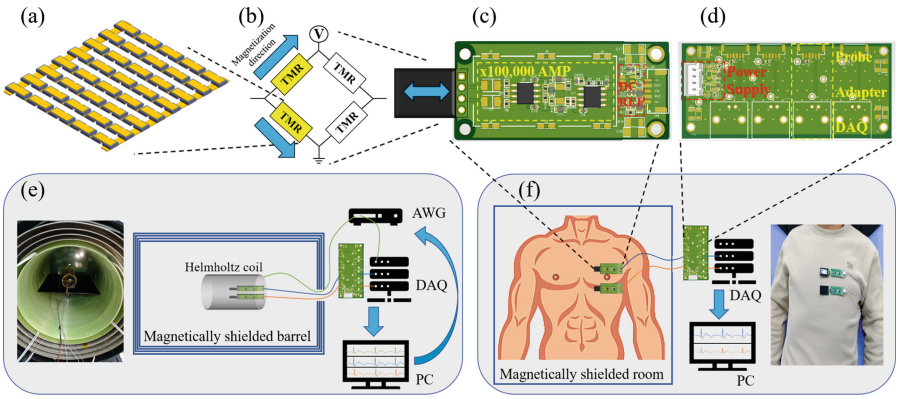


Fig. 2. TMR-based MCG system and experiment setups. (a) A TMR sensor array composed of magnetic tunnel junctions (MTJs). (b) Four TMR elements in a Wheatstone bridge configuration for converting magnetic signals into differential voltage. (c) TMR sensor integrated with a 100 dB amplifier and power supply circuit. (d) 4-channel adapter module connecting the power supply and amplifier to the data acquisition system (DAQ). (e) The experiment setup for obtaining simulated data. (f) The experiment setup for obtaining real data.

in comparison to the captured noise from the TMR system with levels around 5 pT/$\sqrt{\text{Hz}}$. The data is segmented into 10-second samples. Each sample contains several complete cardiac cycles, with the peak-to-peak range linearly scaled between 80 pT and 200 pT. A total of 6957 samples were obtained.

Real MCG Dataset: The experiment was conducted in a magnetically shielded room. Ten healthy male subjects aged between 18 and 35 years participated in the experiment. During data collection, each one remained stationary in a supine position (Fig. 2(f)), and probes were placed above their chest to acquire MCG signals. This study adheres to the ethical guidelines of the Chinese Academy of Sciences Institute of Automation (CASIA), China and was approved by Human Subjects Research Ethics Review Committee of CASIA (Ref: IA21-2302-430203). The acquired signals were filtered using a 1–35 Hz bandpass filter. Subsequently, the signals were averaged over 101 cardiac cycles, using the R-wave peak as a reference to obtain the ground truth. The data from the TMR sensor were then segmented into 10-second episodes, each containing multiple complete cardiac cycles. A total of 2868 samples were obtained.

3.2 Results

We compared MGU-Net with other methods based on four metrics, which are Sum of the Square of the Distances (SSD), Maximum Absolute Distance (MAD), Cosine Similarity (Cosine Sim), and Signal-to-Noise Ratio (SNR) [12]. The "Baseline" in Table 1 was implemented by obtaining an average of the cardiac cycles in a sample and then rearranging it to align with the peaks to form the same

Table 1. Comparison on the Simulated and MCG Real Datasets.

Datasets	Simulated MCG Dataset				Real MCG Dataset			
Metrics	SNR ↑	SSD ↓	MAD ↓	Cosine Sim ↑	SNR ↑	SSD ↓	MAD ↓	Cosine Sim ↑
FIR Filter [9]	−0.5462	6.5051	0.2762	0.6139	4.3598	4.9430	0.2185	0.8231
IIR Filter [9]	−0.5752	6.5444	0.2763	0.6179	4.2813	4.9999	0.2215	0.8347
EMD [11]	−1.2439	7.7633	0.2992	0.5732	3.6089	5.8973	0.2370	0.8195
EEMD [19]	0.4674	5.3164	0.2648	0.5884	3.9274	5.4687	0.2312	0.8306
VMD [5]	0.2766	5.6047	0.3026	0.4544	4.5735	4.7173	0.2179	0.8308
APR-CNN [8]	7.5702	1.0901	0.1993	0.9097	10.0929	1.0469	0.1261	0.9485
TCDAE [2]	2.6863	3.2524	0.2819	0.6683	5.9516	3.4829	0.2118	0.8509
RA-LENet [23]	2.1276	3.8124	0.3207	0.6946	6.3643	2.3523	0.2101	0.8756
DeScoD [12]	7.7682	1.0632	0.2251	0.8988	8.3049	2.0278	0.1593	0.9214
Baseline	4.2304	2.2240	0.2351	0.7968	11.2663	1.0571	0.1470	0.9530
Our Model	**10.5051**	**0.5709**	**0.1511**	**0.9465**	**14.5143**	**0.4717**	**0.0699**	**0.9815**

length as the original sample. Five-fold cross validation was conducted, where 0.1 of the training data were used for validation. All the deep-learning models except for DeScoD were trained for 200 epochs with early stopping and a learning rate of 0.0001. For DeScoD, a higher learning of 0.001 was used according to its original paper. A batch size of 32 was used for all the models. For all the deep-learning based models, we use the Adam optimizer with a learning rate of 0.0001 and batch size was set as 32. The optimization process was guided by the mean squared error (MSE) loss.

As it can be seen in Table 1, the proposed MGU-Net method outperforms all the other methods based on all of the four evaluation metrics. The first five traditional methods have incremental improvement on the data due to their inherent constraints in handling non-stationary noise and preserving subtle cardiac features. All of the comparison methods fail to outperform Baseline on the Real Dataset based on SNR, which indicates that current state-of-the-art methods fails to capture the periodic pattern in the noisy MCG signals. In comparison, our proposed model combining U-net structure with GLU module can successfully capture essential information for denoising periodical MCG signals.

As shown in Fig. 3, the Baseline fails to filter the high-frequency noise from the signal. DeScoD and APR-CNN fail to restore the QRS complex in several cardical cycles. In contrast, only our proposed MGU-Net model achieves the best performance in the denoising task. It accurately restores every QRS complex with high fidelity. Low-amplitude features (like P-waves and T-waves), are also clearly accurately restored in the sample. In aspect of clinical utility, MGU-Net is lightweight and efficient, with 16.13 million parameters, 6.6 GFLOPs, and a real-time inference speed of 5.06 ms (versus DeScoD: 5.31 ms and APR-CNN: 6.42 ms) per sample on an RTX 4090 GPU, making it competitive for practical deployment.

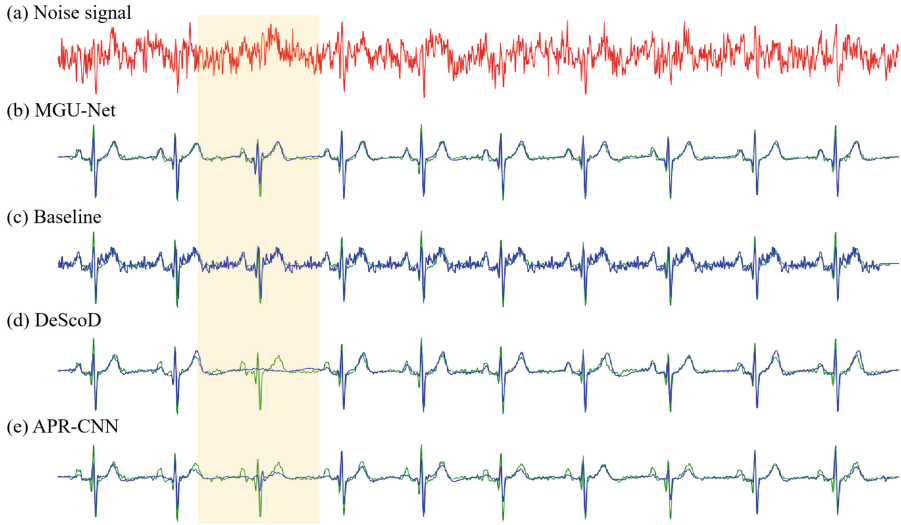


Fig. 3. Denoising performance of different models on a single sample from the simulated dataset. The red line represents the noisy input signal, the green line represents the ground-truth signal, and the blue line represents the denoised signal. (a) The original noisy MCG signal. (b) Denoised signal by MGU-Net. (c) Denoised signal by Baseline. (d) Denoised signal by DeScoD. (e) Denoised signal by the APR-CNN model. (Color figure online)

Table 2. Ablation studies of the proposed model on the simulated and real MCG datasets. The impact of Competitive Gating (CG) and Noise Gating (NG) modules are evaluated.

Dataset	CG	NG	SA	SSD ↓	MAD ↓	Cosine Sim ↑	SNR Input	SNR Output ↑	SNR Improvement ↑
Simulated MCG Dataset	×	×	×	1.6518	0.2366	0.8377	−2.1423	5.6854	7.8277
	✓	×	×	1.1218	0.2037	0.8932	−2.1423	7.3937	9.5360
	×	✓	×	0.8246	0.1666	0.9223	−2.1423	8.9113	11.0536
	×	×	✓	0.6564	0.1664	0.9363	−2.1423	9.8074	11.9498
	✓	✓	×	**0.5709**	**0.1511**	**0.9465**	−2.1423	**10.5051**	**12.6474**
Real MCG Dataset	×	×	×	3.3924	0.2775	0.8213	3.9576	4.8834	0.9258
	✓	×	×	2.1686	0.2356	0.9128	3.9576	7.7366	3.7790
	×	✓	×	1.0615	0.1487	0.9575	3.9576	11.0431	7.0855
	×	×	✓	0.5118	0.1096	0.9769	3.9576	14.1827	10.2251
	✓	✓	×	**0.4717**	**0.0699**	**0.9815**	3.9576	**14.5143**	**10.5567**

Ablation Study. In the ablation study (Table 2), we compare the impact of CG and NG on model performance. Using CG alone improves SNR of the plain model by 1.71 dB (Simulated) and 2.85 dB (Real). Using NG alone results in a more significant SNR improvement of the plain model (3.22 dB for Simulated and 6.16 dB for Real), with reductions in SSD and MAD, indicating NG effectively suppresses noise. Combining CG and NG yields the best results: SNR improves to 12.65 dB (Simulated) and 10.56 dB (Real), demonstrating the synergistic

effect of both modules. The performance of using SA is inferior to that of using the combination of CG and NG.

4 Conclusion

This paper presents a novel deep learning model named MGU-Net for denoising MCG signals acquired using TMR sensors. By integrating noise gating and competitive gating mechanisms to capture the periodic patterns, MGU-Net effectively suppresses noise while preserving key cardiac features of MCG. Experimental results demonstrate that our method outperforms state-of-the-art denoising models in recovery of low-amplitude P-wave and T-waves, which are clinically essential for diagnosing arrhythmias, myocardial ischemia, and other cardiac abnormalities, in a high-noise real MCG recording. This work provides a foundation for the broader application of TMR-based MCG in clinical diagnostics and pathological analysis, offering a cost-effective and scalable alternative for cardiac monitoring.

Our current study was conducted only on a limited of subjects, primarily healthy individuals with normal MCG forms. To address the limitation, we will expand the scope of future research by incorporating a more diverse cohort of participants, including patients with cardiac diseases. Accordingly, we will prioritize improving the model's stability and consistency (e.g., exploration on multi-loss frameworks) across different populations.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (62327805, 62336007) and the STI2030-Major Projects (2021ZD0200201).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alexandrov, E.B., Bonch-Bruevich, V.A.: Optically pumped atomic magnetometers after three decades. Opt. Eng. **31**(4), 711–717 (1992)
2. Chen, M., et al.: Elimination of random mixed noise in ECG using convolutional denoising autoencoder with transformer encoder. IEEE J. Biomed. Health Inform. (2024)
3. Clarke, J.: Squid fundamentals. In: SQUID Sensors: Fundamentals, Fabrication and Applications, pp. 1–62. Springer, Cham (1996)
4. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: International Conference on Machine Learning, pp. 933–941. PMLR (2017)
5. Dragomiretskiy, K., Zosso, D.: Variational mode decomposition. IEEE Trans. Signal Process. **62**(3), 531–544 (2013)
6. Fujiwara, K., et al.: Magnetocardiography and magnetoencephalography measurements at room temperature using tunnel magneto-resistance sensors. Appl. Phys. Express **11**(2), 023001 (2018)

7. Goldberger, A.L., et al.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
8. He, Z., Liu, X., He, H., Wang, H.: Dual attention convolutional neural network based on adaptive parametric RELU for denoising ECG signals with strong noise. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 779–782. IEEE (2021)
9. Heinonen, P., Saramaki, T., Malmivuo, J., Neuvo, Y.: Periodic interference rejection using coherent sampling and waveform estimation **31**(5), 438–446 (1984)
10. Hou, Y., Wang, D.F., Itoh, T.: Maximizing modulation efficiency to minimize 1/f noise in magnetoresistance. Measurement **207**, 112396 (2023)
11. Huang, N.E., et al.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. London. Ser. A: Math. Phys. Eng. Sci. **454**(1971), 903–995 (1998)
12. Li, H., Ditzler, G., Roveda, J., Li, A.: DeScoD-ECG: deep score-based diffusion model for ecg baseline wander and noise removal. IEEE J. Biomed. Health Inform. (2023)
13. Liang, J., et al.: NTIRE 2024 restore any image model (RAIM) in the wild challenge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6632–6640 (2024)
14. Lu, Z., Ji, S., Yang, J.: Measurement of t wave in magnetocardiography using tunnel magnetoresistance sensor. Chin. Phys. B **32**(2), 020703 (2023)
15. Maslennikov, Y.V., et al.: The DC-SQUID-based magnetocardiographic systems for clinical use. Phys. Procedia **36**, 88–93 (2012)
16. Oogane, M., et al.: Sub-PT magnetic field detection by tunnel magneto-resistive sensors. Appl. Phys. Express **14**(12), 123002 (2021)
17. Romero, F.P., Piñol, D.C., Vázquez-Seisdedos, C.R.: Deepfilter: an ECG baseline wander removal filter using deep learning techniques. Biomed. Signal Process. Control **70**, 102992 (2021)
18. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
19. Wu, Z., Huang, N.E.: Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv. Adapt. Data Anal. **1**(01), 1–41 (2009)
20. Yang, K., et al.: Weakly damped squid gradiometer with low crosstalk for magnetocardiography measurement. IEEE Trans. Appl. Supercond. **26**(8), 1–5 (2016)
21. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5728–5739 (2022)
22. Zhao, W., Tao, X., Ye, C., Tao, Y.: Tunnel magnetoresistance sensor with AC modulation and impedance compensation for ultra-weak magnetic field measurement. Sensors **22**(3), 1021 (2022)
23. Zhu, Y., Zhu, D., Liu, J.: Ra-lenet: R-wave attention and local enhancement for noise reduction in ECG signals. In: 2024 International Joint Conference on Neural Networks (IJCNN), pp. 1–9. IEEE (2024)



Multiscale Graph and Multi-step Cross-Frame Mamba for Myocarditis Lesion Segmentation

Chengjin Yu[1], Hao Zhang[2], Yuanting Yan[2(✉)], Dong Zhang[3(✉)], Sangyin Lv[4], and Cailing Pu[5]

[1] School of Big Data and Statistics, Anhui University, Hefei, China
[2] School of Computer Science and Technology, Anhui University, Hefei, China
ytyan2016@163.com
[3] School of Biomedical Engineering, Anhui Medical University, Hefei, China
dongzhang@ahmu.edu.cn
[4] Shaoxing People's Hospital, Shaoxing, China
[5] Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China

Abstract. Myocarditis, an acute cardiac disorder progressing rapidly to life-threatening heart failure, requires precise lesion segmentation from Cine Magnetic Resonance Imaging (Cine-MRI) for timely intervention. Current segmentation accuracy is limited by two key challenges: 1) spatiotemporal discordance between myocardial motion patterns and evolving pathological features and 2) morphological complexity (irregular borders, scattered lesions). In this paper, we propose the MG-Mamba, a framework integrating deep state space models with graph-based spatiotemporal analysis. The architecture employs Mamba blocks to establish initial intra-/inter-frame dependencies in Cine-MRI sequences. For Challenge 1, we improve the detection of subtle abnormal motions through multi-step cross-frame analysis, extending beyond conventional adjacent-frame analysis. For Challenge 2, we further implement multi-scale patch division and constructs inter-patch graphs to concurrently capture global lesion distribution and local geometric patterns. Extensive evaluations on SYC-QC and SYC-SX clinical datasets demonstrate MG-Mamba's superior segmentation accuracy over ten state-of-the-art benchmarks, significantly advancing myocarditis diagnostic precision. The code is available at https://github.com/userZ-CY/MICCAI.

Keywords: Myocarditis · Cine-MRI · Graph · Mamba

1 Introduction

Timely and accurate segmentation of myocarditis lesions is crucial for treatment and patient prognosis [1]. Myocarditis is a group of diseases characterized primarily by inflammation of the myocardium [2]. The myocarditis progresses

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 438–447, 2026.
https://doi.org/10.1007/978-3-032-04947-6_42

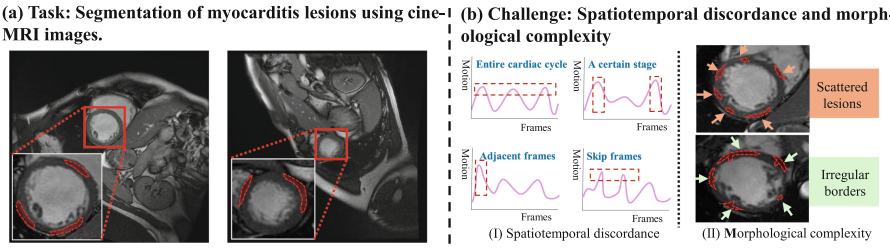


Fig. 1. (a): Segmenting myocarditis lesions using Cine-MRI images can determine the size, distribution, and severity of inflammation. This can help doctors provide targeted treatment. (b): (I) The abnormal myocardial motion caused by myocarditis can occur at different times. (II) The myocarditis lesions are unevenly distributed and irregularlyshaped

rapidly and it might lead to severe complications, including arrhythmias, heart failure, and cardiogenic shock [3]. The Lake Louise criteria are commonly used in clinical practice to diagnose myocarditis, including T2-weighted imaging, early gadolinium-enhanced imaging, and late gadolinium-enhanced imaging [4]. However, the nephrotoxicity of gadolinium contrast agents and the prolonged scan time limit their clinical application [5]. In contrast, Cine-MRI imaging technology eliminates the need of contrast agents and offers shorter scan time [6], providing a potentially safer and more effective approach for the segmentation of myocarditis lesions, as shown in Fig. 1(a).

However, due to the inherent technical limitations ingrained in Cine-MRI, it demonstrates insufficient contrast when imaging myocarditis lesions. In recent years, the researches on using deep learning to extract spatiotemporal information from Cine-MRI images for segmenting cardiomyopathy regions has achieved inspiring results [7]. Nevertheless, given the morphological complexity of the myocarditis lesions and spatiotemporal discordance of the myocarditis motion patterns, existing methods might not be directly applicable to the myocarditis segmentation. First, the motion patterns of myocarditis patients exhibit significant spatiotemporal inconsistency: the occurrence of abnormal motion is unpredictable; the compensatory mechanisms of healthy myocardium can mask subtle abnormalities in affected regions, as shown in Fig. 1(b) I. Then, as shown in Fig. 1(b) II, due to the infiltration patterns of inflammatory cells (either localized or diffuse) [8], myocardial inflammation lesions typically exhibit irregular shapes and scattered spatial distributions, which pose challenges in learning both the local structure information and global distribution information.

To address the aforementioned challenges, we propose a MG-Mamba model for myocarditis lesion segmentation. We first divides the image into different patches at multiple scales. Then, for Challenge 1, we employ a Multi-step Cross-frame Mamba Motion Analysis (MCMMA) module in each scale. This module designs multi-step cross-frame scanning sequences, utilizing the Mamba network's state space modeling capability to scan inter-frame relationships in vary-

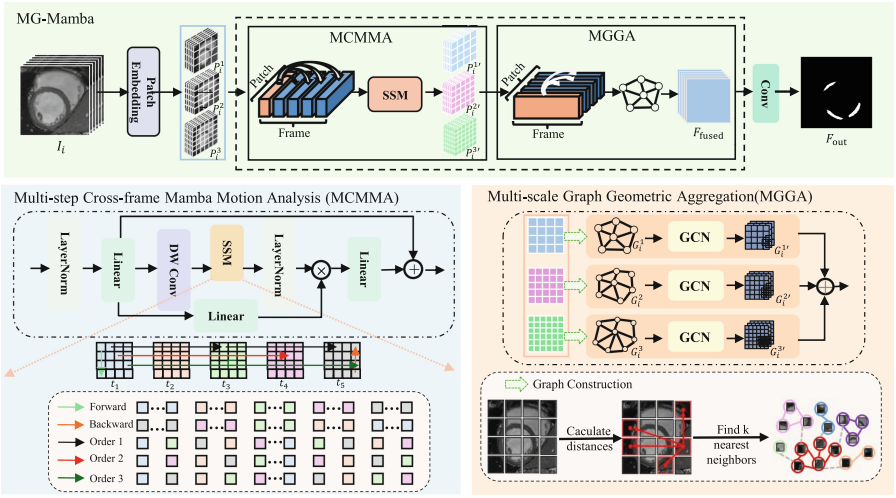


Fig. 2. The proposed MG-Mamba model consists of two main modules: MCMMA and MGGA. The MCMMA module captures abnormal myocardial motion features, while the MGGA module aggregates multi-scale geometric features of the lesion

ing orders. This approach effectively detects subtle pathological motion patterns caused by myocarditis. For Challenge 2, we propose a Multi-scale Graph Geometric Aggregation (MGGA) module. This module first divides the image into different patches and constructs a graph by computing the Euclidean distances between intra-frame patches. Subsequently, we employ Graph Convolutional Networks (GCN) to aggregate geometric features across different scales, enabling the simultaneous capture of both local structural patterns and global morphological distributions of myocardial lesions.

Our contributions are summarized as follows: 1) We present a clinical method for segmentation of myocarditis lesions using Cine-MRI images. This method has significant clinical potential for myocarditis lesion segmentation without contrast agents and shorter scan times. 2) We propose a MCMMA module to address the spatiotemporal inconsistencies in myocardial motion through multi-step cross-frame state space modeling, and a MGGA module to address the morphology complexity by constructing multi-scale graphs that aggregate global lesion distribution and local geometric information. 3) Extensive experiments conducted on two clinical datasets (10,400 Cine-MRI images in total) demonstrate that the MG-Mamba is effective and superior to ten state-of-the-art methods.

2 Method

Overall Pipeline. Our MG-Mamba framework combines two key components: the Multi-step Cross-frame Mamba Motion Analysis (MCMMA) module (Sect. 2.1), designed to identify abnormal motion patterns via multi-step cross-frame state space modeling, and the Multi-scale Graph Geometric Aggregation

(MGGA) module (Sect. 2.2), which extracts lesion geometric features through multi-scale graph learning. The overall workflow is depicted in Fig. 2.

2.1 Multi-step Cross-Frame Mamba Motion Analysis (MCMMA)

To detect subtle pathological motion patterns caused by myocarditis, we propose the MCMMA module, which captures motion abnormalities by multi-step cross-frame state space modeling. We first create multi-step inter-frame scanning sequences by spanning varying numbers of frames. Then, we employ Mamba [9] model to scan created sequences extract motion features.

Specifically, for each frame $I_i \in \mathbb{R}^{H\times W}$ (with $i = 1, 2, \ldots, T$), H denotes the height, W denotes the width and T denotes the frames(in this experiment, $T = 20$.). Then we divide it into image patches at different scales of 8×8, 10×10, and 16×16. Through Patch Embedding (PE) operations, we generate patch embeddings $P_i^1 \in \mathbb{R}^{n_1\times d_1}$, $P_i^2 \in \mathbb{R}^{n_2\times d_2}$, and $P_i^3 \in \mathbb{R}^{n_3\times d_3}$, where n_1, n_2, and n_3 denote the number of patches, and d_1, d_2, and d_3 denote the embedding dimensions.

$$P_i^1 = PE^{8\times 8}(I_i), \quad P_i^2 = PE^{10\times 10}(I_i), \quad P_i^3 = PE^{16\times 16}(I_i) \tag{1}$$

Next, for the multi-scale patches P_i^1, P_i^2, P_i^3, we sequentially process the features $P \in \mathbb{R}^{T\times n\times d}$ obtained from the T frames of data at each scale, where T denotes the number of frames, n denotes the number of patches, d denotes the embedding dimensions.

Afterward, at each scale, we employ the Mamba model to learn inter-frame dependencies of the P_i^j=$\{P_1^j, P_2^j, P_3^j, \ldots, P_T^j\}$. To enhance the motion modeling capability of our block, we introduce a triple-branch architecture that operates in parallel: a forward SSM branch, a backward SSM branch, and an inter-frame SSM branch.

At j-th scale, we first perform forward SSM to scan from the first frame to the T-th frame, while backward SSM to scan from the T-th frame to the first frame.

$$\widetilde{P}_f = SSM(f) : P_1^j \rightarrow P_2^j \rightarrow \ldots \rightarrow P_{T-1}^j \rightarrow P_T^j \tag{2}$$

$$\widetilde{P}_b = SSM(b) : P_T^j \rightarrow P_{T-1}^j \rightarrow \ldots \rightarrow P_2^j \rightarrow P_1^j \tag{3}$$

For inter-frame SSM, we scan across frames by varying the frame interval ($\Delta_t = 2, 3, 5$):

$$\widetilde{P}_{\Delta_t} = SSM(\Delta_t) : P^j_{[1][x]} \rightarrow P^j_{[1+\Delta_t][x]} \rightarrow P^j_{[1+2*\Delta_t][x]} \rightarrow \ldots P_T^j \tag{4}$$

where $[a][b]$ indicate the b-th patch embedding in a-th frame.

After, we fuse the output of the forward, backward and inter-frame SSM:

$$X = \mathrm{LN}(\widetilde{P}_f \odot \widetilde{P}_b \odot \widetilde{P}_{\Delta_{t=2}} \odot \widetilde{P}_{\Delta_{t=3}} \odot \widetilde{P}_{\Delta_{t=5}}) \tag{5}$$

where $\odot$ denotes the Hadamard product. Given three different scales, we can get ($X^1 \in \mathbb{R}^{T\times n_1\times d_1}, X^2 \in \mathbb{R}^{T\times n_2\times d_2}, X^3 \in \mathbb{R}^{T\times n_3\times d_3}$).

2.2 Multi-scale Graph Geometric Aggregation(MGGA)

To capture the morphological complexity of myocarditis lesions, we propose the MGGA module, which learns geometric features through multi-scale graph learning. Specifically, we further model the geometric features of the multi-scale embeddings (X^1, X^2, X^3) in each frame. Take t-th frame $X_t^1 = \{x_{t1}^1, x_{t2}^1, ...x_{tn_1}^1\}$ as an example, each patch embedding x_{t1}^1 can be seen as an vertex. Then we compute the Euclidean distance between embeddings within the same frame. Then, we employ the k-nearest neighbor (k-NN) algorithm to identify the k nearest neighbors for each vertex. Based on these neighbors, we construct an intra-patch graph $G_i^j = (X_t^1, E)$, where E denotes the corresponding set of edges, and each vertex $x_{ti}^j \in P$ only connects to its k-nearest neighbors. Then, we use a Graph Convolution Network [10] on the intra-patch graph G_i^j to aggregate geometric features of the lesion regions. It aggregates small local features of the lesion into global distribution features. For three scale space, we can get:

$$G_t^{1'} = GCN(G_t^1), \quad G_t^{2'} = GCN(G_t^2), \quad G_t^{3'} = GCN(G_t^3) \tag{6}$$

where G_i^1, G_i^2, G_i^3 represent the graphs constructed at different scales. $G_i^{1'}$, $G_i^{2'}$, $G_i^{3'}$ denote the outputs of graph convolution.

Finally, we aggregate the feature maps of different scales to obtain the final multi-scale fused feature map F_{fused}. This is achieved by performing a weighted summation of the normalized feature maps $G_i^{1'}$, $G_i^{2'}$, and $G_i^{3'}$, where the weights are learned during training.

$$F_{fused} = w_1 \cdot G_i^{1'} + w_2 \cdot G_i^{2'} + w_3 \cdot G_i^{3'} \tag{7}$$

where w_1, w_2, and w_3 are the learnable weights for each scale.

2.3 Loss Function

For the fused feature map F_{fused}, we generate the lesion segmentation map F_{out} through a conv operation with a kernel size of $20 \times 1 \times 1$. Then, we employ combining Cross-entropy loss and Dice loss to optimize our model.

$$L = \alpha \cdot L_{CE} + (1 - \alpha) \cdot L_{Dice} \tag{8}$$

where α is a weighting parameter that balances the influence of the two losses, α is 0.6.

3 Experiments

3.1 Dataset and Implementation Details

Dataset. Our datasets consist of Cine-MRI images from short-axis views of the left ventricle, collected from two hospitals in China and referred to as SYC-QC

and SYC-SX. These datasets include 320 and 200 myocarditis cases, respectively, totaling of 10,400 Cine-MRI images. Due to variations in MRI scanners, the number of frames in cine-sequences varies between 20 and 25. To standardize the data, all sequences were uniformly downsampled to 20 frames at equal intervals. This ensures consistency across different scanners and patients, facilitating feature extraction and model training.

Evaluation Metrics. We employ the Dice score(Dice)[%], Precision(Pre)[%], Recall(Rec)[%] and Hausdorff Distance(HD)[mm] as metrics.

Implementation Details. Our model MG-Mamba is implemented using PyTorch 2.2.1 with CUDA 12.0. All experiments were run on an NVIDIA GeForce RTX 4090 GPU. During training, we used a batch size of 8 per GPU for 200 epochs. Optimization is performed using Adam with a Cosine Annealing with Warm Restarts scheduler (initial learning rate of 1e-3).

Table 1. Comparison with other methods on the SYC-QC and SYC-SX. The best-performing results are highlighted in bold. Natural denotes natural image segmentation methods. Medical denotes medical image segmentation methods

	Methods	SYC-QC				SYC-SX			
		Dice ↑	Pre ↑	Rec ↑	HD ↓	Dice ↑	Pre ↑	Rec ↑	HD↓
Natural	OW-VISF(IJCV'24) [11]	74.74	68.62	84.39	5.16	73.56	65.10	86.96	5.85
	Video-kMaX(WACV'24) [12]	73.81	65.94	87.03	5.49	72.15	65.79	82.61	5.49
	TarViS(CVPR'23) [13]	77.28	73.98	82.77	4.66	73.90	66.35	86.39	5.78
	TDSNet(INS'24) [14]	71.01	59.15	**91.97**	7.02	70.32	61.28	86.52	7.83
	CFFM++(TPAMI'24) [15]	78.92	75.42	84.90	5.70	75.47	68.90	86.31	5.01
Medical	CAS-Net(MedIA'23) [16]	78.48	71.64	88.51	6.06	76.62	67.94	**90.23**	5.34
	PolypNext(IJCARS'23) [17]	75.16	70.81	82.87	4.94	74.96	67.03	88.03	6.56
	MemSAM(CVPR'24) [18]	80.74	75.93	87.84	5.64	80.26	75.37	87.62	5.84
	ZePT(CVPR'24) [19]	81.50	79.28	85.77	3.99	80.09	76.31	86.02	4.01
	Vivim(arXiv'24) [20]	83.68	80.97	88.21	4.68	82.41	77.73	89.32	5.01
	Ours	**86.92**	**88.51**	86.54	**3.84**	**86.35**	**87.56**	86.45	**3.99**

3.2 Comparison with State-of-the-Art Methods

We validated our model on the SYC-QC and SYC-SX datasets, comparing it with ten state-of-the-art methods, including five natural image segmentation methods and five medical image segmentation methods. We compared our model with OW-VISF [11], Video-kMaX [12], TarViS [13], TDSNet [14] and CFFM++ [15] for natural image segmentation, and with CAS-Net [16], PolypNext [17], MemSAM [18], ZePT [19] and Vivim [20] for medical image segmentation. As shown in Table 1, our method outperforms others on both datasets, achieving a Dice of 86.92% and HD of 3.84 mm on SYC-QC, and a Dice of 86.35% and HD of 3.99 mm on SYC-SX. For visualization, we selected three top-performing methods from each category for comparison, including CFFM++, TarViS, OW-VISF, MemSAM, ZePT

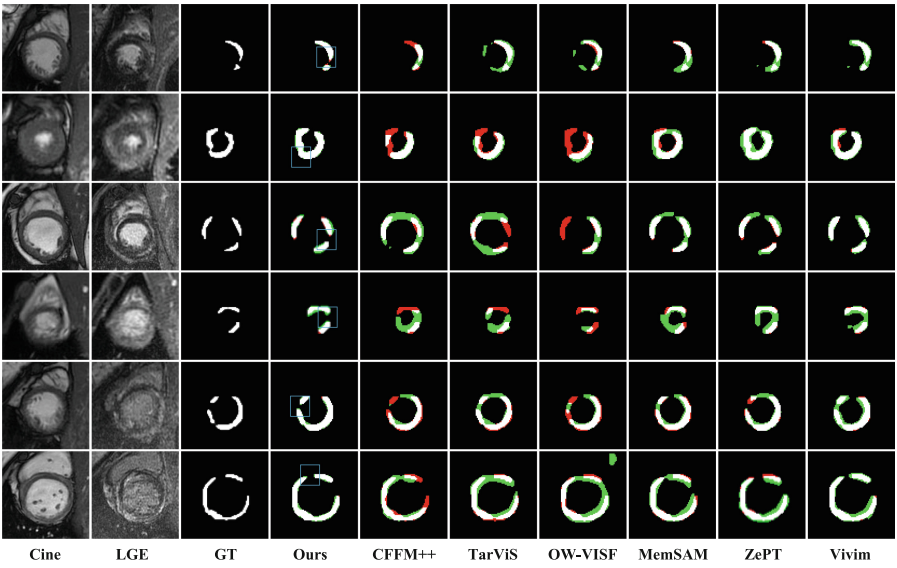


Fig. 3. Qualitative comparison of six different methods on SYC-QC. The green area in the figure indicates false positives, and the red area indicates false negatives. The blue boxed area indicates that we segmented the lesion details better (Color figure online)

and Vivim. As illustrated in Fig. 3, our model demonstrates superior segmentation performance, particularly in capturing fine details of myocarditis lesions.

3.3 Ablation Studies and Analysis

Effectiveness of Different Components. We conducted ablation studies on the SYC-QC dataset to validate the contributions of the MCMMA and MGGA modules in MG-Mamba. Three model versions were compared: M1 (backbone), M2 (backbone + MCMMA), and M3 (backbone + MGGA). Results in Table 2 show that both M2 and M3 outperform M1, while the MG-Mamba model (MCMMA + MGGA) achieves the best performance across all metrics. This demonstrates the complementary and synergistic effects of MCMMA and MGGA in enhancing lesion segmentation accuracy and robustness.

Feature Visualization Analysis. To further validate the effectiveness of our proposed MG-Mamba model, we perform feature visualization analysis on this model and several comparison models with better performance. In the experiment, we use the t-SNE technique to reduce the dimensions of normal and lesion patches in the features for visualization. The experimental results are shown in Fig. 4(a). We can see that our MG-Mamba model exhibits clearer class separation and tighter intra-class clustering compared to other models, demonstrating its superior discriminative capability.

Table 2. Ablation Studies of different components of our method

Methods	Modules		SYC-QC			
	MCMMA	MGGA	Dice ↑	Pre ↑	Rec ↑	HD ↓
M1	-	-	82.64	81.58	85.23	4.48
M2	✓	-	$83.61_{\uparrow 0.97}$	$82.56_{\uparrow 0.98}$	$86.28_{\uparrow 1.05}$	$4.03_{\downarrow 0.45}$
M3	-	✓	$84.49_{\uparrow 1.85}$	$84.90_{\uparrow 3.32}$	$85.61_{\uparrow 0.38}$	$3.92_{\downarrow 0.56}$
Ours	✓	✓	$\mathbf{86.92}_{\uparrow 4.28}$	$\mathbf{88.51}_{\uparrow 6.93}$	$\mathbf{86.54}_{\uparrow 1.31}$	$\mathbf{3.84}_{\downarrow 0.64}$

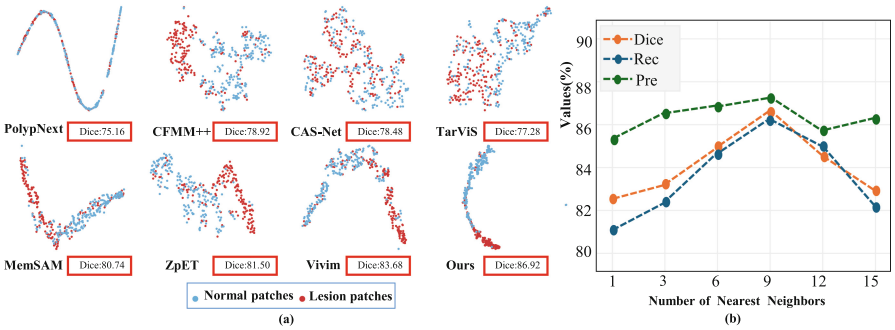


Fig. 4. (a) shows the t-SNE results of our MG-Mamba model and other comparative models. (b) shows the segmentation results of MG-Mamba with respect to different numbers k of nearest neighbors in k-NN graph

Stability Analysis. In the MG-Mamba model, the neighborhood size K of the k-NN algorithm influences the learning of geometric features of myocarditis lesions. To evaluate the effect of different values of K on the model performance, we trained the model on the SYC-QC dataset with K values set to {1, 3, 6, 9, 12, 15}. As shown in Fig. 4(b), the model's performance initially improves and then declines as K increases, peaking at K = 9. This value provides an optimal local space for capturing detailed geometric information.

4 Conclusion

In our work, we propose the MG-Mamba framework for myocarditis lesion segmentation in Cine-MRI images. The framework integrates two key modules: MCMMA for multi-step cross-frame motion analysis to address spatiotemporal inconsistencies, and MGGA for multi-scale graph geometric aggregation to capture both local structural patterns and global lesion distributions. Experimental results demonstrate significant performance gains over state-of-the-art methods, highlighting the effectiveness of our approach.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (No.62376002, No.62101491, No.62401016), in part by the Foundation of Sichuan Provincial People's Hospital (Grant Number 2023QN25).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Lasica, R., et al.: Update on myocarditis: from etiology and clinical picture to modern diagnostics and methods of treatment. Diagnostics **13**(19), 3073 (2023)
2. Seidman, M.A., McManus, B.: Myocarditis. Cardiovasc. Pathol. 553–575 (2022)
3. Brociek, E., et al.: Myocarditis: etiology, pathogenesis, and their implications in clinical practice. Biology **12**(6), 874 (2023)
4. Luetkens, J.A., et al.: Comparison of original and 2018 lake Louise criteria for diagnosis of acute myocarditis: results of a validation cohort. Radiol. Cardiothorac. Imaging **1**(3), e190010 (2019)
5. Coimbra, S., et al.: Toxicity mechanisms of gadolinium and gadolinium-based contrast agents–a review. Int. J. Mol. Sci. **25**(7), 4071 (2024)
6. Wendell, D.C., Judd, R.M.: Cardiac cine imaging. Basic Principles Cardiovasc. MRI: Phys. Imaging Tech. 145–159 (2015)
7. Isensee, F., Jaeger, P.F., Full, P.M., Wolf, I., Engelhardt, S., Maier-Hein, K.H.: Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In: Pop, M., et al. (eds.) STACOM 2017. LNCS, vol. 10663, pp. 120–129. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75541-0_13
8. Bière, L., Piriou, N., Ernande, L., Rouzet, F., Lairez, O.: Imaging of myocarditis and inflammatory cardiomyopathies. Arch. Cardiovasc. Dis. **112**(10), 630–641 (2019)
9. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
10. Zhang, S., Tong, H., Xu, J., Maciejewski, R.: Graph convolutional networks: a comprehensive review. Comput. Soc. Netw. **6**(1), 1–23 (2019). https://doi.org/10.1186/s40649-019-0069-y
11. Thawakar, O., et al.: Video instance segmentation in an open-world. Int. J. Comput. Vis. 1–12 (2024)
12. Shin, I., et al.: Video-kmax: a simple unified approach for online and near-online video panoptic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 229–239 (2024)
13. Athar, A., Hermans, A., Luiten, J., Ramanan, D., Leibe, B.; Tarvis: a unified approach for target-based video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18738–18748 (2023)
14. Yuan, H., Peng, J., Cai, Z.: Tdsnet: a temporal difference based network for video semantic segmentation. Inf. Sci. **686**, 121335 (2024)
15. Sun, G., Liu, Y., Ding, H., Wu, M., Van Gool, L.: Learning local and global temporal contexts for video semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. (2024)
16. Dong, C., Songhua, X., Dai, D., Zhang, Y., Zhang, C., Li, Z.: A novel multi-attention, multi-scale 3d deep network for coronary artery segmentation. Med. Image Anal. **85**, 102745 (2023)

17. Bhattacharya, D., Reuter, K., Behrendt, F., Maack, L., Grube, S., Schlaefer, A.: Polypnextlstm: a lightweight and fast polyp video segmentation network using convnext and convlstm. Int. J. Comput. Assist. Radiol. Surg. **19**(10), 2111–2119 (2024)
18. Deng, X., Wu, H., Zeng, R., Qin, J.; Memsam: taming segment anything model for echocardiography video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9622–9631 (2024)
19. Jiang, Y., Huang, Z., Zhang, R., Zhang, X., Zhang, S.; Zept: zero-shot pan-tumor segmentation via query-disentangling and self-prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11386–11397 (2024)
20. Yang, Y., Xing, Z., Zhu, L.; Vivim: a video vision mamba for medical video object segmentation. arXiv preprint arXiv:2401.14168 (2024)



OsteoOpt: A Bayesian Optimization Framework for Enhancing Bone Union Likelihood in Mandibular Reconstruction Surgery

Hamidreza Aftabi[1(✉)], John E. Lloyd[1], Amanda Ding[2], Benedikt Sagl[3], Eitan Prisman[2], Antony Hodgson[4], and Sidney Fels[1]

[1] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada
aftabi@student.ubc.ca
[2] Department of Surgery, University of British Columbia, Vancouver, Canada
[3] Center of Clinical Research, Medical University of Vienna, Vienna, Austria
[4] Department of Mechanical Engineering, University of British Columbia, Vancouver, Canada

Abstract. Mandibular reconstruction is crucial after oral tumor resection, yet current methods rely on premorbid geometric approximations and struggle with achieving reliable donor-native bone union. We propose a Bayesian optimization framework that enhances predicted bone union likelihood and facilitates computer-aided intervention by systematically varying key surgical parameters—resection plane orientation, donor bone positioning, and graft length—across three mandibular regions. Reconstruction performance is evaluated using two cost functions, coupled with a sensitivity analysis on modeling parameters. We validated the model using longitudinal patient-specific data from 5-day and 1-year postoperative CT and MRI scans. Our results show that optimization significantly enhances the predicted likelihood of bone union, with a relative improvement of up to 329% compared to the standard surgical practice. Additionally, validation shows a Dice coefficient of up to 0.76 between union prediction and actual postoperative imaging data. This study suggests that modifying the standard surgical plan can significantly improve bone union, underscoring the need for advanced optimization frameworks in surgical planning. The open-source code is available on GitHub.

Keywords: Mandibular Reconstruction Surgery · Optimization · Computer-Assisted Intervention · Computational Modeling

1 Introduction

The incidence of oral cancer has risen significantly in recent years. The American Cancer Society estimates that about 59,660 new cases of oral cavity and oropharyngeal cancer will be diagnosed in the U.S. in 2025 [1]. Many cases involve

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 448–458, 2026.
https://doi.org/10.1007/978-3-032-04947-6_43

the mandible, necessitating resection and reconstruction to restore the jaw's contour and function [2]. Autologous bone grafting, primarily using fibula or scapula grafts, is the gold standard for mandibular reconstruction [3]. However, nonunion at the graft-host interface remains a major challenge [4,5], with rates reaching up to 37% [4]. Such complications can cause pain, functional impairment, and, in severe cases, necessitate revision surgery.

Several studies have assessed the likelihood of bone union in mandibular reconstruction using simulations based on established bone remodeling algorithm [6–9]. While clinically evaluated in specific scenarios such as denture fixation [9] or mandibular reconstruction [6–8], these studies have some limitations. Existing methods lack automated design optimization for predicting bone union—typically focusing on implants or plates [10] while neglecting bone graft structure. In contrast to hardware modifications, which may lead to complications such as fracture or exposure [11], adjusting bone shape offers greater surgical flexibility. These studies also use simplified biomechanical models of the temporomandibular joint (TMJ) [12], muscle [3], and bone-to-bone contact, relying on static rather than dynamic simulations.

Building on prior works [13,14], (i) We develop a Bayesian optimization framework that enhances predicted bone union by efficiently exploring key design parameters. Integrating medical imaging and dynamic simulation, our method supports next-generation computer-aided intervention for personalized surgical planning that optimizes functional parameters beyond conventional geometry replication [15]. The open-source code is available on GitHub. (ii) We optimize three key but less explored parameters—resection plane orientation, donor vertical offset, and donor bone length—evaluating them across three mandibular defect types with two optimization cost functions. Robustness is assessed through sensitivity analysis. We hypothesize that our automated optimization can significantly improve bone union likelihood. (iii) Finally, we validate the model by comparing its predictions with longitudinal patient-specific imaging data.

2 Methods and Materials

2.1 Workflow

Our workflow (Fig. 1) comprises three automated steps—Virtual Planning, Mesh Refinement, and Physics-based Simulation—integrated into the optimization process. The core implementation is in MATLAB, integrating external tools.

Virtual Planning: We used the ArtiSynth API [16] to reconstruct three defect cases based on Urken's Classification [17]. Using imaging data from [13], we generated three synthetic defects: body (B), symphysis (S), and ramus-body (RB). Surgeons define the resection plane positions based on defect type, while other angular and length-based variables are optimized through the program described in Sect. 2.2. The number and position of donor segments are determined using the Ramer-Douglas-Peucker (RDP) algorithm [18], which approximates the mandible's contour using straight-line segments. One segment is used

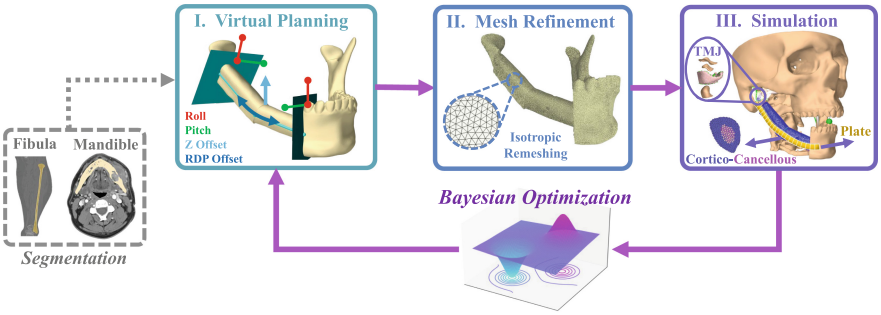


Fig. 1. The Bayesian optimization workflow includes virtual planning, mesh refinement, and simulation (RB baseline example with optimization parameters).

to reconstruct B and S defects, while RB defects require two segments; however, the framework generalizes to any number of segments. RDP point offset controls the position of the intermediate cut between the two donor segments along the mandible's curvature. To increase bony contact, a brute-force search is performed, analyzing overlapping planar polygons from resection and donor meshes to maximize apposition. Finally, a coarse finite element (FE) plate deforms to match the mandibular shape via axial springs in a forward simulation.

Mesh Refinement: We used PYMeshLab API [19] to refine the mesh before physics-based simulations. A mesh sensitivity analysis on maximum principal stress (MPS) determined the optimal edge length, stopping when changes fell below 5% to balance computational cost and accuracy. The process applies isotropic explicit remeshing to key anatomical structures with a target edge length of 0.50 mm, 50 refinement iterations. Additional steps merge close vertices, snap mismatched borders, remove duplicate faces, and repair non-manifold elements. Small holes are automatically closed using edge-based refinement to maintain structural integrity. The refined meshes are then used in physics-based simulations at each optimization iteration.

Physics-Based Simulation: The reconstructed and refined meshes are reimported into ArtiSynth for physics-based biomechanical simulation. This study builds on the model from [13], which includes rigid bodies (resected mandible, maxilla, hyoid), 24 Hill-type muscles, ligaments, a food bolus with 110 N resistance, and a validated TMJ with FE capsule and rigid disc [12]. To model different defects, the corresponding muscles attached to each resection area were removed, as detailed in [13], which also provides information on muscle and ligament properties. We further extended the model by incorporating FE plate and a donor bone with cortical and cancellous layers. We used TetGen library [20] to generate a volumetric mesh from the refined surface mesh in the previous stage. We observed uniform cortical thickness and linearly mapped average Hounsfield unit (HU) to bone density (0.7–1.8 g/cm^3). Linear elastic material properties were set as $E = 13.7$ GPa, $\nu = 0.3$ for cortical bone and $E = 1.1$ GPa, $\nu = 0.3$

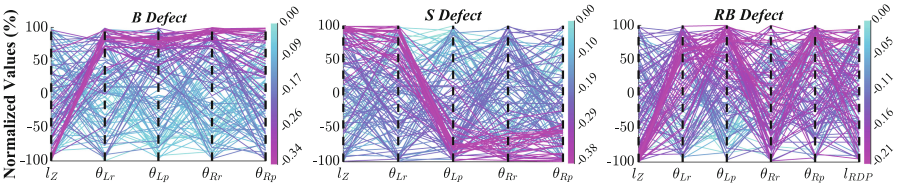


Fig. 2. Parallel chart of 150 iterations for three defects with F_{opt} objective. Parameters are normalized to $[-100, 100]$. Magenta marks the optimal region. (Color figure online)

for cancellous bone [9]. To secure the donor segment in place, we incorporated a linear elastic titanium plate ($\rho = 4.42$ g/cm^3, $E = 100$ GPa, $\nu = 0.3$) with rigid locking screws at segment midpoints, connecting the nearest nodes of the plate and donor bone. The plate was attached to the resected mandible by directly linking its nodes to the mandible's surface. Our simulation models the contact between the mandible and donor bone using vertex penetration and the Aharonov–Jones–Landau (AJL) method [16]. To improve accuracy, we apply a linear elastic contact [13,21], with $E = 30$ KPa, $v = 0.3$, and a thickness of $t_{contact} = 0.2$ mm [10]. These parameters account for near-void soft tissue regions forming in early reconstruction and the typical gap at the interface [10]. To predict bone union, we used strain energy density (SED) normalized by apparent bone density, the primary mechanical stimulus for bone remodeling. Based on established remodeling algorithm [6–10], when the normalized SED, S, exceeds $S_0(1+\S)$ with a lazy zone of $2\S = 0.2$ and $S_0 = 0.036$ mJ/g, bone growth occurs, enhancing bone apposition. At each time step, the normalized SED of each element is computed as $S = (1/2)(\sigma : \varepsilon)/\rho$, where σ and ε are the element stress and strain tensors, respectively, and ρ is the element apparent bone density. To assess remodeling, we overlaid a single element layer on each donor resection side and defined the apposition fraction as the ratio of elements exceeding the threshold to the total elements in the layer. We ran the ArtiSynth forward dynamics simulation of chewing with the non-reconstructed side and muscle activation set to pre-reconstruction levels to represent the early stage of reconstruction [13]. The simulation used a constrained backward Euler integrator (0.001 s step) on a PC (16 GB RAM, RTX 3060, i7-10700F@2.9 GHz), averaging 6 min per iteration.

2.2 Optimization Framework

Parameters Setup: Optimization parameters include the angular variables for roll and pitch angles of the right and left resection planes, with $\theta_{Rp} = \theta_{Lp} = \theta_{Rr} = \theta_{Lr} = 0$, and two length parameters: the vertical donor offset (l_Z) and, for the two segment case, the RDP point offset (l_{RDP}), both initially set to zero (see Fig. 1). These baseline values are commonly used by surgeons. The parameters are grouped as follows:

$$\boldsymbol{\theta} = (\theta_{Lr}, \theta_{Lp}, \theta_{Rr}, \theta_{Rp})^T \in \mathbb{R}^4 \quad \text{(angular)}, \quad \boldsymbol{l} = (l_Z, l_{RDP})^T \in \mathbb{R}^2 \quad \text{(length)}. \tag{1}$$

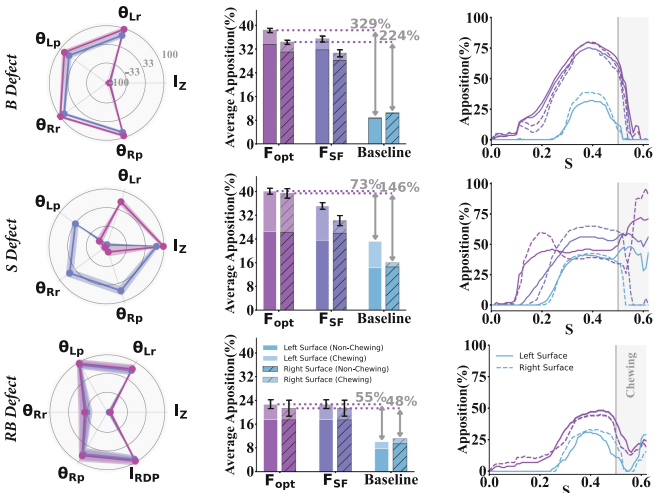


Fig. 3. Average of five trials showing mapped optimized parameters $[-100, 100]$ (col. 1), apposition average (col. 2), and apposition over time (col. 3) for three defects (rows) and objectives: F_{opt}, F_{SF}, and Baseline (parameters set to zero).

The full parameter vector is $\boldsymbol{\phi} = (\boldsymbol{\theta}^T, \boldsymbol{l}^T)^T \in \mathbb{R}^6$, with the feasible region:

$$\mathcal{X} = \left\{ \boldsymbol{\phi} \in \mathbb{R}^6 \mid \theta_{Lr},\, \theta_{Lp} \in [-\alpha, \alpha],\, \theta_{Rr},\, \theta_{Rp} \in [-\beta, \beta],\, l_Z \in [-z, z],\, l_{RDP} \in [-r, r] \right\}. \tag{2}$$

The parameter ranges for each defect type are chosen based on physical constraints, ensuring at most one tooth root is sacrificed and a 10–25 mm vertical gap between the fibula and maxillary teeth for implant placement [22]. The vector $(\alpha°, \beta°, z\,\text{mm}, r\,\text{mm})$ is set to $(25°, 20°, 3.5, -)$ for the B defect, $(15°, 15°, 5, -)$ for the S defect, and $(25°, 15°, 5, 7)$ for the RB defect. Jaw reconstruction effectiveness is assessed using apposition fractions over a chewing cycle with $n = 62$ data point (spanning 0.62 s). Here, $L_i(\boldsymbol{\phi})$ and $R_i(\boldsymbol{\phi})$ denote the left and right fractions at each time step i, respectively, while $M_i(\boldsymbol{\phi})$ is defined as the minimum fraction from the contacting middle resection planes as they show similar behaviour. Their cycle averages are given by $\overline{X}(\boldsymbol{\phi}) = \frac{1}{n}\sum_{i=1}^{n} X$ for $X \in \{L, R, M\}$. We use $\overline{X}(\boldsymbol{\phi})$ in our cost functions, since sustained contact is key to union [2,23]. Two cost functions are then defined for optimization.

Cost Functions: The first cost function, F_{opt}, determines the optimal parameters by solving $\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} -(F_{opt})$ where

$$F_{opt} = W_1 \sum\nolimits_{X \in \{L,R,M\}} \overline{X}(\boldsymbol{\phi}) - W_2 \sum\nolimits_{X \neq Y \in \{L,R,M\}} |\overline{X}(\boldsymbol{\phi}) - \overline{Y}(\boldsymbol{\phi})|. \tag{3}$$

In this equation, the first term rewards high overall apposition by maximizing the sum of the average left, right, and middle apposition fractions, while the second term penalizes asymmetry between them. The coefficients $W_1 = W_2 = 0.5$ balance these contributions.

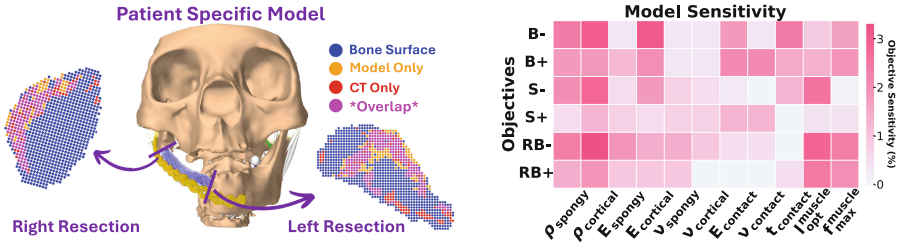


Fig. 4. Patient-specific validation against imaging data and sensitivity analysis. (Color figure online)

The second cost function, $F_{SF} = F_{opt} - \overline{C}$, builds upon the first by introducing a penalty term $\overline{C}$ for deviations in safety factor:

$$\overline{C} = \frac{1}{n} \sum_{i=1}^{n} \sum_{s \in \{\text{left,right}\}} w^s \left(\max(0, SF_{\text{desired}} - SF^s_{\text{worst},i})\right)^2. \tag{4}$$

The safety factor (SF) quantifies structural integrity, defined as the ratio of a bone's compressive yield strength to its maximum experienced stress, with $SF_{\text{desired}} = 1$. A higher SF indicates a lower failure risk. The worst-case safety factor for each side is computed as:

$$SF^s_{\text{worst},i} = \min \left\{ \frac{\sigma_{\text{yield, cortical}}}{\sigma^s_{\text{maxP, cortical},i}}, \frac{\sigma_{\text{yield, cancellous}}}{\sigma^s_{\text{maxP, cancellous},i}} \right\}, \quad s \in \{\text{left}, \text{right}\}. \tag{5}$$

Here, $\sigma^s_{\text{maxP, cortical/cancellous},i}$ represents MPS in each region on side s at time i during the chewing cycle. The yield stress values are set as $\sigma_{\text{yield, cortical}} = 100\,\text{MPa}$ and $\sigma_{\text{yield, cancellous}} = 5\,\text{MPa}$ [24].These parameters are selected to ensure optimization under worst-case scenarios, taking into account the region most prone to failure. The weighting factors $w^s = 0.5$ apply equal penalty contributions for both sides. These computationally expensive cost functions are then used within a Bayesian optimization.

Bayesian Optimization: To efficiently explore the parameter space $\boldsymbol{\phi} \in \mathbb{R}^6$ within the feasible region $\mathcal{X}$, we first generate $n_{\text{samples}} = 25$ points using a Sobol' sequence [25] (with a skip of 1000 and a leap of 100). Each sample is then linearly mapped to the feasible region by scaling the Sobol' values to the parameter limits. This initialization establishes quasi-uniform coverage of $\mathcal{X}$, reducing bias prior to Bayesian Optimization. The complete dataset at iteration N is defined by $\mathcal{D}_N = \{(\boldsymbol{\phi}_j, f_j)\}_{j=1}^{n_{\text{samples}}} \cup \{(\boldsymbol{\phi}_j, f_j)\}_{j=n_{\text{samples}}+1}^{N}$, where $f \in \{F_{opt}, F_{SF}\}$ and the first evaluations provide the initial sampling of $\mathcal{X}$. The subsequent evaluations are incorporated iteratively to refine the search. The computationally expensive function $f(\boldsymbol{\phi})$ is then modeled as a Gaussian Process (GP). Conditioned on the dataset $\mathcal{D}_N$, the posterior distribution of $f(\boldsymbol{\phi})$ is modeled as $f(\boldsymbol{\phi}) \mid \mathcal{D}_N \sim \mathcal{N}\left(\mu_N(\boldsymbol{\phi}), \sigma_N^2(\boldsymbol{\phi}) + \sigma^2_{\text{noise}}\right)$, where $\mu_N(\boldsymbol{\phi})$ denotes the

posterior mean, $\sigma_N^2(\boldsymbol{\phi})$ the model-derived posterior variance, and σ_{noise}^2 accounts for observation noise. To balance exploration and exploitation, the Expected Improvement Plus (EI+) acquisition function is used, prioritizing areas with the highest potential for improvement. The minimum observed function value among the dataset is defined as $\tilde{f}_{\min} = \min\{f_j : (\boldsymbol{\phi}_j, f_j) \in \mathcal{D}_N\}$. with EI+ computation then follows [26]:

$$\text{EI+}(\boldsymbol{\phi}) = \left(\tilde{f}_{\min} - \mu_N(\boldsymbol{\phi})\right)\Phi\left(\frac{\tilde{f}_{\min} - \mu_N(\boldsymbol{\phi})}{\sqrt{S(\boldsymbol{\phi})}}\right) + \sqrt{S(\boldsymbol{\phi})}\,\varphi\left(\frac{\tilde{f}_{\min} - \mu_N(\boldsymbol{\phi})}{\sqrt{S(\boldsymbol{\phi})}}\right). \tag{6}$$

where we define $S(\boldsymbol{\phi}) = \sigma_N^2(\boldsymbol{\phi}) + \sigma_{\text{noise}}^2$, and denote by $\Phi(\cdot)$ and $\varphi(\cdot)$ the cumulative distribution function and probability density function, respectively, of the standard normal distribution. The next sampling point is selected by solving $\boldsymbol{\phi}_{N+1} = \arg\max_{\boldsymbol{\phi}\in\mathcal{X}} \text{EI+}(\boldsymbol{\phi})$.

2.3 Sensitivity Analysis and Validation

We analyzed model sensitivity by applying ±10% changes to 11 biomechanical parameters, including bone properties, elastic contact, and muscle parameters, and evaluated their impact on F_{opt} across all defect types. Each case was repeated five times, yielding a total of 330 simulations, with the average effect reported.

Moreover, beyond analyzing synthetic defect cases (B, S, and RB), we performed patient-specific validation using CT scans at 5 days and 1 year post-surgery, along with post-operative MRI. The patient, a 65-year-old female, underwent right ramus and body reconstruction with a scapular bone. The 5-day CT was used to segment the mandible, maxilla, hyoid, and plate in Slicer, capturing the early reconstruction stage. T1-weighted MRI was registered to CT using Elastix [27] for TMJ disc segmentation. TotalSegmentator [28] was used to segment four key masticatory muscles, and their physiological cross-sectional areas were determined using the method described in [6]. Muscle insertions and origins were manually assigned based on these segmentations, and the maximum lengths were obtained by scaling the resting lengths by a factor of 1.2 to 1.5 [13]. Several muscles/ligaments, similar to those excluded in the RB defect (except temporalis), were omitted. FE components for the donor bone, fixation plate, and TMJ capsule as well as simulation method followed Sect. 2.1. We used the method in Sect. 2.1 to estimate a uniform cortical thickness of 0.78 mm. Cortical and cancellous bone densities were set to $\rho = 1.45$ g/cm^3 and $\rho = 0.7$ g/cm^3, respectively, with other parameters unchanged. We assessed bone union likelihood over a full chewing cycle by tracking apposition fraction at the donor-mandible interface (resection sides). To assess prediction accuracy, the 1-year CT was segmented, and a 1-voxel-thick tensor quantified bone formation at the mandible-donor interface. Cortical bone percentage (*cort*%) was computed using an HU threshold of 1000 [29], a known marker of bone formation and successful union [30], generating a binary label map. We then used this percentage in the simulation to rank and select the top *cort*% of elements on the resection

surface, prioritizing those that reached the apposition threshold and maintained contact for a longer duration over the chewing cycle [2,23], representing regions with prolonged bone-to-bone interaction. Finally, the centroids of these elements were interpolated using the Gaussian method to align with the imaging voxel grid, allowing direct correlation between simulation and clinical imaging data.

3 Results and Discussion

Figure 3 illustrates the optimized parameters and the average apposition behavior over one chewing cycle for the three defect cases (B, S, and RB). The results show that peak apposition occurs during the mid-phase (lateral movement) and the end-phase (bolus chewing, gray background). The results, averaged over five simulations with random initial values and 50 iterations, demonstrate the robustness and efficiency of the optimization process, achieving optimal values with minimal observations. The optimization resulted in substantial relative improvements in predicted bone union—up to 329% for B, 146% for S, and 55% for RB—compared to the surgeon's common practice (baseline), while keeping the standard deviation low. Changing the cost function from F_{opt} to F_{SF} had minimal impact on B, no impact on RB, but altered the optimized values for S, which exhibited higher apposition during the bolus chewing phase, suggesting a more conservative optimization approach may be necessary to reduce failure risk. Moreover, the lower improvement in RB may result from extensive resection and significant muscle loss, reducing load at the mandible-donor interface. Additionally, to demonstrate the impact of each parameter on the cost function F_{opt}, we ran 150 iterations. Figure 2 presents a parallel plot of the normalized variables and their effect on optimization. As this is a minimization problem, lower values indicate more favorable outcomes, shown in magenta. The results highlight the significant contribution of key surgical parameters, particularly for B and S defects, while for RB, the right resection parameter had less influence due to the lack of muscle and ligament support. While the Bayesian optimization was run sequentially, its parallelization capability could further reduce runtime for larger iterations. The sensitivity analysis further demonstrates the robustness of our modeling, as shown in Fig. 4. Varying the modeling parameters by $\pm 10\%$ resulted in less than a 4% change in the objective function.

Although the model we built upon had previously been validated against the literature for jaw movement, bite force, muscle activation, condyle motion, and bone remodeling [6–9,13], we further validated it using longitudinal patient-specific data. Because assigning different resection or donor positions within a single patient is not feasible, we assessed simulation accuracy by comparing the predicted bone likelihood with patient-specific observations. Figure 4 illustrates the patient-specific model, with the right and left resection surfaces in blue. The simulation's predicted bone formation (orange), based on Sect. 2.3, was compared to cortical bone formation observed in imaging (red). The overlap between prediction and imaging data (pink) yielded a Dice coefficient of 0.7 for the left surface and 0.76 for the right surface, indicating a strong correlation.

However, some discrepancies may stem from missing patient-specific data, such as jaw movement, and modeling simplifications, particularly in screw modeling near the resections.

4 Conclusion and Future Directions

This study introduces an automated Bayesian optimization framework for key surgical variables, achieving up to a 329% improvement in predicted bone union. While model predictions have been validated against patient-specific clinical data, larger datasets are needed for broader confirmation. To date, the approach has been tested in a single patient because obtaining longitudinal imaging at 5 days and 1 year post-surgery is challenging and not part of routine care; additional CT scans also raise radiation-exposure concerns, particularly for patients receiving adjuvant therapy.

Future work will refine the model by incorporating volumetric muscle representations and patient-specific data on mandibular kinematics, muscle activation, and bite force. We will also evaluate additional variables, such as screw positioning, and conduct usability testing with surgeons. Although outcome variability has not yet been fully studied across different patients, we hypothesize that individuals with similar defect types—classified according to Urken—will exhibit comparable responses; this will be investigated in future studies.

Acknowledgments. We gratefully acknowledge funding from the Terry Fox Research Institute and extend our sincere thanks to the ISTAR Group at the University of British Columbia for their invaluable support.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. American Cancer Society. Oral cavity and oropharyngeal cancer: Key statistics (2025). https://www.cancer.org/cancer/types/oral-cavity-and-oropharyngeal-cancer/about/key-statistics.html. Accessed 18 Feb 2025
2. Kumar, B.P., Venkatesh, V., Jeevan Kumar, K.A., Yashwanth Yadav, B., Ram Mohan, S.: Mandibular reconstruction: overview. J. Maxillofac. Oral Surg. **15**, 425–441 (2016)
3. Aftabi, H., et al.: Computational models and their applications in biomechanical analysis of mandibular reconstruction surgery. Comput. Biol. Med. **169**, 107887 (2024)
4. Swendseid, B., et al.: Natural history and consequences of nonunion in mandibular and maxillary free flaps. Otolaryngol.-Head Neck Surg. **163**(5), 956–962 (2020)
5. Sabiq, F., et al.: Evaluating the benefit of virtual surgical planning on bony union rates in head and neck reconstructive surgery. Head Neck **46**(6), 1322–1330 (2024)
6. Zheng, K., et al.: Investigation on masticatory muscular functionality following oral reconstruction-an inverse identification approach. J. Biomech. **90**, 1–8 (2019)
7. Zheng, K., et al.: Bone remodeling following mandibular reconstruction using fibula free flap. J. Biomech. **133**, 110968 (2022)

8. Wan, B., et al.: On interaction between fatigue of reconstruction plate and time-dependent bone remodeling. J. Mech. Behav. Biomed. Mater. **136**, 105483 (2022)
9. Field, C., Li, Q., Li, W., Thompson, M., Swain, M.: Prediction of mandibular bone remodelling induced by fixed partial dentures. J. Biomech. **43**(9), 1771–1779 (2010)
10. Ferguson, B.M., Entezari, A., Fang, J., Li, Q.: Optimal placement of fixation system for scaffold-based mandibular reconstruction. J. Mech. Behav. Biomed. Mater. **126**, 104855 (2022)
11. Onoda, S., et al.: Prevention points for plate exposure in the mandibular reconstruction. J. Cranio-Maxillofac. Surg. **40**(8), e310–e314 (2012)
12. Sagl, B., Schmid-Schwap, M., Piehslinger, E., Kundi, M., Stavness, I.: A dynamic jaw model with a finite-element temporomandibular joint. Front. Physiol. **10**, 1156 (2019)
13. Aftabi, H., Sagl, B., Lloyd, J.E., Prisman, E., Hodgson, A., Fels, S.: To what extent can mastication functionality be restored following mandibular reconstruction surgery? A computer modeling approach. Comput. Methods Programs Biomed. **250**, 108174 (2024)
14. Aftabi, H., Lloyd, J.E., Sagl, B., Ding, A., Prisman, E., Hodgson, A., Fels, S.: Optimizing bone cuts enhances predicted bone union propensity in mandibular body reconstruction. In: 2025 IEEE 22th International Symposium on Biomedical Imaging (ISBI). IEEE (2025)
15. Vyas, K., Gibreel, W., Mardini, S.: Virtual surgical planning (VSP) in craniomaxillofacial reconstruction. Facial Plast. Surg. Clin. **30**(2), 239–253 (2022)
16. Lloyd, J.E., Stavness, I., Fels, S.: ArtiSynth: a fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In: Soft Tissue Biomechanical Modeling for Computer Assisted Surgery, pp. 355–394 (2012)
17. Urken, M.L., Weinberg, H., Vickery, C., Buchbinder, D., Lawson, W., Biller, H.F.: Oromandibular reconstruction using microvascular composite free flaps: report of 71 cases and a new classification scheme for bony, soft-tissue, and neurologic defects. Arch. Otolaryngol.-Head Neck Surg. **117**(7), 733–744 (1991)
18. Shkedy, Y., Howlett, J., Wang, E., Ongko, J., Durham, J.S., Prisman, E.: Predicting the number of fibular segments to reconstruct mandibular defects. Laryngoscope **130**(11), E619–E624 (2020)
19. Muntoni, A., Cignoni, P.: PyMeshLab (2021)
20. Hang, S.: TetGen, a delaunay-based quality tetrahedral mesh generator. ACM Trans. Math. Softw. **41**(2), 11 (2015)
21. Bei, Y., Fregly, B.J.: Multibody dynamic simulation of knee contact mechanics. Med. Eng. Phys. **26**(9), 777–789 (2004)
22. Tran, K.L., et al.: Dental implantability of mandibular reconstructions: comparing freehand surgery with virtual surgical planning. Oral Oncol. **140**, 106396 (2023)
23. Prasad, J., Goyal, A.: An invertible mathematical model of cortical bone's adaptation to mechanical loading. Sci. Rep. **9**(1), 5890 (2019)
24. Morgan, E.F., Unnikrisnan, G.U., Hussein, A.I.: Bone mechanical properties in healthy and diseased states. Annu. Rev. Biomed. Eng. **20**(1), 119–143 (2018)
25. Renardy, M., Joslyn, L.R., Millar, J.A., Kirschner, D.E.: To sobol or not to sobol? The effects of sampling schemes in systems biology applications. Math. Biosci. **337**, 108593 (2021)
26. Qin, C., Klabjan, D., Russo, D.: Improving the expected improvement algorithm. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
27. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2009)

28. Wasserthal, J., et al.: TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol.: Artif. Intell. **5**(5) (2023)
29. Mahesh, M.: The essential physics of medical imaging. Med. Phys. **40**(7), 077301 (2013)
30. Knitschke, M., et al.: Osseous union after mandible reconstruction with fibula free flap using manually bent plates vs. patient-specific implants: a retrospective analysis of 89 patients. Curr. Oncol. **29**(5), 3375–3392 (2022)

PD-INR: Prior-Driven Implicit Neural Representations for TOF-PET Reconstruction

Yuxuan Long[1], Yulin Zhang[1], Hong Wang[3], Xiaodong Kuang[2], Hailiang Huang[2], Fan Rao[2], Huafeng Liu[1], Yefeng Zheng[4], and Wentao Zhu[1,2(✉)]

[1] The College of Optical Science and Engineering, Zhejiang University, Hangzhou, China
zhuwentao.ee@gmail.com
[2] Zhejiang Lab, Hangzhou, China
[3] Xi'an Jiaotong University, Xi'an, China
[4] Westlake University, Hangzhou, China

Abstract. Positron emission tomography (PET) reconstruction is a challenging inverse problem, where projection data often contain low statistics. While current supervised learning methods offer strong noise suppression abilities, they may suffer from generalization issues and are in many cases not accurate quantitatively. To overcome these challenges, we propose a novel self-supervised Time-of-Flight PET (TOF-PET) reconstruction framework that utilizes Implicit Neural Representations (INR) to model PET images. Specifically, we introduce a differentiable forward projection model based on the imaging mechanism for TOF-PET and reformulate TOF-PET reconstruction problem using INR. To enhance image smoothness, we develop a ray-based total variation (TV) regularization term, distinct from the traditional TV. For the internal structure of our INR, we integrate a multi-resolution hash encoder with our designed prior-image encoder, where the latter provides sufficient image prior and always delivers reliable initial reconstructions for arbitrary network depth. Experiments on brain and chest datasets show that our method outperforms traditional iterative algorithms and self-supervised approaches in noise suppression and contrast recovery. Compared to conventional NeRF-based architectures, our model is more compact and converges faster, providing an efficient solution for TOF-PET reconstruction. The source code repository is hosted on GitHub: https://github.com/zyl123300/PD-INR.git.

Keywords: TOF-PET Reconstruction · Inverse Problem · Implicit Neural Representation · Neural Radiance Fields

Y. Long and Y. Zhang—Contributed equally (co-first authors).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 459–469, 2026.
https://doi.org/10.1007/978-3-032-04947-6_44

1 Introduction

Positron emission tomography (PET) is a widely used functional imaging technique for clinical applications [1,9,17], such as tumor detection, by tracking the spatial distribution of radioactive tracers. However, PET systems face challenges like low spatial resolution and high noise sensitivity. Time-of-Flight (TOF) PET addresses these limitations by incorporating the time differences of annihilation photons, enhancing spatial resolution, especially in low-dose scenarios.

Although the TOF kernel's finite width (typically modeled as a Gaussian) introduces some uncertainty in localization, TOF-PET significantly improves imaging performance. However, effective reconstruction algorithms remain crucial for further enhancing PET image quality.

To address the limitations of PET imaging, iterative reconstruction techniques were introduced. One of the most prominent iterative methods is the Maximum Likelihood Expectation Maximization (MLEM) algorithm [11,24], which iteratively optimizes the solution by incorporating a non-negativity constraint to prevent negative values in the reconstructed image. Compared to analytical methods, MLEM offers better noise suppression, significantly improving image quality in low-signal regions. As a result, iterative reconstruction has become the standard approach in most modern PET systems.

In recent years, deep learning has achieved significant advancements in PET image reconstruction, particularly in enhancing image quality and accelerating the reconstruction process. Deep learning applications in PET reconstruction can be broadly categorized into three main approaches: (1) image post-processing [7,14], where traditional iterative algorithms are used to generate a preliminary image that may contain noise or artifacts, and a denoising network is then trained to refine the image; (2) end-to-end deep learning models [10,29] directly map sinogram data to image space, enabling the networks to learn key features from large datasets and reduce reliance on manual intervention. However, these approaches typically require a large amount of paired data and suffer from significant generalization issues when handling new data. (3) hybrid methods that integrate traditional iterative algorithms with deep learning [5,6], replacing specific steps such as data fitting and regularization with deep neural networks. This combination retains the stability and theoretical rigor of traditional methods while benefiting from the nonlinear mapping capabilities of deep learning, but still faces issues such as the need for pre-training and generalization problems. In light of these challenges, a self-supervised reconstruction method needs to be proposed, which can optimize the image reconstruction process by leveraging the inherent properties of the data itself, without relying on large amounts of labeled data.

Recently, Implicit Neural Representation (INR) [18] has gained significant attention, particularly through its success in Neural Radiance Fields (NeRF). NeRF uses multilayer perceptrons (MLPs) to model the mapping from spatial coordinates to color and density in continuous volumetric fields, enabling high-quality image reconstruction from sparse or noisy data. This framework has been extended to medical imaging [3,4,15,16,19,22,23,27], with applications in

reconstruction, segmentation, and registration. For example, IREM [19] leverages implicit functions of 3D spatial coordinates to reconstruct high-resolution images from sparse inputs, while D-NAF [16] employs continuous 4D implicit neural representation for dynamic 4D-CBCT reconstruction, achieving superior image quality and efficiency. Other methods, such as CuNeRF [3], enable zero-shot super-resolution through cubic sampling and hierarchical rendering, further advancing medical image reconstruction. However, the application of INR for TOF-PET reconstruction remains unexplored so far.

In this work, our main contributions are summarized as follows:

- We propose a novel self-supervised method for TOF-PET reconstruction. To the best of our knowledge, it is the first exploration of INR's potential in TOF-PET.
- We develop a novel framework called prior-driven INR (PD-INR), which incorporates image prior into the conventional INR. With only a single-layer MLP, PD-INR could still achieve promising reconstruction quality and convergence.
- We introduce a novel ray-wise TV regularization term in the feature domain, specially customized for our PD-INR.

2 Method

2.1 TOF-PET Image Reconstruction

In TOF-PET, assuming that the TOF events have been already pre-corrected, the expected true count $\bar{y}_{ik}$ along line-of-response (LOR) i at TOF-bin k can be expressed as $\bar{y}_{ik} = \sum_j p_{ijk} x_j$, where x_j is the activity at voxel j, p_{ijk} is the sensitivity of the detector at (i, k) for activity in j. Assuming that TOF-PET data $\mathbf{y}$ are independently Poisson distributed, the log-likelihood function can be expressed as:

$$\mathcal{L}(\mathbf{y}|\mathbf{x}) = \sum\nolimits_{i,k} y_{ik} \log \bar{y}_{ik} - \bar{y}_{ik} - \log y_{ik}!. \tag{1}$$

The activity image is usually reconstructed by maximizing the Poisson log-likelihood:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x} \geq \mathbf{0}} \mathcal{L}(\mathbf{y}|\mathbf{x}). \tag{2}$$

2.2 Prior-Driven INR for TOF-PET Reconstruction

Mathematical Formulation. Conventional algorithms reconstruct the image in discretized image space. Here, we adopt INR to model our activity image. Besides the continuity of INR, the internal structure of INR would also bring some implicit prior into the image. For the INR, the activity image is expressed as:

$$x_j = f_\Theta(S_\Phi(\mathbf{c}_j)) \ \forall j, \tag{3}$$

where $\mathbf{c}_j \in \mathbb{R}^3$ is 3D coordinate of the voxel j, $S_\Phi(\cdot)$ encodes the spatial coordinates into image features, and $f_\Theta(\cdot)$ decodes the image features into the image intensities. Here, Φ and Θ are both learnable parameters for our image representation.

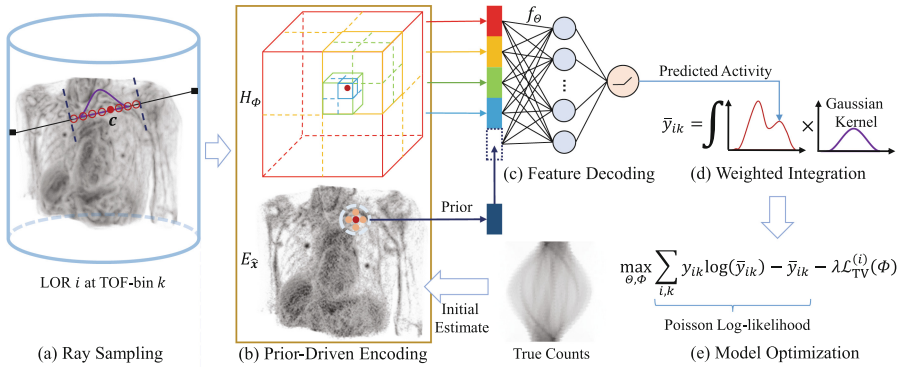


Fig. 1. Overview of the proposed PD-INR framework. (a) Along LOR i, numerous points are sampled within TOF-bin k. (b) For each point, the hash encoder H_Φ maps its coordinate $\mathbf{c}$ to multi-resolution features while $E_{\hat{\mathbf{x}}}$ extracts the neighboring intensities from the initial estimate $\hat{\mathbf{x}}$ as the prior feature. (c) All features are concatenated, then decoded by a tiny MLP f_Θ. (d) The predicted activities are integrated with the TOF profile, yielding the estimated counts $\bar{y}_{ik}$. (e) For all LORs and TOF-bins, the model parameters are optimized and regularized by our designed TV term in feature domain

To adapt to such image representation, the traditional system matrix may need to be re-defined in order to fully benefit from the expressive power of neural networks. Like NeRF, we would like to reformulate the forward projection in a ray-wise manner, in order to ease batch computation during iterative optimization.

For TOF-PET, the forward projection along the LOR i at TOF-bin k is modelled as a line integral, i.e. ,

$$\bar{y}_{ik} = \int_{-\infty}^{\infty} \Psi(\mathbf{x}, \mathbf{s}_i + t\mathbf{d}_i) h(t - t_{ik})\, dt, \tag{4}$$

where the LOR i has the starting point $\mathbf{s}_i \in \mathbb{R}^3$ and the unit direction $\mathbf{d}_i \in \mathbb{R}^3$, t_{ik} is the TOF variable, $h(\cdot)$ is the TOF profile often modeled as a truncated Gaussian function, and Ψ queries the image intensity at the specified spatial position. For conventional forward projection, Ψ linearly interpolates a discrete image. In the case of INR, we have $\Psi(\mathbf{x}, \mathbf{s}_i + t\mathbf{d}_i) = f_\Theta(S_\Phi(\mathbf{s}_i + t\mathbf{d}_i))$. By substituting it into (2) and further introducing a penalty term, the reconstruction

problem is finally re-formulated as:

$$\begin{aligned} \max_{\Theta,\Phi} \quad & \mathcal{L}(\mathbf{y}|\Theta,\Phi) - \lambda \sum_i \mathcal{L}_{\mathrm{TV}}^{(i)}(\Phi), \\ \text{s.t.} \quad & \bar{y}_{ik} = \int_{-\infty}^{\infty} f_\Theta(S_\Phi(\mathbf{s}_i + t\mathbf{d}_i))h(t - t_{ik})\, dt, \ \forall i, k, \end{aligned} \tag{5}$$

where $\mathcal{L}_{\mathrm{TV}}^{(i)}(\Phi)$ is a TV term that enhances the image smoothness, and λ is some small non-negative weight that controls the magnitude of penalty. Here, $\mathcal{L}_{\mathrm{TV}}^{(i)}(\Phi)$ is defined as:

$$\mathcal{L}_{\mathrm{TV}}^{(i)}(\Phi) = \frac{1}{(P-1)F} \sum\nolimits_{p=1}^{P-1} \left\| S_\Phi(\mathbf{v}_{p+1}^{(i)}) - S_\Phi(\mathbf{v}_p^{(i)}) \right\|_1, \tag{6}$$

where $\{\mathbf{v}_p^{(i)}\}_{p=1}^P$ are the coordinates for the sampled points along LOR i, with $\mathbf{v}_p^{(i)} = \mathbf{s}_i + t_p\mathbf{d}_i$, and the distances $\{t_p\}_{p=1}^P$ are of ascending order. F is the feature dimension, i.e. , $S_\Phi(\mathbf{v}_p^{(i)}) \in \mathbb{R}^F$. Those sampled points are also used for numerical approximation of the integral in (5), hence avoiding re-computations. Conventional TV penalty is computed in discrete image space, and may be time-consuming for INR as all voxels' activities are to be evaluated. Our TV term (6) is ray-wise (or LOR-wise), implying small computation load. It only penalizes on the difference of neighbouring features along each LOR. This is nearly equivalent to a ray-wise TV penalty in image domain since all coordinates share the same feature decoder when evaluating the activities.

Prior-Driven Encoding. TOF-PET reconstruction problem has been formulated using INR, as shown in (5). Here, we would like to discuss our prior-driven INR (PD-INR) framework in details, especially the prior-driven encoding function S_Φ. Specifically, S_Φ concatenates the features computed independently from two encoders, i.e. , $S_\Phi(\mathbf{c}) = \left[H_\Phi(\mathbf{c})^T, E_{\hat{\mathbf{x}}}(\mathbf{c})^T\right]^T$, where $H_\Phi(\cdot)$ is the learnable hash encoder [20] and $E_{\hat{\mathbf{x}}}(\cdot)$ is the prior-image encoder. These two encoders are respectively expressed as:

$$H_\Phi(\mathbf{c}) = \oplus_{l=1}^L \Psi\left(H^{(l)}, \mathbf{c}\right), \ E_{\hat{\mathbf{x}}}(\mathbf{c}) = \oplus_{\mathbf{c}' \in \mathcal{N}(\mathbf{c})} \Psi\left(\hat{\mathbf{x}}, \mathbf{c}'\right), \tag{7}$$

where $\oplus$ denotes feature concatenation and Ψ is the linear interpolation function that retrieves the features given the 3D coordinate.

For the hash encoder, there are a total of L levels of features for concatenation, and $H^{(l)}$ is the l^{th} level cube-like feature map stored in a trainable hash table. The introduction of hash tables not only compresses the number of parameters but also induces the many-to-one property due to hash collisions, which coincides with the sparsity nature of human organs, as claimed in NAF [28]. The hash encoder also concatenates multi-resolution features from each level, which facilitates learning of both low- and high-frequency details, thereby enhancing both image contrast and smoothness.

For the prior-image encoder, as shown in (7), $\mathcal{N}(\mathbf{c})$ denotes the set of neighbors (including itself) of the coordinate $\mathbf{c}$, and $\hat{\mathbf{x}}$ is the prior image obtained from the projection data. $E_{\hat{\mathbf{x}}}(\mathbf{c})$ extracts the image intensities from the neighbours and merges them as one feature vector, hence providing the image prior to our PD-INR. It is expected to facilitate convergence and strengthen the expressive power of our spatial encoding.

We design the feature decoder $f_\Theta(\cdot)$ as a tiny MLP, which is sufficient to handle the reconstruction task. Here, the feature decoder is initialized in the way that our PD-INR start with a prior image. Specifically, the weights of the final layer are initialized as $[\mathbf{0}_{LF}^T, \mathbf{w}^T]^T$, such that $\mathbf{1}^T\mathbf{w} = 1$; for the other layers, the weights are simply initialized as identity matrices, while there are no bias in our MLP. And ReLU is used as the activation function in the MLP and also guarantees a non-negative image. For such an MLP of arbitrary number of layers, it could be found that the image intensity $f_\Theta(S_\Phi(\mathbf{c}))$ is initially set as $\mathbf{w}^T E_{\hat{\mathbf{x}}}(\mathbf{c})$, which is just a weighted average of neighbouring intensities in the prior image. In this way, our PD-INR could always start from a prior image since the very beginning, without any need for warm-up fitting.

Since then, our TOF-PET formulation and the internal design of PD-INR have been discussed in details. The overall framework is displayed in Fig. 1.

3 Experiments

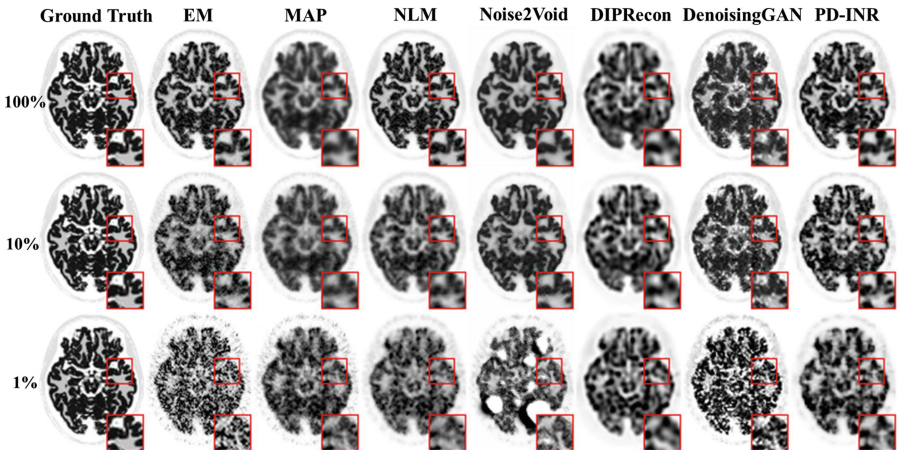


Fig. 2. TOF-PET reconstruction of different methods on brain data at various doses

3.1 Data and Settings

Data. We conduct experiments on two datasets: 1) Brain data, which are publicly available 3D PET images from BrainWeb [12]; 2) Chest data, which are

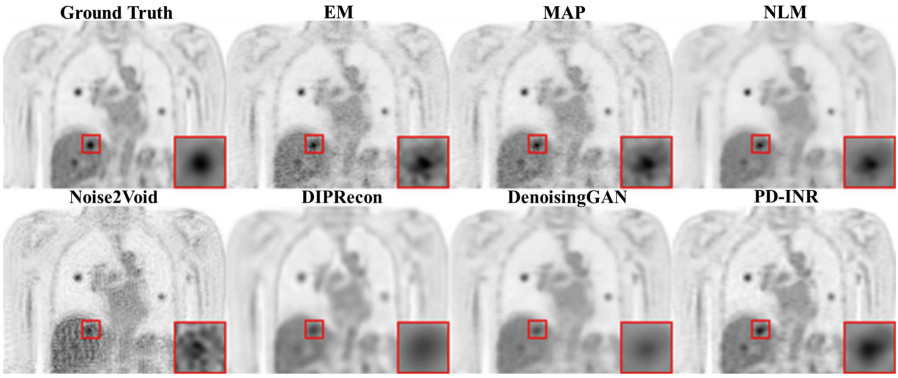


Fig. 3. TOF-PET reconstruction results of different methods on chest data with enlarged tumor ROI

clinical PET images collected from hospitals for patient diagnosis. Specifically, to investigate the algorithm's performance in tumor recovery, we insert multiple solid spheres with Gaussian textures into the chest images. The projection data are generated using the Joseph3D algorithm from the PyTomography library [21], with the number of TOF bins set to 13, and the TOF time resolution set to 200 ps. For all data, the image size is $128 \times 128 \times 96$, with a voxel size of $1.5 \times 1.5 \times 1.5$ mm^3, and the sinogram size is $224 \times 449 \times 4096 \times 13$.

Baselines. We compare our **PD-INR** with six baseline methods. **OSEM** [11] is a classical iterative reconstruction techniques for PET. **MAP** [13], is an iterative algorithm with enforced TV penalty. Non-Local Mean (**NLM**) [2] is an image post-processing method that preserves fine details while reducing noise. We also consider self-supervised learning methods, **Noise2Void** [25] and **DIPRecon** [5], which perform image restoration without requiring labeled data. **DenoisingGAN** [8] is used as a post-processing denoising module and serves as a representative of supervised learning methods.

Implementation Details. We implement our framework in PyTorch and use Adam optimizer to solve (5). All experiments are run on a single NVIDIA V100 GPU with 32 GB of RAM. The learning rate is set at 2×10^{-3} and progressively reduced to 2×10^{-4}, with a batch size of 16,384 rays (LORs) per iteration. For our framework, the prior image is obtained by performing an initial OSEM reconstruction on the raw projection data. We take single-layer MLP as the feature decoder and set $L = 4$ for H_Φ. Through hyperparameter analysis, as shown on the left of Fig. 5, our TV weight λ is set to 5×10^{-5}.

To evaluate the reconstruction quality, we employ PSNR and SSIM metrics [26]. PSNR measures artifact suppression, while SSIM assesses perceptual differences, with higher values indicating more accurate reconstruction. For local lesion recovery, we analyze CR and STD curves [5]. CR evaluates the contrast

Table 1. PSNR (dB) and SSIM of different methods on brain data at various doses

Method	**EM** [24]	**MAP** [13]	**NLM** [2]	**N2V** [25]	**DIP** [5]	**GAN** [8]	**PD-INR**
100%dose	31.4/.936	30.8/.948	31.5/.942	31.2/.680	30.8/.939	30.4/.954	**31.7**/**.964**
10%dose	20.2/.905	23.6/.927	23.8/.916	25.1/.653	25.2/.926	25.3/.916	**27.5**/**.942**
1%dose	19.6/.823	22.8/.875	22.5/.871	18.2/.642	24.4/.888	24.0/.844	**25.0**/**.893**

recovery between the tumor region and the background, and STD reflects the variability of background noise, together providing a comprehensive assessment of lesion recovery quality. All metrics are averaged over 20 independent noise realizations to ensure statistical reliability.

3.2 Results

Quantitative Evalution. Table 1 presents the experimental results on the brain data. Our proposed **PD-INR** consistently achieves highest PSNR and SSIM. Notably, as the dose decreases, the performance of classical iterative algorithms (e.g. **EM**) declines significantly, primarily due to the low statistics in the data. In contrast, **PD-INR** demonstrates a clear advantage in the case of low dose, outperforming even the supervised-learning method **DenoisingGAN** (abbreviated as **GAN** in the table). Thanks to our prior-driven encoding and designed TV regularization.

Figure 4 depicts the experimental results for small lesion recovery on the chest data, displaying CR and STD of several iterative algorithms. Results for lung and liver tumors are presented respectively. **PD-INR** shows relatively stable results, where its CR approaches to 1 in few epochs. This is explained by our inserted image prior, which enables fast convergence. Compared with other methods, **PD-INR** better suppresses noise with relatively small STD.

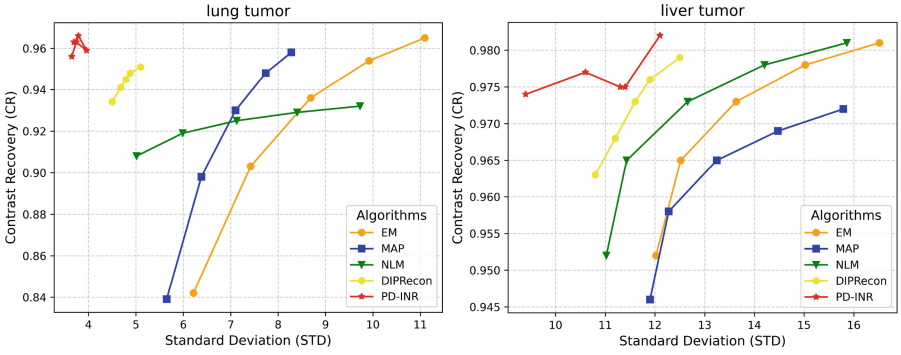


Fig. 4. Comparison of different methods with CR and STD for tumor recovery

Qualitative Evaluation. We also conduct visual comparison of different methods for qualitative evaluation. The reconstruction results are shown in Fig. 2 and Fig. 3. It could be observed that our **PD-INR** demonstrates superior overall reconstruction quality and more accurate recovery of small lesion. Compared to other methods, **PD-INR** could well preserve the image details while exhibiting robustness to data noises.

Ablation Studies. We conducted ablation studies on the chest data. We first found that removal of our prior-image encoder in PD-INR led to failure in convergence, possibly due to lack of expressive power. Then, we compare our PD-INR with following settings: 1) conventional INR framework, i.e. hash encoder + four-layer MLP; 2) removal of the TV regularization from our PD-INR. The results are shown on the right of Fig. 5. As expected, removal of regularization (i.e. the orange curve) encouraged PD-INR to overfit to statistical noises, hence PSNR starts to continuously drop after reaching the peak. With TV penalty (i.e. red curve), PD-INR converged to a steady estimate. Besides, it could be observed that our PD-INR (i.e. the orange curve) always performed better than conventional INR (i.e. the blue curve), demonstrating the apparent advantage of our prior-driven encoding.

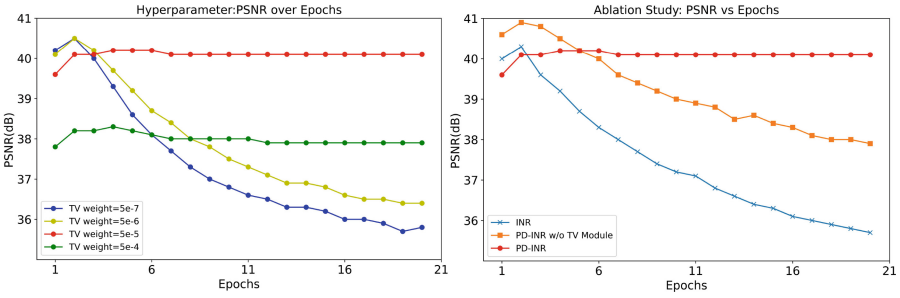


Fig. 5. Results of hyperparameter analysis (left) and ablation studies (right) on chest data using PD-INR

4 Conclusion

This paper proposes a self-supervised learning framework based on INR for TOF-PET reconstruction. We re-formulate TOF-PET reconstruction problem under INR framework, considering the imaging physics behind TOF-PET. To enhance image smoothness while preserving structural details, we introduce ray-based TV regularization that is specially customized for INR. Besides, we propose to employ multi-resolution hash encoder for effective feature learning, combined with our designed image-prior encoder. Experiments on brain and chest data

demonstrate that our PD-INR method outperforms existing baselines in noise suppression and contrast recovery, particularly for low-dose PET imaging. In the future, it is highly possible to extend our method to other imaging modalities.

Acknowledgments. This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62401516), and the National Natural Science Foundation of China (Grant No. 62427807).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Beyer, T., et al.: A combined PET/CT scanner for clinical oncology. J. Nucl. Med. **41**(8), 1369–1379 (2000)
2. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2, pp. 60–65. IEEE (2005)
3. Chen, Z., Yang, L., Lai, J.H., Xie, X.: Cunerf: cube-based neural radiance field for zero-shot medical image arbitrary-scale super resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21185–21195 (2023)
4. Corona-Figueroa, A., Frawley, J., Bond-Taylor, S., Bethapudi, S., Shum, H.P., Willcocks, C.G.: Mednerf: medical neural radiance fields for reconstructing 3d-aware CT-projections from a single x-ray. In: 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, pp. 3843–3848. IEEE (2022)
5. Gong, K., Catana, C., Qi, J., Li, Q.: Pet image reconstruction using deep image prior. IEEE Trans. Med. Imaging **38**(7), 1655–1665 (2018)
6. Gong, K., et al.: Iterative pet image reconstruction using convolutional neural network representation. IEEE Trans. Med. Imaging **38**(3), 675–685 (2018)
7. Gong, K., Guan, J., Liu, C.C., Qi, J.: Pet image denoising using a deep neural network through fine tuning. IEEE Trans. Radiat. Plasma Med. Sci. **3**(2), 153–161 (2018)
8. Gong, Y., et al.: Parameter-transferred Wasserstein generative adversarial network (PT-WGAN) for low-dose pet image denoising. IEEE Trans. Radiat. Plasma Med. Sci. **5**(2), 213–223 (2020)
9. Gunn, R.N., Slifstein, M., Searle, G.E., Price, J.C.: Quantitative imaging of protein targets in the human brain with pet. Phys. Med. Biol. **60**(22), R363 (2015)
10. Häggström, I., Schmidtlein, C.R., Campanella, G., Fuchs, T.J.: Deeppet: a deep encoder-decoder network for directly solving the pet image reconstruction inverse problem. Med. Image Anal. **54**, 253–262 (2019)
11. Hudson, H.M., Larkin, R.S.: Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans. Med. Imaging **13**(4), 601–609 (1994)
12. Kwan, R.S., Evans, A.C., Pike, G.B.: MRI simulation-based evaluation of image-processing and classification methods. IEEE Trans. Med. Imaging **18**(11), 1085–1097 (1999)
13. Levitan, E., Herman, G.T.: A maximum a posteriori probability expectation maximization algorithm for image reconstruction in emission tomography. IEEE Trans. Med. Imaging **6**(3), 185–192 (1987)

14. Liu, C.C., Qi, J.: Higher snr pet image prediction using a deep learning model and MRI image. Phys. Med. Biol. **64**(11), 115004 (2019)
15. Liu, J., Bai, X.: Volumenerf: CT volume reconstruction from a single projection view. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 743–753. Springer (2024). https://doi.org/10.1007/978-3-031-72104-5_71
16. Long, Y., Lyu, T., Rao, F., Yuan, P., Chen, Y., Zhu, W.: D-NAF: dynamic neural attenuation fields for 4d CBCT reconstruction in pulmonary imaging. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 3549–3554. IEEE (2024)
17. Machac, J.: Cardiac positron emission tomography imaging. In: Seminars in Nuclear Medicine, vol. 35, pp. 17–36. Elsevier (2005)
18. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2021)
19. Molaei, A., et al.: Implicit neural representation in medical imaging: a comparative survey. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2381–2391 (2023)
20. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. (TOG) **41**(4), 1–15 (2022)
21. Polson, L.A., et al.: Pytomography: a python library for medical image reconstruction. SoftwareX **29**, 102020 (2025)
22. Ran, Y., et al.: Ct-nerf: incremental optimizing neural radiance field and poses with complex trajectory. arXiv preprint arXiv:2404.13896 (2024)
23. Shen, L., Pauly, J., Xing, L.: NeRP: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. IEEE Trans. Neural Netw. Learn. Syst. **35**(1), 770–782 (2022)
24. Shepp, L.A., Vardi, Y.: Maximum likelihood reconstruction for emission tomography. IEEE Trans. Med. Imaging **1**(2), 113–122 (2007)
25. Song, T.A., Yang, F., Dutta, J.: Noise2void: unsupervised denoising of pet images. Phys. Med. Biol. **66**(21), 214002 (2021)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
27. Xu, J., et al.: Nesvor: implicit neural representation for slice-to-volume reconstruction in MRI. IEEE Trans. Med. Imaging **42**(6), 1707–1719 (2023)
28. Zha, R., Zhang, Y., Li, H.: NAF: neural attenuation fields for sparse-view CBCT reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 442–452. Springer, Cham (2022)
29. Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, B.R., Rosen, M.S.: Image reconstruction by domain-transform manifold learning. Nature **555**(7697), 487–492 (2018)

PFESA: FFT-Based Parameter-Free Edge and Structure Attention for Medical Image Segmentation

Mingqian Li[1], Zhiqian Yan[1], Miaoning Yan[2], Yaodong Liang[1], Qingmao Zhang[1(✉)], and Qiongxiong Ma[1(✉)]

[1] Guangdong Provincial Key Laboratory of Nanophotonic Functional Materials and Devices, School of Optoelectronic Science and Engineering, South China Normal University, Guangzhou, China
zhangqm@scnu.edu.cn , maqx@m.scnu.edu.cn
[2] Riton 3D., Ltd., Guangzhou, China

Abstract. The U-Net architecture remains pivotal in medical image segmentation, yet its skip connections often propagate redundant noise and compromise edge information. We propose a **P**arameter-**F**ree **E**dge and **S**tructure **A**ttention (**PFESA**) based on Fast Fourier Transform (FFT) to address these limitations. PFESA employs frequency-domain feature decoupling to separate high-frequency (edge details) and low-frequency (structural components) representations. Leveraging feature Signal-to-Noise Ratio(SNR) analysis, we devise dual attention paths: a High-frequency Edge Attention (EA) enhances gradient-sensitive regions to preserve anatomical contours, while a Low-frequency Structure Attention (SA) suppresses noise through energy redistribution. This frequency-aware attention mechanism enables adaptive feature refinement in skip connections without introducing trainable parameters. The parameter-free design ensures robustness against overfitting in medical datasets with scarce data. Extensive experiments on multi modal 2D/3D medical image datasets demonstrate PFESA's superiority over existing attention methods, achieving SOTA performance with statistically significant improvements in Dice Similarity Coefficient (DSC: +3.3% vs. baseline) and Hausdorff Distance metrics. Code is available at: https://github.com/59-lmq/PFESA.

Keywords: Medical Image Segmentation · Parameter-Free Attention · FFT-based Feature Decoupling

1 Introduction

As a cornerstone of computer-aided diagnosis, medical image segmentation accuracy directly determines the reliability of lesion localization and surgical planning. U-Net [1] and its variants, which employ skip connections to integrate

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 470–480, 2026.
https://doi.org/10.1007/978-3-032-04947-6_45

multi-scale features, have become benchmark architectures in this domain. However, these direct feature concatenation mechanisms present three inherent limitations [2]: 1) Low signal-to-noise ratio (SNR) background noise in shallow encoder features interferes with semantic reconstruction in decoders; 2) Progressive downsampling induces attenuation of high-frequency edge information; 3) Existing fusion mechanisms lack interpretability for establishing correspondence between feature weights and anatomical structures.

To address these challenges, attention mechanisms enhance critical regions through dynamic feature weighting. SENet [3] models global dependencies via channel compression but introduces overfitting risks in small-sample scenarios through fully-connected layers. CBAM [4] combines spatial-channel attention at the cost of quadratic computational complexity growth with channel numbers. ECA-Net [5] reduces parameters through 1D convolution but remains data-dependent. Attention U-Net [2] selectively enhances salient features via dual attention in skip connections. While improving performance, such parameterized modules risk overfitting in medical images' data-scarce context [6] and obscure clinical credibility due to their black-box nature - clinicians cannot verify spatial correlations between attention weights and anatomical landmarks [7]. Recent parameter-free attention methods (e.g., SimAM [8]'s energy function-based attention) enhance generalizability but neglect the frequency characteristics of medical images, which high-frequency components encode anatomical edges while low-frequency components represent organ morphology. Conventional convolutions struggle to decouple frequency information due to limited receptive fields, whereas spectral processing enables physically interpretable feature enhancement through global analysis. Liu et al. [6] enhances ViT high frequencies via Laplacian operators but lost low-frequency structures. SF-Net [7] uses rectangular masks extracting frequencies after FFT, causing discontinuous/noisy high frequencies. WaveSNet [9] replaces downsampling with wavelets but ignores high-frequency in skip connections, introducing extra noise. PFD-Net [10] applies 1×1 convolutions in frequency domain without frequency ban analysis. MC-FDN [22] neglects frequency continuity via channel attention on real/imaginary components after FFT. SASAN [23] fuses denoised low frequencies but discarded high-frequency feature.

To overcome these limitations, we propose a novel **P**arameter-**F**ree **E**dge-**S**tructure **A**ttention (**PFESA**) mechanism based on Fast Fourier Transform (FFT). Our contributions include three key aspects: 1) We implements Gaussian filtering to separate high-frequency (edge details) and low-frequency (structural components) features, avoiding optimization bias from trainable filters, and decoupling frequency-domain feature; 2) We propose a SNR-Driven Parameter-Free Attention which enhances edge SNR through local energy maximization in high-frequency components and reinforces structural morphology via contrast-enhanced normalization in low-frequency components. The statistical computation process eliminates trainable parameters, ensuring mathematically traceable weight generation. Grad-CAM visualization further enhances interpretability; 3) When integrated into skip connections of both CNN-based and Transformer-

based models, our method demonstrates state-of-the-art performance across multiple 2D/3D multi-modal medical image datasets, confirming its robustness.

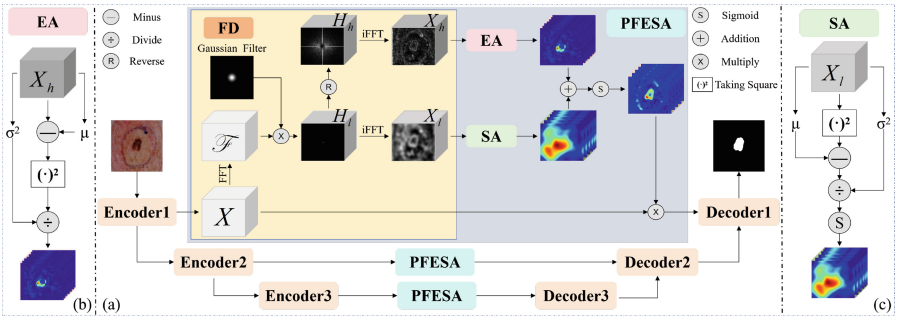


Fig. 1. An overview of the PFESA (a), which comprises Frequency-Domain **F**eature **D**ecouping, **E**dge **A**ttention (b) and **S**tructure **A**ttention (c).

2 Methodology

Figure 1 illustrates the overall framework of the proposed PFESA and shows an example usage of it. Since it is a parameter-free plug-to-play attention module with the same channels input and output channels, PFESA can be plugged into any skip connection. In this section, We first introduce the Frequency-Domain Feature Decouping (Sect. 2.1) and then concentrate on the details of Parameter-Free SNR-Driven Attention (Sect. 2.2).

2.1 Frequency-Domain Feature Decouping

Medical image frequency components exhibit distinct separability in the spectral domain, which high-frequency elements consist edges and textures while low-frequency components contain organ morphology. Using 2D data as an example (with 3D cases following the same principle), we first transform input feature maps $X \in \boldsymbol{R}^{C\times H\times W}$ into the frequency domain $\mathcal{F}(X) \in R^{C\times H\times W}$ via Fast Fourier Transform (FFT). Gaussian filtering $G_\sigma(u,v) = \exp\left(-\frac{u^2+v^2}{2\sigma^2}\right)$ is then applied for frequency-band decoupling: $H_l = G_\sigma \odot \mathcal{F}(x), H_h = 1 - H_l$, where the parameter $\sigma = r \cdot \frac{\min(H,W)}{\max(H,W)}$, $r = 0.1$ controls the cutoff frequency (see Sect. 3.3 for ablation studies). The high-frequency features X_h and low-frequency features X_l are subsequently reconstructed through inverse FFT.

2.2 Parameter-Free SNR-Driven Attention

The signal-to-noise ratio (SNR) of medical image features can be mathematically defined as the ratio of signal energy to noise energy [11]: $SNR = \frac{E[S^2]}{E[N^2]} = \frac{\mu_S^2}{\sigma_N^2}$, where σ_N^2 denotes noise variance and μ_s^2 approximates the signal energy intensity. This fundamental distinction necessitates differentiated computational frameworks for modeling SNR in high-frequency edge features versus low-frequency structural features.

Edge Attention (EA). High-frequency features S_h exhibit intense gradient variations where local energy (quantified as variance) significantly surpasses noise energy in anatomical edge regions of medical images. Assuming additive Gaussian noise distribution $N_h \sim N(0, \sigma_{n,h}^2)$, let $\sigma_{s,h}^2$ and $\sigma_{n,h}^2$ denote signal and noise variances respectively. In edge regions, $\sigma_{s,h}^2 \gg \sigma_{n,h}^2$ will be present. The edge SNR is formulated as:

$$SNR_{edge} = \frac{E[S_h^2]}{\sigma_{n,h}^2} \approx \frac{E[(X_h - \mu_h)^2]}{\sigma_{n,h}^2} \tag{1}$$

By substituting $\sigma_{n,h}^2$ with σ_h^2 (feature variance), we derive the Edge Attention (EA) map normalized to [0, 1].

$$EA = \frac{(X_h - \mu_h)^2}{\sigma_h^2}, \tag{2}$$

where σ_h^2 is feature variance. For high-frequency feature S_h, when $(X_h - \mu_h)^2 \approx \sigma_{s,h}^2, \sigma_h^2 \approx \sigma_{s,h}^2$ in edge regions, $EA \approx 1$; In noise-dominated regions, when $(X_h - \mu_h)^2 \approx \sigma_{n,h}^2, \sigma_h^2 \approx \sigma_{s,h}^2 + \sigma_{n,h}^2$, $EA \approx \frac{\sigma_{n,h}^2}{\sigma_{n,h}^2 + \sigma_{s,h}^2} \to 0$.

Structure Attention(SA). Low-frequency artifacts noise N_l in medical image typically exhibit energy distributions comparable to anatomical signals S_l [24]. To amplify their discriminability, we strategically enhance dominant regions through the squared mean-to-variance difference:

$$SNR_{structure} = \frac{E[S_l^2]}{\sigma_{n,l}^2} \approx \frac{E[(X_l^2 - \mu_l)]}{\sigma_{n,l}^2} \tag{3}$$

Substituting $\sigma_{n,l}^2$ with σ_l^2 and applying Sigmoid normalization yields the Structure Attention (SA):

$$SA = Sigmoid\left(\frac{X_l^2 - \mu_l}{\sigma_l^2}\right) \tag{4}$$

where σ_l^2 is feature variance. By applying Sigmoid for $SNR_{structure}$, SA is constrained to the range of [0.5, 1]. For low-frequency feature S_l, when $X_l^2 - \mu_l \gg 0, \sigma_l^2 \approx \sigma_{s,l}^2$, $SNR_{structure}$ and SA approaches 1; In noise-dominated regions, when $X_l^2 - \mu_l \approx 0$, $SNR_{structure}$ approaches 0 and SA approaches 0.5.

The high-frequency Edge Attention (EA) enhances boundary signal-to-noise ratio through localized energy normalization, while the low-frequency Structure Attention (SA) amplifies morphological discriminability via global variance contrast. These dual mechanisms operate complementarily in the spectral domain to collectively refine spatial-domain representations. The parameter-free design ensures mathematically traceable weight generation and anatomical coherence with medical image characteristics, guaranteeing physically plausible and interpretable feature enhancement. The final output after PFESA is obtained by fusing EA and SA through a Sigmoid-activated fusion gate, followed by element-wise multiplication with input features:

$$A = Sigmoid(EA + SA), X_{out} = A \odot X \tag{5}$$

3 Experiments

3.1 Datasets and Implementation

Datasets. We conduct extensive experiments on two 2D datasets (GlaS [12] and ISIC-2017 [13]) and two 3D datasets (LA [14] and Tooth [15]). GlaS is a gland segmentation dataset, including 165 Hematoxylin and Eosin (H&E) stained images of benign and malignant tissue. The number of training, valid and test images are 85, 20 and 60, respectively. ISIC-2017 is curated from the ISIC Challenge Dataset 2017 provided by the International Skin Imaging Collaboration (ISIC), comprising 2000 images for training, 150 images for validation and 600 images for test. LA is a left atrial segmentation dataset from 2018 Atrial Segmentation Challenge. It consists of 100 3D LGI-MRI images, with a resolution of $0.625 \times 0.625 \times 0.625$ mm. Following [16], we use 80 images for training and 20 images for testing. Tooth dataset is collected from a subset of a large scale dataset [15], which is used for segmentation and reconstruction of individual teeth and alveolar bone to aid in dental treatment. Following [17], we use 103 images for training and 26 images for testing.

Implementation Details. We implemented our framework using PyTorch. Training and inference of all models were performed on one NVIDIA GeForce RTX 4090. The input resolutions of 2D datasets and 3D datasets were 256^2 and 128^3, respectively. The loss function was DiceCE Loss. We ran 500 epochs for all 2D datasets and 10000 iterations for all 3D datasets. We used SGD optimizer, with initial learning rates of 0.1 and 0.01, decays of 1×10^{-5} and 1×10^{-4}, and momentum of 0.9 and 0.99 for 2D and 3D datasets, respectively. Performance was evaluated using the dice similarity coefficient (DSC,%), and hausdorff distance (HD, mm) for all the experiments.

3.2 Comparison with the State-of-the-Art Methods

The baseline frameworks comprise CNN-based models (U-Net [1] and 3D U-Net [18]) and Transformer-based frameworks (TransUNet [19] for 2D and

UNetr++ [20] for 3D). We compared our method against with parameterized attention mechanisms (CBAM [4], SE [3], ECA [5]) and parameter-free attention mechanisms (SimAM [8] and SIAM [21]). All attention modules are integrated into every skip connection of the baseline architectures. For a fair comparison, we used publicly available codes for all methods. We reported segmentation performance solely on individual model accuracy, without using pre-training, model ensembling, or extra data.

Table 1. Comparison experimental results obtained from four datasets, highlighting the best scores with **bold**.

2D Methods	GlaS		ISIC-2017		3D Methods	LA		Tooth	
	DSC↑	HD↓	DSC↑	HD↓		DSC↑	HD↓	DSC↑	HD↓
U-Net [1]	88.61	11.51	84.09	11.40	3D U-Net [18]	84.55	15.68	84.88	25.27
+CBAM [4]	87.31	16.23	83.43	10.67	+CBAM [4]	86.01	16.26	86.88	9.26
+SE [3]	84.21	17.49	83.05	12.01	+SE [3]	86.29	14.35	87.36	8.66
+ECA [5]	84.34	18.32	83.47	11.62	+ECA [5]	87.09	15.12	85.00	15.62
+SimAM [8]	87.44	12.07	83.87	11.77	+SimAM [8]	86.63	14.23	84.99	31.81
+SIAM [21]	88.05	14.66	83.79	12.43	+SIAM [21]	83.72	20.30	80.08	26.40
+PFESA (ours)	**90.18**	**8.13**	**84.25**	**10.12**	+PFESA (ours)	**87.17**	**12.91**	**88.21**	**8.22**
TransUNet [19]	88.83	13.76	84.55	10.16	UNetr++ [20]	89.78	6.75	86.27	2.47
+CBAM [4]	87.92	14.79	83.50	10.91	+CBAM [4]	89.47	7.82	84.20	4.99
+SE [3]	86.26	14.75	83.60	11.20	+SE [3]	89.40	10.13	87.30	2.38
+ECA [5]	88.06	13.25	84.44	10.36	+ECA [5]	89.52	11.98	86.97	2.40
+SimAM [8]	85.88	15.96	84.84	10.36	+SimAM [8]	90.09	6.74	87.22	2.34
+SIAM [21]	86.48	15.65	84.15	10.14	+SIAM [21]	88.65	7.41	81.44	9.45
+PFESA (ours)	**89.33**	**13.21**	**85.27**	**9.54**	+PFESA (ours)	**90.16**	**6.72**	**87.77**	**2.14**

Quantitative Analysis. Table 1 presents detailed quantitative comparisons, with **bold** indicating the best results. On the GlaS [12] dataset, other attention methods degrade performance versus baseline, which DSC reduced at least 0.56%, HD is increased by at least 0.56 mm, indicating that these methods are overfitting on scarce dataset. However, our method significantly reduces HD (reduced by 3.4 mm compared to U-Net [1] and 0.55 mm compared to TransUNet [19]) while improving DSC. In the ISIC-2017 [13] dataset, it is noted that except for the improvement of SimAM [8] added to TransUNet [19] over TransUNet [19], the DSC of other methods has decreased (DSC decreased by at least 0.11%). Our method has improved the performance of other methods (DSC increased by 0.16% compared to U-Net [1] and 0.72% compared to TransUNet [19]). It shows that our method can also be effective in the field of dermoscopic images and natural-like images. In the LA [14] dataset and Tooth [15] dataset, it can be found that when using 3D U-Net [18] as the baseline, our

method has significantly decreased in HD compared to other methods (HD of the Tooth [15] dataset decreased from 25.27 mm of the baseline to 8.22 mm), which highlights the superiority of our performance. When using UNetr++ [20] as the baseline, it can be found that the baseline is better than our method in 3D U-Net [18] (DSC of the LA [14] dataset increased by 2.61%), but after adding our method, the indicator is still improved (DSC from 89.78% to 90.16%). The results show that our method can achieve SOTA level on both CNN-Based Model and Transformer-Based Model, also achieving SOTA on datasets of different modalities, which verifies the robustness of our method.

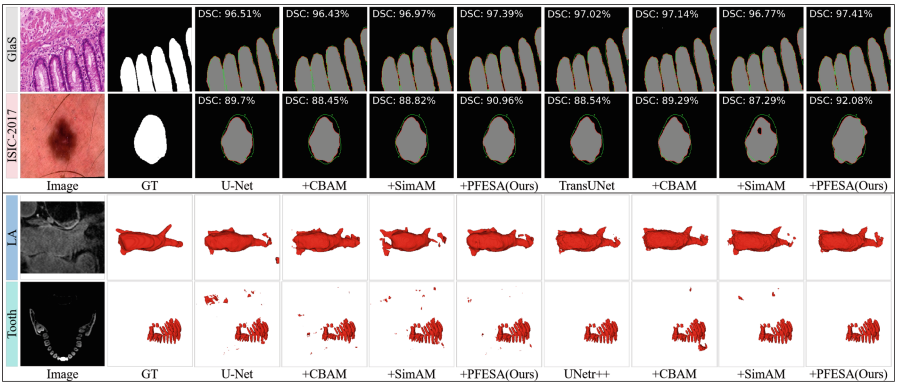


Fig. 2. Qualitative examples of four datasets. The first and second columns are the raw images and labels, and the other columns are the predictions of different methods.

Qualitative Analysis. Figure 2 shows some specific examples from the comparative experiments. Upon visually inspecting the results of other attention mechanisms, it is apparent that our method pays more attention to the boundary regions. Figure 3 presents the Grad-Cam visualizations of different layers of Skip Connections (SC-1, SC-2, SC-3) in TransUNet [19], applied to the GlaS [12] and ISIC-2017 [13] datasets using different attention mechanisms. It can be seen that in the lower layers with high background noise (SC-1), other methods lose edge information, while our method emphasizes feature edges and details. In the deeper layers (SC-2 and SC-3), while we focus on the edges, we also retain attention on the main structural features, whereas other methods only focus on certain regions of the target features. As shown in Table 1 and Fig. 3, our method achieves lower HD values compared to other approaches and effectively preserves high-frequency edge information at skip connections, enhancing the feature signal-to-noise ratio, which intuitively supports the reliability and interpretability of Eq. 2 and Eq. 4.

3.3 Ablation Study

All ablation experiments were conducted on the ISIC-2017 [13] dataset using U-Net [1]. Table 2 demonstrates the impact of varying cutoff frequencies (r) on frequency-band separation and segmentation performance. As r increases from 0.1 to 0.5, the Hausdorff Distance (HD) progressively rises (from 10.12 mm to 11.02 mm), indicating that broader low-frequency bandwidth captured by the Gaussian filter reduces high-frequency signal energy. This degradation in the SNR of skip connection directly correlates with performance decline (DSC decreased by 0.35%). Table 3 shows the impact of different components (FD: Frequency-Domain Feature Decoupling, SA: Structure Attention, EA: Edge Attention) on segmentation performance. When only FD is used, the sum of high- and low-frequency features replaces the original skip connection features. The results in Table 3 show that with only FD, DSC increases by 0.05% and HD decreases by 0.3 mm compared to the baseline. When only SA is used, the model focuses more on structural information at the skip connections, reducing some redundant features. When only EA is used, HD decreases by 0.45 mm compared to the baseline, indicating that the model focuses more on edge information at the skip connections. Our method achieves the best results, with HD reduced by 1.28 mm and DSC increased by 0.14%, showing that the model, while emphasizing edge information at the skip connections, also captures structural information and suppresses background noise. Table 4 compares feature fusion strategies. Direct replacement slightly degrades DSC (−0.31%) but improves HD (−0.46 mm), validating edge-specific enhancements. Additive fusion achieves peak DSC (84.26%) at the cost of compromised HD due to feature distribution distortion.

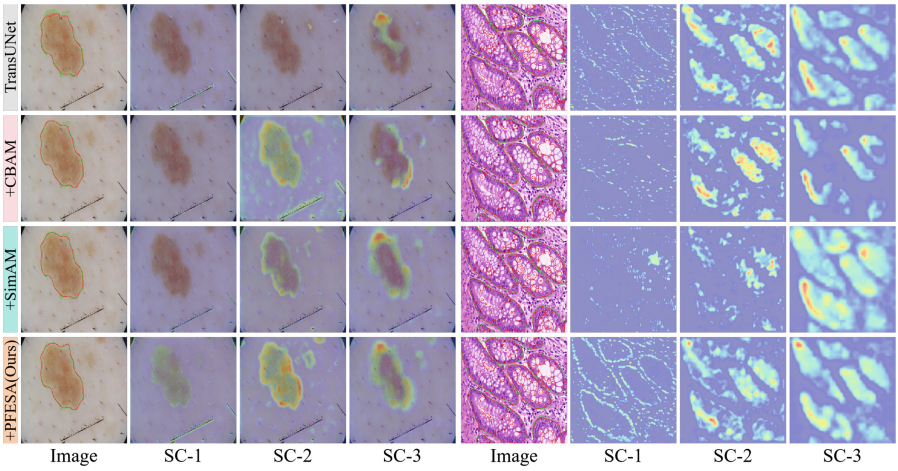


Fig. 3. Visualization of each skip connection feature using the trained TransUNet with different attention. The green and red line in Image represent Ground Truth and Prediction respectively. SC-1 means that low-level skip connection. SC-2 and SC-3 represent the high-level skip connection. (Color figure online)

Table 2. Ablation of the r value.

r	DSC↑	HD↓
0.1	**84.25**	**10.12**
0.2	83.21	11.53
0.3	84.19	11.49
0.4	84.08	10.78
0.5	83.90	11.02

Table 3. Ablation of the each component.

FD	SA	EA	DSC↑	HD↓
			84.09	11.40
✓			84.14	11.10
✓	✓		84.17	11.02
✓		✓	84.21	10.95
✓	✓	✓	**84.25**	**10.12**

Table 4. Ablation of the Feature fusion strategies.

Method	DSC↑	HD↓
baseline	84.09	11.40
Replace	83.78	10.94
ADD	**84.26**	10.59
Mul	84.25	**10.12**

Multiplicative fusion balances accuracy and robustness (DSC: 82.5%, HD: 10.12), preserving original feature statistics while integrating attention guidance.

4 Conclusion

In this study, we addresses noise interference and edge degradation in skip connections by proposing a Fast Fourier Transform (FFT)-based parameter-free edge and structural attention mechanism. Through frequency-domain decomposition, we disentangle edge and structural features, then construct SNR-driven attention modules to collaboratively suppress noise while preserving critical edge information. Experimental results demonstrate state-of-the-art performance across multi modal datasets with varying baselines, confirming the method's robustness. Our work establishes a novel frequency feature decoupling paradigm for medical image segmentation, offering fresh insights into architectural design through interpretable frequency-space operations.

Acknowledgements. This work was supported by the Guangdong Province Key Field R&D Plan (2023B0909030003), the Guangdong Basic and Applied Basic Research Foundation (2023A1515012966 and 2023A1515011641) and the National Excellent Engineers Innovation Institute of Guangdong-Hong Kong-Macao Greater Bay Area (Foshan) Advanced Manufacturing Industry (JBGS2024001).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
2. Schlemper, J., et al.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53**, 197–207 (2019)
3. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. IEEE, Salt Lake City (2018)

4. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
5. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: ECA-net: efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11534–11542. IEEE, Seattle (2020)
6. Qiong, L., Chaofan, L., Jinnan, T., Liping, C., Jianxiang, S.: Medical image segmentation based on frequency domain decomposition SVD linear attention. Sci. Rep. **15**(1), 2833 (2025)
7. Zhou, Z., He, A., Wu, Y., Yao, R., Xie, X., Li, T.: Spatial-frequency dual domain attention network for medical image segmentation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 4076–4081. IEEE, Lisbon (2024)
8. Yang, L., Zhang, R. Y., Li, L., Xie, X.: SimAM: a simple, parameter-free attention module for convolutional neural networks. In: Proceedings of the 38th International Conference on Machine Learning, pp. 11863–11874. PMLR (2021)
9. Li, Q., Shen, L.: WavesNet: wavelet integrated deep networks for image segmentation. In: Yu, S., et al. (eds.) PRCV 2022, LNCS, vol. 13537, pp. 325–337. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-18916-6_27
10. Yang, C., Zhang, Z.: PFD-net: pyramid Fourier deformable network for medical image segmentation. Comput. Biol. Med. **172**, 108302 (2024)
11. Hoogeboom, E., Heek, J., Salimans, T.: simple diffusion: end-to-end diffusion for high resolution images. In: Proceedings of the 40th International Conference on Machine Learning, pp. 13213–13232. PMLR, Hawaii (2023)
12. Sirinukunwattana, K., et al.: Gland segmentation in colon histology images: the GlaS challenge contest. Med. Image Anal. **35** 489–502 (2017)
13. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomed iCAL imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172. IEEE (2018)
14. Xiong, Z., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Med. Image Anal. **67** 101832 (2021)
15. Cui, Z., et al.: A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. Nat. Commun. **13**(1), 2096 (2022)
16. Zhou, Y., Huang, J., Wang, C., Song, L., Yang, G.: XNet: wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21085–21096. IEEE, Paris (2023)
17. Zhong, J., et al.: PMFSNet: polarized multi-scale feature self-attention network for lightweight medical image segmentation. Comput. Methods Programs Biomed. **261**, 108611 (2025)
18. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49

19. Chen, J., et al.: TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med. Image Anal. **97**, 103280 (2024)
20. Shaker, A.M., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: UNETR++: delving into efficient and accurate 3D medical image segmentation. IEEE Trans. Med. Imaging **43**(9) (2024)
21. Han, G., Huang, S., Zhao, F., Tang, J.: SIAM: a parameter-free, spatial intersection attention module. Pattern Recogn. **153**, 110509 (2024)
22. Caixia, D., Duwei, D., Zongfang, L., Songhua, X.: A novel deep network with triangular-star spatial–spectral fusion encoding and entropy-aware double decoding for coronary artery segmentation. Inf. Fusion **112**, 102561 (2024)
23. Xingru, H., Jian, H., Kai, Z., Tianyun, Z., Zhi, L., Changpeng, Y.: SASAN: spectrum-axial spatial approach networks for medical image segmentation. IEEE Trans. Med. Imaging **43**(8), 3044–3056 (2024)
24. Tanguay, J., Kim, J., Kim, H.K., Iniewski, K., Cunningham, I.A.: Frequency-dependent signal and noise in spectroscopic X-ray imaging. Med. Phys. **47**(7), 2881–2901 (2020)



Physics-Informed Implicit Neural Representations for Joint B0 Estimation and Echo Planar Imaging

Wenqi Huang[1(✉)], Nan Wang[2], Congyu Liao[2,3], Yimeng Lin[2], Mengze Gao[2], Daniel Rueckert[1,4], and Kawin Setsompop[2]

[1] Technical University of Munich, Munich, Germany
wenqi.huang@tum.de
[2] Stanford University, Stanford, USA
[3] University of California, San Francisco, USA
[4] Imperial College London, London, UK

Abstract. Echo Planar Imaging (EPI) is widely used for its rapid acquisition but suffers from severe geometric distortions due to B0 inhomogeneities, particularly along the phase encoding direction. Existing methods follow a two-step process: reconstructing blip-up/down EPI images, then estimating B0, which can introduce error accumulation and reduce correction accuracy. This is especially problematic in high B0 regions, where distortions align along the same axis, making them harder to disentangle. In this work, we propose a novel approach that integrates Implicit Neural Representations (INRs) with a physics-informed correction model to jointly estimate B0 inhomogeneities and reconstruct distortion-free images from rotated-view EPI acquisitions. INRs offer a flexible, continuous representation that inherently captures complex spatial variations without requiring predefined grid-based field maps. By leveraging this property, our method dynamically adapts to subject-specific B0 variations and improves robustness across different imaging conditions. Experimental results on 180 slices of brain images from three subjects demonstrate that our approach outperforms traditional methods in terms of reconstruction quality and field estimation accuracy (Code available: https://github.com/wenqihuang/PINR).

Keywords: MRI · Echo Planar Imaging · B0 Inhomogeneity · Implicit Neural Representations · Distortion Correction

1 Introduction

Magnetic Resonance Imaging (MRI) is widely used in clinical and research settings for non-invasive imaging with excellent soft tissue contrast. Echo Planar Imaging (EPI), a commonly used fast MRI technique, enables rapid acquisition of images, making it particularly advantageous for functional and diffusion imaging [14,18]. However, its high-speed acquisition causes geometric distortions

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 481–490, 2026.
https://doi.org/10.1007/978-3-032-04947-6_46

along the phase encoding direction due to B0 field inhomogeneities, leading to stretching, compression, intensity modulation, and misalignment, especially near air-tissue interfaces [1,16]. These artifacts compromise spatial accuracy, affecting quantitative analysis and clinical diagnosis.

Several techniques have been developed to correct B0-induced distortions in EPI. Some use pre-calculated field maps acquired before the scan, but these require extra scan time and may suffer from physiological B0 fluctuations (e.g., breathing, cardiac pulsation, motion) and error accumulation [22]. Others, such as blip-up/down acquisitions, estimate B0 by capturing images with opposite phase encoding gradients [1,2,9,10,16]. The distortions in the images with opposite phase encoding appear also in opposite directions, allowing B0 estimation through displacement analysis. While avoiding pre-scanning, this approach captures distortions along a single axis, making it difficult to resolve complex B0 variations, especially in high-susceptibility regions. Some studies have explored multi-shot EPI [3,11], but all of the existing methods are restricted to orthogonal directions. Acquiring phase encoding in rotated directions can capture more information of image distortion in multiple orientations, but the reconstruction and post-processing of off-grid data is challenging with conventional discrete representations. They could also encounter accumulated errors, as any inaccuracies in distorted image reconstruction, B0 estimation, propagate into the reconstructed image.

Implicit neural representations (INRs) have gained significant attention in computer vision and medical imaging for modeling complex signals. Unlike grid-based methods, INRs encode signals as continuous functions, capturing fine details without fixed resolution constraints. They have been successfully applied to CT and MRI reconstruction [4,7,8,13,15,17], making them a promising approach for B0 estimation and image reconstruction in EPI.

In this work, we propose a novel joint B0 estimation and image reconstruction framework that utilizes two INR networks, one for modeling the B0 field and another for the distortion-free image. Our method integrates an MR physics-informed forward model that warps the distortion-free image into distorted EPI images, which are then compared with multi-coil rotated-view EPI acquisitions to supervise the training of both networks. The rotated EPIs distribute distortions across multiple orientations, facilitating the disentanglement and correction of distortions. Due to the continuous representation capability of INRs, rotation is handled directly in the coordinate space rather than requiring interpolation in the image space. By jointly optimizing B0 estimation and image reconstruction, our approach mitigates error accumulation and enhances robustness across diverse imaging conditions. Our key contributions are as follows:

1. We propose a novel **P**hysics-informed **I**mplicit **N**eural **R**epresentation B0 estimation and image reconstruction framework, termed as PINR, for rotated multi-view EPI. We realized joint parallel imaging reconstruction across multi-views at the same time as estimating B0 field inhomogeneity all in one go rather than separately, and no need for interpolation thanks to the continuous representation capability of INRs.

2. We leverage rotated-view EPI acquisitions to improve B0 field estimation by distributing distortions across multiple orientations, facilitating better disentanglement and correction.
3. Our approach eliminates the need for additional calibration scans, making the method more time-efficient while maintaining high reconstruction accuracy. By integrating INR-based B0 estimation directly into the reconstruction process through a physics-informed forward model, we reduce error accumulation and enhance robustness across diverse imaging conditions.
4. We evaluated our method on high-resolution MRI simulated rotated-view EPI data and prospectively sampled blip-up/down acquisitions, both shows superier image reconstruction and B0 estimation comparing to state-of-the-art methods.

2 Methods

2.1 EPI and B0 Inhomogeneity

EPI achieves fast acquisition through its unique single-shot or multi-shot echo train sampling trajectory. Unlike Cartesian sequences that segment k-space acquisition, EPI samples k-space in a zig-zag pattern using alternating readout gradients. With a homogeneous B0 field, the acquired k-space directly corresponds to a distortion-free image after inverse Fourier transformation. However, B0 inhomogeneity introduces an additional phase term at each k-space point, dependent on local field variations. Mathematically, the acquired EPI k-space data. Mathematically, the acquired EPI k-space data $d(k_x, k_y, t)$ is expressed as:

$$d(k_x, k_y, t) = \int_{\Omega} \mathcal{S}(x, y) m(x, y) e^{-i(k_x x + k_y y)} e^{-i\gamma B_0(x,y)t} dx dy, \tag{1}$$

here $m(x, y)$ is the true image intensity, $\mathcal{S}(x, y)$ represents coil sensitivity maps, γ is the gyromagnetic ratio, and $B_0(x, y)$ represents the B0 field inhomogeneity at location (x, y). The integration domain Ω represents the spatial region over which the MRI signal is acquired, encompassing all imaged tissue areas. The undesired phase term $e^{-i\gamma B_0(x,y)t}$ accrual alters the spatial encoding, leading to geometric distortions and signal misregistration in the reconstructed image, particularly along the phase encoding direction.

Correcting distortions requires estimating the underlying B0 field $B_0(x, y)$. Traditional approaches involve acquiring field maps or using blip-up/blip-down acquisitions, but face limitations in scan time and resolving distortions in highly inhomogeneous regions. We propose to use rotated-view EPI acquisitions, which distribute distortions across multiple orientations, allowing for improved B0 estimation and correction. Let θ_n represent the phase encoding direction for the n-th rotated EPI acquisition, with N total rotations. The B0 field estimation and image reconstruction problem can then be formulated as a joint optimization across multiple rotated views:

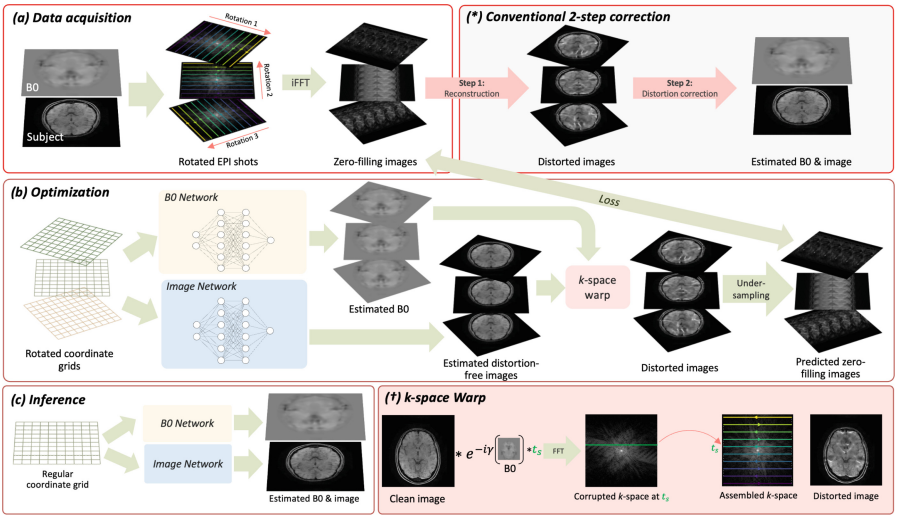


Fig. 1. Overview of the proposed PINR. **(a) Rotated-view EPI acquisition:** Acquiring EPI images at different phase encoding directions distributes distortions and aliasing due to B0 inhomogeneity and k-space undersampling. **(*) Existing methods** use two views (blip-up/down) to reconstruct a distorted image, then estimate B0 for correction. **(b) Optimization:** Our method inputs rotated coordinate grids into INR networks for B0 and distortion-free image estimation, enforcing MRI physics for training. k-space warping is indicated by (†). **(c) Inference:** A regular spatial grid is fed into trained INR networks to obtain the B0 field and distortion-free image

$$\widehat{B_0}, \hat{m} = \arg\min_{B_0,m} \sum_{n=1}^{N} \|d_n(k_x,k_y) - \sum_s \Lambda_{n,s} \mathcal{A}_n(m(x,y)e^{-i\gamma B_0(x,y)t_s})\|^2 + \mathcal{R}(B_0,m), \tag{2}$$

where $\mathcal{A}_n = \mathcal{M}_n \mathcal{F} \mathcal{S}$ represents the forward operator of MRI incorporating the coil sensitivity maps $\mathcal{S}$, the Fourier transform $\mathcal{F}$, and the undersampling mask $\mathcal{M}_n$ corresponding to the n-th rotated acquisition. t_s denotes the echo time of segment s, while the selection mask $\Lambda_{n,s}$ extracts k-space segments, summing over s to reconstruct the full k-space. The term $\mathcal{R}(B_0, m)$ acts as a regularization to enforce prior constraints on B_0 and m. When $N = 2$ and $|\theta_1 - \theta_2| = 180°$, this formulation reduces to the standard joint B0 estimation and image reconstruction problem for blip-up/down EPI acquisitions.

2.2 Implicit Neural Representations for Rotated-View EPI

Typical approaches involve sequentially estimating B0 before jointly reconstructing the k-space data. This limits the acceleration of each EPI shot, as highly undersampled acquisitions make it difficult to estimate a reliable B0 field upfront.

Additionally, the reliance on grid-based processing has restricted exploration to four primary directions (up, down, left, right), limiting the flexibility of EPI acquisitions. When only two opposing phase encoding directions are used, resolving distortions is further constrained by limited directional information. To overcome these limitations, we leverage INRs to model both the image and B0 field as continuous functions, enabling a more flexible and accurate reconstruction framework. INRs are coordinate-based deep learning models that map spatial coordinates (e.g., (x, y)) to signal values (e.g., image intensity) through a neural network. Unlike traditional grid-based representations, INRs provide continuous modeling, naturally capturing fine details without pixel grid constraints. This property makes them particularly well-suited for rotated-view EPI. We propose using an INR to represent both the image $m(x, y)$ and the B0 field $B_0(x, y)$ as:

$$m(x, y) = f_\psi(x, y), \quad B_0(x, y) = g_\phi(x, y), \tag{3}$$

where f_ψ and g_ϕ are neural networks parameterized by ψ and ϕ, respectively. These networks take spatial coordinates as input and output the corresponding intensity and field values. A key advantage of INRs is their ability to handle rotated-view acquisitions without interpolation. Instead of rotating the image or B0 field directly, INRs allow us to simply rotate the coordinate grid and query the function at the new locations:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta_n & -\sin\theta_n \\ \sin\theta_n & \cos\theta_n \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \tag{4}$$

This property ensures that rotated views can be handled seamlessly without introducing interpolation artifacts, thereby preserving spatial fidelity.

By incorporating INRs into our joint B0 estimation and image reconstruction framework, we gain a flexible and high-fidelity approach to modeling spatial variations. The continuous representation allows our method to generalize across different imaging conditions while maintaining robustness to distortions.

2.3 Model Optimization and Inference

To jointly estimate the B0 field and reconstruct the distortion-free image, we optimize the parameters ψ and ϕ of the INR networks $f_\psi(x, y)$ and $g_\phi(x, y)$. By incorporating these networks into the joint optimization framework in Eq. (2), the learning objective becomes:

$$\widehat{\psi}, \widehat{\phi} = \arg\min_{\psi,\phi} \sum_{n=1}^{N} \|\mathcal{A}_n^H d_n(k_x, k_y) - \mathcal{A}_n^H \sum_s \Lambda_{n,s} \mathcal{A}_n(f_\psi(x, y) e^{-i\gamma g_\phi(x,y) t_s})\|_{sl1} + \mathcal{R}(f_\psi, g_\phi), \tag{5}$$

where $\mathcal{A}_n$ is the MRI forward operator incorporating sampling, Fourier transform, and coil sensitivity maps for the n-th view, $\mathcal{A}_n^H$ is its adjoint operator that convert k-space to zero-filled images. The inclusion of $\mathcal{A}_n^H$ allows loss calculation in image space, mitigating issues caused by the large intensity range of k-space values. We choose smoothed L1 loss $\|\cdot\|_{sl1}$ for the data consistency to prevent

exploding gradients like [5]. $\mathcal{R}(f_\psi, g_\phi)$ serves as a regularization term to enforce smoothness and prior constraints. The process is illustrated in Fig. 1b.

Once the models are well-trained, B0 and the distortion-free image can be obtained by querying the INR networks over a regular grid. Specifically, by passing a uniform spatial coordinate grid (x, y) into $g_\phi(x, y)$, we obtain the estimated B0 field $\widehat{B_0}$. Similarly, by querying $f_\psi(x, y)$, we obtain the final high-fidelity distortion-free image $\hat{m}$ (see Fig. 1c):

$$\widehat{B_0}(x, y) = g_{\widehat{\phi}}(x, y), \quad \hat{m}(x, y) = f_{\widehat{\psi}}(x, y). \tag{6}$$

3 Experimental Setup

Dataset. We simulated distorted EPI data at different rotated angles using undistorted images from an accelerated multi-echo GRE dataset acquired with EPTI [20]. The B0 map was estimated from multi-echo images [21], and coil sensitivity was obtained using ESPIRiT [19]. Data were collected from three healthy volunteers on a 3T scanner (UHP, GE Healthcare) with a $240 \times 216 \times 240$ mm^3 FOV, 1-mm isotropic resolution, 48 echoes (1 ms echo spacing, TE: 553 ms, TR: 60 ms), and a 45-s acquisition time ($R = 48$), with ethical consent. Each scan included 90 brain slices, totaling 270 slices. Distorted images were simulated with a 0.25 ms echo spacing for fully sampled data and $R \times 0.25$ ms for undersampling. Two datasets were generated: (1) A two-view set ($R = 4$, 0° and 180°) simulating blip-up/down, and (2) A three-view set ($R = 6$, 0°, 120°, 240°) for rotated-view EPI. Ground truth images in the three-view set were interpolated onto a rotated grid, introducing potential systematic errors.

Model Details. The B0 network and image network are both multi-layer perceptrons (MLPs) with hash-grid encoding [12]. The encoding uses a hashmap of size 20 with 16 levels, 2 features per level, a base resolution of 16, and a per-level scale of 1.19. Each MLP has 2 hidden layers with 256 nodes. The B0 network outputs a scalar, while the image network outputs two channels representing the real and imaginary parts of the distortion-free image. The model was trained with a smoothed L1 loss for data consistency and total variation (TV) regularization to enforce smooth B0 estimation. The TV weight started at 1×10^{-5}, decaying by 0.1 every 1000 iterations, and was removed in the last 1000 iterations to fully utilize the INR flexibility. Training used AdamW with a 3×10^{-3} learning rate for 6000 iterations. Our proposed method implemented in PyTorch on an NVIDIA A6000 GPU took around 4 min per slice.

Evaluation Settings. We evaluated our method against the state-of-the-art TOPUP method from FSL [1,16], using GRAPPA [6] to reconstruct blip-up/down images before correction. As TOPUP does not support rotated acquisitions, we conducted two experiments: (1) Comparing B0 estimation and image correction using two EPI views (blip-up/down) to assess our method's effectiveness, and (2) Evaluating our method with three-view EPI and comparing it to

both TOPUP and our two-view method. To ensure fair k-space coverage, we set $R = 4$ for two-view and $R = 6$ for three-view experiments, maintaining the same total acquired k-space lines. Quantitative evaluation included NRMSE, PSNR, and SSIM for images, and absolute difference for B0 estimation. The TOPUP method processed a full volume of 90 slices in about 10 h for this high-resolution data.

Table 1. Comparison of Image and B0 Metrics (Mean ± Std)

Method	Image Metrics			B0 Metrics
	NRMSE ↓	PSNR (dB) ↑	SSIM ↑	Diff. (Hz) ↓
TOPUP	0.0534 ± 0.0153	37.1972 ± 2.0703	0.9666 ± 0.0171	3.0505 ± 1.3535
PINR-2v	**0.0143** ± 0.0169	**51.9853** ± 7.1921	**0.9955** ± 0.0054	**0.6017** ± 0.3742
PINR-3v	0.0441 ± 0.0104	38.7221 ± 1.8140	0.9805 ± 0.0097	1.0657 ± 0.4628

4 Results

Table 1 summarizes the mean and standard deviation of quantitative metrics for TOPUP, PINR with two views (PINR-2v), and three views (PINR-3v). Our method outperforms TOPUP across all metrics. While PINR-3v shows slightly higher B0 error and lower image quality metrics than PINR-2v, likely due to interpolation artifacts in multi-view image generation, it still surpasses the two-step, image-domain-based TOPUP method.

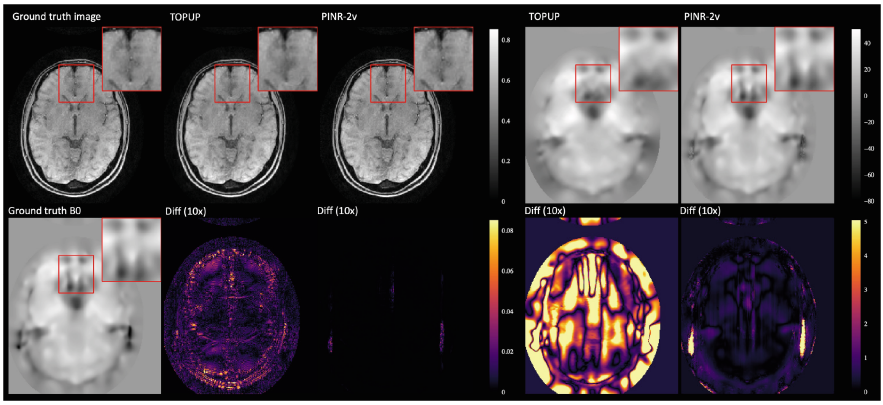


Fig. 2. Comparison with two-view (blip-up/down) acquisition. The first column shows the ground truth image and B0 field. The second and third columns present TOPUP-corrected results and our method's results, along with their error maps. The right two columns display B0 estimates from TOPUP and our method, with corresponding error maps below. The red box highlights the zoomed-in region of interest (ROI) (Color figure online)

Figure 2 compares TOPUP and our two-view PINR. The ROI shows that TOPUP produces blurrier images, while PINR preserves finer structures. Error maps reveal residual aliasing in TOPUP, likely from GRAPPA reconstruction errors. Since TOPUP operates in the image domain with separate reconstruction and B0 correction steps, errors in distorted EPI propagate into B0 estimation, degrading the final image. PINR consistently achieves lower B0 error, particularly in the ROI, benefiting from INRs' flexibility and joint optimization. In contrast, TOPUP's Gaussian-based B0 estimation oversmooths high-susceptibility regions. These results highlight the advantage of our joint framework for precise B0 estimation and distortion correction.

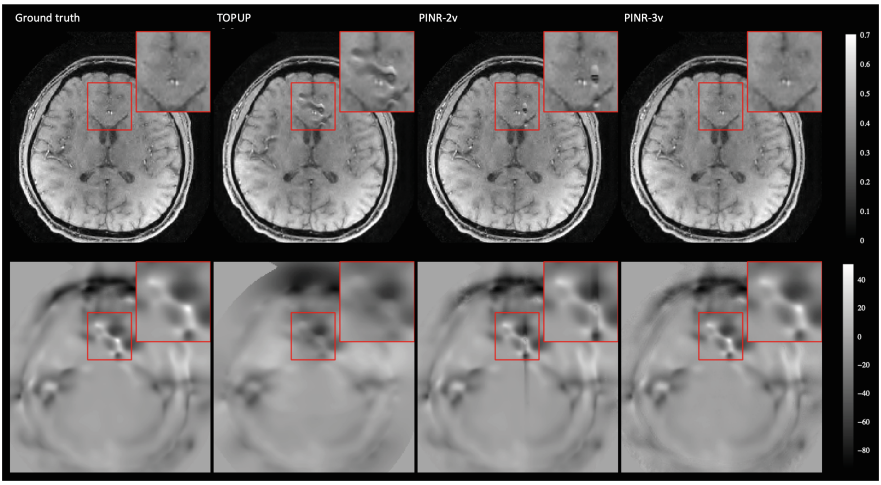


Fig. 3. Comparison of two-view and three-view results. The first row presents the ground truth and corrected images, while the second row shows the corresponding B0 field estimates. The red box highlights the region of interest (ROI) (Color figure online)

Figure 3 shows a challenging case where a high B0 peak in the ROI severely distorts the image. With only two views, distortions and aliasing align along the same phase encoding direction, making correction difficult. TOPUP fails to restore the region, leaving unresolved tissue compression and oversmoothed B0 estimation. PINR-2v captures more structural details but struggles at the high peak, causing local misestimation. In contrast, PINR-3v leverages additional rotated views to better disentangle distortions, accurately estimating the B0 peak and preserving anatomical structures. These results highlight the benefits of multi-view PINR for handling severe distortions beyond conventional blip-up/down methods.

5 Discussion and Conclusion

We proposed PINR, a joint B0 estimation and image reconstruction framework using INRs for distortion correction in EPI images. Unlike conventional methods relying on explicit field maps or blip-up/down acquisitions, PINR optimized B0 and the distortion-free image within a physics-informed framework. Rotated-view EPI acquisitions helped distribute distortions across multiple orientations, improving B0 estimation and reconstruction, while INRs naturally handled off-grid multi-view data, making them well-suited for rotational sampling. Our experiments showed that PINR outperformed TOPUP across multiple metrics. Currently, the method operates slice-wise, limiting inter-slice consistency and efficiency. Additionally, validation was performed on simulated B0 data due to challenges in acquiring high-resolution field maps. More complex field imperfection such T2 blurring and eddy current are also not discussed. Future work will address these limitations with a 3D extension and in-vivo evaluation. Our results highlight the potential of INRs for MRI field imperfection correction, advancing physics-informed deep learning in medical imaging.

Acknowledgments. This research was partially supported by European Research Council (Deep4MI project, Grant Agreement Nr.884622) and National Institutes of Health research grants (R01HD114719, K99EB035178, R01MH116173). Wenqi Huang's research visit to Stanford University was also partly funded by TUM Graduate School Internationalization Support Grant.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andersson, J.L., Skare, S., Ashburner, J.: How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. Neuroimage **20**(2), 870–888 (2003)
2. Andersson, J.L., Sotiropoulos, S.N.: An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. Neuroimage **125**, 1063–1078 (2016)
3. Bhushan, C., Joshi, A.A., Leahy, R.M., Haldar, J.P.: Improved b0-distortion correction in diffusion MRI using interlaced q-space sampling and constrained reconstruction. Magn. Reson. Med. **72**(5), 1218–1232 (2014)
4. Feng, J., et al.: Spatiotemporal implicit neural representation for unsupervised dynamic MRI reconstruction. IEEE Trans. Med. Imaging (2025)
5. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
6. Griswold, M.A.: Generalized autocalibrating partially parallel acquisitions (GRAPPA). Magn. Reson. Med.: Off. J. Int. Soc. Magn. Reson. Med. **47**(6), 1202–1210 (2002)
7. Huang, W., Li, H.B., Pan, J., Cruz, G., Rueckert, D., Hammernik, K.: Neural implicit k-space for binning-free non-cartesian cardiac MR imaging. In: Frangi,

A., de Bruijne, M., Wassermann, D., Navab, N. (eds.) IPMI 2023. LNCS, vol. 13939, pp. 548–560. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-34048-2_42
8. Kunz, J.F., Ruschke, S., Heckel, R.: Implicit neural networks with Fourier-feature inputs for free-breathing cardiac MRI reconstruction. IEEE Trans. Comput. Imaging (2024)
9. Liao, C., et al.: Distortion-free, high-isotropic-resolution diffusion MRI with gSlider BUDA-EPI and multicoil dynamic B0 shimming. Magn. Reson. Med. **86**(2), 791–803 (2021)
10. Liao, C., Cao, X., Cho, J., Zhang, Z., Setsompop, K., Bilgic, B.: Highly efficient MRI through multi-shot echo planar imaging. In: Wavelets and Sparsity XVIII, vol. 11138, pp. 353–365. SPIE (2019)
11. Mani, M., Jacob, M., Kelley, D., Magnotta, V.: Multi-shot sensitivity-encoded diffusion data recovery using structured low-rank matrix completion (mussels). Magn. Reson. Med. **78**(2), 494–507 (2017)
12. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. (TOG) **41**(4), 1–15 (2022)
13. Reed, A.W., et al.: Dynamic CT reconstruction from limited views with implicit neural representations and parametric motion fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2258–2268 (2021)
14. Setsompop, K., et al.: Improving diffusion MRI using simultaneous multi-slice echo planar imaging. Neuroimage **63**(1), 569–580 (2012)
15. Shen, L., Pauly, J., Xing, L.: NeRP: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. IEEE Trans. Neural Netw. Learn. Syst. **35**(1), 770–782 (2022)
16. Smith, S.M., et al.: Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage **23**, S208–S219 (2004)
17. Spieker, V., et al.: ICoNIK: generating respiratory-resolved abdominal MR reconstructions using neural implicit representations in k-space. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Zhu, D., Yuan, Y. (eds.) MICCAI 2023. LNCS, vol. 14533, pp. 183–192. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-53767-7_18
18. Stehling, M.K., Turner, R., Mansfield, P.: Echo-planar imaging: magnetic resonance imaging in a fraction of a second. Science **254**(5028), 43–50 (1991)
19. Uecker, M., et al.: ESPIRiT–an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA. Magn. Reson. Med. **71**(3), 990–1001 (2014)
20. Wang, F., et al.: Echo planar time-resolved imaging (EPTI). Magn. Reson. Med. **81**(6), 3599–3615 (2019)
21. Wang, N., et al.: Spherical echo-planar time-resolved imaging (sEPTI) for rapid 3D quantitative T 2* and susceptibility imaging. Magn. Reson. Med. **93**(1), 121–137 (2025)
22. Zeng, H., Constable, R.T.: Image distortion correction in EPI: comparison of field mapping with point spread function mapping. Magn. Reson. Med.: Off. J. Int. Soc. Magn. Reson. Med. **48**(1), 137–146 (2002)

Physiological Neural Representation for Personalised Tracer Kinetic Parameter Estimation from Dynamic PET

Kartikay Tehlan[1,2(✉)] and Thomas Wendler[1,2]

[1] Department of diagnostic and interventional Radiology and Neuroradiology, University Hospital Augsburg (UKA), Stenglinstr. 2, 86156 Augsburg, Germany
kartikay.tehlan@med.uni-augsburg.de

[2] Computer-Aided Medical Procedures and Augmented Reality, Technical University of Munich, Boltzmannstr. 3, 85748 Garching bei München, Germany

Abstract. Dynamic positron emission tomography (PET) with tracer [^{18}F]FDG enables non-invasive quantification of glucose metabolism by means of kinetic analysis, often modelled by the two-tissue compartment model (TCKM). However, voxel-wise kinetic parameter estimation using conventional methods is computationally intensive and limited by spatial resolution. Deep neural networks (DNNs) offer an alternative but require large training datasets and significant computational resources. To address these limitations, we propose a physiological neural representation based on implicit neural representations (INRs) for personalized kinetic parameter estimation. INRs, which learn continuous functions, allow for efficient, high-resolution parametric imaging with reduced data requirements. Our method also integrates anatomical priors from a 3D CT foundation model to enhance robustness and precision in kinetic modelling. We evaluate our approach on an [^{18}F]FDG dynamic PET/CT dataset and compare it to state-of-the-art DNNs. Results demonstrate superior spatial resolution, lower mean-squared error, and improved anatomical consistency, particularly in tumour and highly vascularized regions. Our findings highlight the potential of INRs for personalized, data-efficient tracer kinetic modelling, enabling applications in tumour characterization, segmentation, and prognostic assessment. The code is available at: https://github.com/tkartikay/PhysNRPET.

Keywords: Implicit Neural Representations · Tracer Kinetic Modelling · Dynamic PET

1 Introduction

Dynamic positron emission tomography (PET) with [^{18}F]FDG has allowed for non-invasive assessment of glucose metabolism in tissues through kinetic analysis, in particular with the irreversible two-compartment-kinetic model (TCKM). The kinetic parameters thus obtained, K_1, k_2, and k_3 represent the rates of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 491–500, 2026.
https://doi.org/10.1007/978-3-032-04947-6_47

transport of the tracer between blood and cells (influx and efflux) and the rate of phosphorylation of glucose, respectively. V_B describes the blood volume fraction, to correct for the partial volume effects arising from the limitations of the scanner spatial resolutions [6]. Since glucose metabolism differs for metastatic lesions from non-metastatic tissues with increased uptake, kinetic parameters have been advantageous in successful diagnosis, complementing Standard Uptake Values (SUV) values from static PET images [22]. Phosphorylation rate (k_3) is significant metric for tumour differentiation in colorectal cancers [19] and breast cancers [8], and the perfusion parameter (K_1) can be a marker of angiogenesis in breast cancers [3]. Kinetic parameters, carrying more information than SUV, can play a vital role as additional biomarkers for tumour heterogeneity analysis, precise segmentations, and prognostic predictions.

Conventional methods based on non-linear curve fitting (CF) used to perform kinetic analysis on volumes of interest lack the spatial neighbourhood understanding and are slow when applied to derive voxel-based parametric images. These challenges can be overcome through the use of deep neural networks (DNNs) to derive the parametric images [4,7,9]. Self-supervised DNNs have recently been proposed to generate the K_1, k_2, k_3, V_B parametric images which satisfy the TCKM [5]. These methods, in turn, require large amount of training data to achieve accurate results, and while they are quick to predict the outputs once trained, the training of these networks itself is a slow and memory-intensive process, balanced by reducing the resolution of the training dataset, and slicing it into 2D+t data instead of processing the entire 3D+t data at once.

Implicit neural representations (INRs) are neural networks that learn continuous functions, which can parametrise signals, such as mapping continuously the image intensities from the input spatial coordinates. With steady adoptions in the medical imaging field due to their memory efficiency, improved reconstructions, interpolation capabilities, and personalisation [10], INRs could overcome the data limitation issues faced by DNNs and speed issues of CF in generation of parametric images in dynamic PET modelling. Since INRs can be optimised on single patient acquisitions, they can be adapted to individual patient anatomy and physiology [17], leading to personalised neural networks.

Encoding the dynamic physiological processes in the INR also allows their integration with downstream models that are trained for classification of malignancies, or early detection of metastases, etc., as well as with foundation models that extract features from the associated anatomical data such as CT scans, enabling feature-rich end-to-end pipelines for aiding the clinical routines.

Thus, guided by their promise, we propose the use of INRs for modelling dynamic PET to generate kinetic parameters (Fig. 1). Our contribution is:

- the introduction of a physiological neural representation for tracer kinetic modelling for a [^{18}F]FDG dynamic PET/CT dataset,
- its extension by integrating anatomical information and features extracted from corresponding CT scan using a 3D CT foundation model, and
- a comparative analysis against state-of-the-art DNNs and CF in terms of goodness of fit, memory and speed.

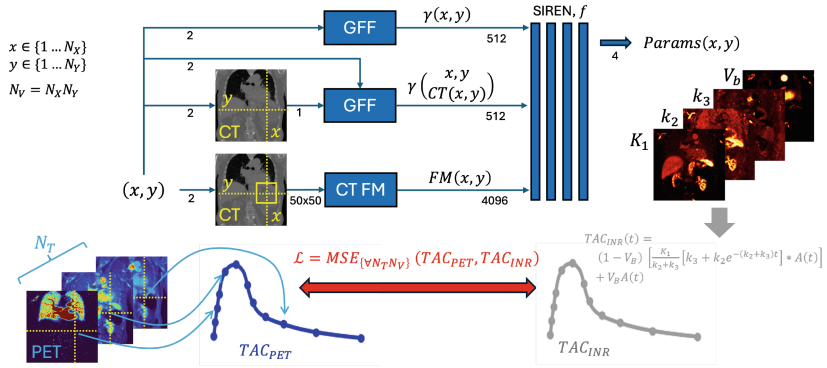


Fig. 1. Schematic overview of the proposed pipeline (illustrated in 2D for clarity; works in 3D spatial coordinates). Spatial positions are encoded using either positional coordinates alone, with local CT Hounsfield units, or with neighbourhood features from a CT foundation model. These features are mapped via a SIREN-based INR to estimate voxelwise kinetic parameters of the irreversible two-tissue compartment model. Predicted parameters generate modelled time-activity curves (TAC), which are optimised by minimising the MSE with measured PET curves

1.1 Related Works

Voxel-wise estimation of kinetic parameters have been explored by solving the non-linear optimisation problem of fitting the model-predicted curves onto the measured data [1,21]. The data in these studies was limited to smallest frames of 10 s resolution. With the advancements in long axial field of view scanners, and the availability of improved temporal resolution, the kinetic parameters have also been estimated for increased resolution of 2 s frames [14]. Sari et al. provide reference values for kinetic parameters for volumes of interest in oncological patients, and have also shown the voxel-wise generation of TCKM. Building upon this work, de Benetti el al. proposed DNNs for the generation of parametric images from the dynamic PET data, relying on self-supervised training of the DNN through the integration of the kinetic model into the loss function [5]. To parametrise tracer kinetics through encoding the PET reconstruction and one compartment model in neural networks, a hybrid physics- and data-driven approaches is proposed to incorporate the mathematical expressions defining the kinetics [23]. Shao et al. propose employing DNNs to discretise the data, and solving the ODEs that define the kinetic model, where the tracer concentration in each compartment is represented with separate fully connected DNNs, and the kinetic parameters are optimised while minimising the ODE loss [15].

Since INRs are based on the assumption of continuity as an *a priori* [10], these representations can be made robust to noise and further constrained anatomically through image-derived priors, such as extracted from large-scale "foundation" models. Large-scale models trained on high-resolution oncological data can provide knowledge of vascular architecture, tissue heterogeneity, and image-based biomarkers as embeddings or input features to the task-specific models.

Neural networks that encode the tracer kinetics could similarly benefit from the anatomical understanding of macroscopic blood supply routes, sharp gradients marking organ bounds, regions of tissue similarities and anomalies, etc. A foundation model trained on high-resolution 3D CT oncological datasets with understanding of underlying biology, and shown to improve performance of downstream tasks, particularly with limited dataset, and resilient to input variation has been presented by Pai et al. [11].

2 Materials and Methods

2.1 Dataset

The dataset consists of 24 patients with various tumour types undergoing dynamic PET imaging (average injected activity: 235 ± 51 MBq of [^{18}F]FDG), immediately followed by a CT scan using a Siemens Biograph Vision Quadra long-field-of-view PET/CT device. The PET acquisition lasted 65 min for all patients and consisted of 62 frames with varying durations: 2 s for the initial frames, 30 s after the first 2 min, and up to 5 min per frame toward the end of the acquisition. PET images were reconstructed with a voxel size of 1.65 × 1.65 × 1.65 mm^3 and subsequently filtered using a 2 mm full-width at half maximum (FWHM) Gaussian filter to generate **high-resolution images (HiRes)**. To match the resolution used by De Benetti et al., we downsampled the PET images to 2.5 × 2.5 × 2.5 ^{3}, producing **low-resolution images (LoRes)**. The CT scans were acquired in low-dose mode (voltage: 120 kV, tube current: 25 mA) and reconstructed with a voxel size matching that of the high-resolution PET images.

2.2 Preprocessing and Architecture

Dynamic PET data were first normalised to the range $[0, 1]$ by dividing by the maximum of the image-derived input function, which was on the order of $\sim$ 200 kBq/ml. The corresponding CT data were similarly mapped from $[-1024, 2048]$ to the interval $[0, 1]$. Registration of PET and CT was assumed valid and no motion correction between frames was applied.

Using the normalised CT volume, the CT foundation model [11] extracted 4096 features for each voxel, by taking as input 50 × 50 × 50 neighbourhood centred on that voxel. These features served as an optional input to the network, depending on the variant of the INR under consideration.

Spatial coordinates $x \in \mathrm{R}^3$ were first normalised to the interval $[-1, 1]$. Each normalised coordinate was encoded using Gaussian Fourier Features (GFF) [20] [12] drawn from a zero-mean Gaussian distribution, with a standard deviation chosen to control frequency bandwidth. A total of 256 frequencies were sampled, and a multiplier of 10 was applied to spread the frequency spectrum. Specifically, for each coordinate $\boldsymbol{x}$, the encoding was formed by computing: $\gamma(\boldsymbol{x}) = [sin(\mathbf{2\pi}\boldsymbol{B}\boldsymbol{x}), cos(\mathbf{2\pi}\boldsymbol{B}\boldsymbol{x})]^T$ where $\boldsymbol{B}$ is a randomly generated matrix of size $\mathbf{256 \times 3}$. The choice of 256 frequencies and a multiplier of 10 followed

typical recommendations in the literature and was not varied [20]. In all cases, the spatial domain (2D or 3D) was sampled, and the resulting coordinates were normalised to the interval $[-1, 1]$ prior to being processed with GFF. Where relevant (CTHU variant, see below), the voxel-wise H.U. value, also normalised to $[0, 1]$, was appended to the normalised coordinates to form an extended input vector. This augmented input was then encoded with GFF. However, in the variant utilising the foundation model features (CTFM variant, see below), the 4096-dimensional feature vector was incorporated directly, without GFF encoding, while GFF was still applied to the spatial coordinates.

The encoded features were passed to a SIREN [18] network with three hidden layers. Each hidden layer comprised 512 units and used sine activations, following the standard weight-initialisation scheme described in the SIREN work to preserve gradient stability. The final layer was a linear layer producing four outputs, the predicted kinetic parameters: K_1, k_2, k_3, and V_b. These outputs correspond to the parameters in the TCKM governed by a system of ordinary differential equations (ODEs). Maintaining differentiability from inputs to outputs facilitated direct gradient-based training in conjunction with the ODE formulation.

Parameter estimation was performed by minimising the mean-squared error (MSE) between the Time Activity Curve (TAC) generated from the predicted parameters and real TAC values measured at the input coordinate in the dynamic PET. Training used the Adam optimiser with a learning rate of $1e^{-5}$. The network was trained for 100 epochs and all hyperparameters, including the number of frequencies, the sine activation design, and the network depth, were adapted from SIREN and GFF examples without additional ablation experiments [2,13,16].

Using the estimated kinetic parameters from the INR for each voxel coordinate, the predicted TAC was generated by solving the TCKM ODEs Eq. 1.

$$\mathrm{TAC}_{\mathrm{INR}}(t) = (1 - V_b)\left[\frac{K_1}{k_2 + k_3}\left[k_3 + k_2 e^{-(k_2+k_3)t}\right] * A(t)\right] + V_b A(t) \quad (1)$$

To train the INR, the predicted TAC was then compared using MSE against the observed TAC at the discrete time points where measurements were available. The mean and standard deviation of these reconstruction errors were computed across the full 3D spatial domain with time (3D+t), selected 2D slices with time (2D+t), and at each individual voxel with time (voxel+t). The blood-to-tissue transport constant K_i was calculated using $K_i = K_1 k_3/(k_2 + k_3)$.

3 Results

We evaluated the proposed architecture in several variants: HiRes and LoRes, 2D (single slice) and 3D (a stack of 10 adjacent slices) - resulting in the variants inr-HiRes-2D, inr-LoRes-2D, inr-HiRes-3D and inr-LoRes-3D. Due to incomplete data of one patient, we compared inr-HiRes-2D to conventional CF in 23 patients (Table 1). In those 23 patients, we manually segmented liver, spleen, both lungs, and both kidneys for analysis of kinetic parameters (Fig. 2). Additionally, we

tested configurations incorporating the Hounsfield unit (HU) values from the CT scan (only for HiRes and in 2D - variant CTHU-2D) and those using features extracted with the foundation model (variant CTFM-2D), comparing them to the neural network proposed by De Benetti et al. in both 2D (single-slice) and 3D (stack of 10 adjacent slices) configurations (Table 2) [5].

As comparison metrics we used MSE with its standard deviation (SD), and training time, inference time, and memory utilization. All experiments were conducted on a MacBook with an M3 Pro (12-core CPU, Metal 3, 36 GB RAM).

Table 1. Quantitative evaluation of INR against CF: differences are statistically significant ($p \ll 0.05$)

Method	Mean MSE (± SD)	SD MSE	Runtime (± SD)
INR	**5.48E-03 ± 2.43E-03**	**1.13E-02**	**1336.13** ± 38.54 s
CF	1.16E-02 ± 8.23E-03	2.27E-02	40933.44 ± 1651.57 s

Table 2. Quantitative evaluation of INR variants and comparison with the spatio-temporal neural network of De Benetti et al. (FDB112-2D and FDB112-3D)

Model	Voxel Resolution	MSE(±SD)	Training Time	Inference Time	Memory Utilisation
FDB112-3D	2.5 mm	0.066 ± 0.255	–	62 s	–
INR-LoRes-3D	2.5 mm	0.009 ± 0.093	24.3 min	66 s	6.7 GB
FDB112-2D	2.5 mm	0.073 ± 0.268	–	0.15 s	–
INR-LoRes-2D	2.5 mm	0.009 ± 0.093	24.1 min	6 s	4.3 GB
INR-2D-HiRes	1.65 mm	0.009 ± 0.094	25.2 min	8 s	11.9 GB
CTHU-INR-2D-HiRes	1.65 mm	0.009 ± 0.094	25.4 min	8 s	12.4 GB
CTFM-INR-2D-HiRes	1.65 mm	0.009 ± 0.094	25.7 min	9 s	13 GB

To qualitatively visualize the impact of the different INR variants, we visualize the mean MSE and its SD (Fig. 3) for an exemplary patient. This gives an idea of which organs the measured and the calculated TACs depart in average (mean MSE), and its spread (SD) is. Additionally, Fig. 4 shows for a particular segment across the tumour when in time and space the MSE is highest (Fig. 4).

4 Discussion

This study applies implicit neural representations (INRs) to tracer kinetic modelling in dynamic [^{18}F]FDG PET. Unlike non-linear CF or large-scale DNNs,

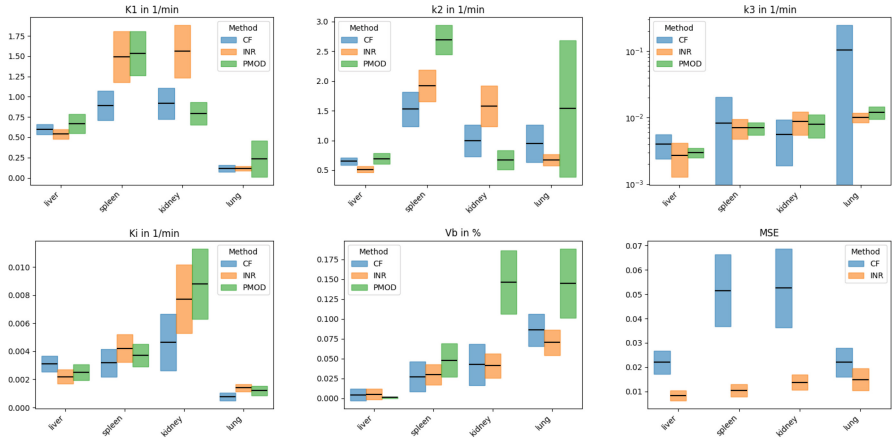


Fig. 2. Box plots (average ± SD for calculated parameters and MSE for four organs across 23 patients using INR and CF. Ranges obtained with PMOD as reported by Sari et al. [14] are plotted alongside for reference

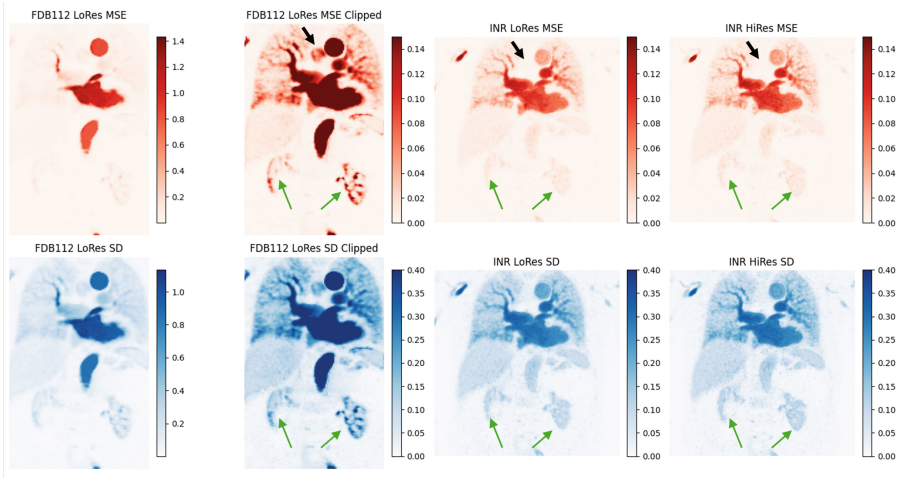


Fig. 3. MSE (top row, red) and SD (lower row, blue) for a coronal slice of an exemplary patient. MMSE is higher on average over the complete slice (see FDB112 LoRes MSE Clipped vs. INR LoRes MSE), particularly in the tumour (black arrow) and kidneys (green arrow), ignoring the blood pool (visible larger vessels). SD images show less variation on the MSE over time for our model. LoRes and HiRes variants of our model (right two columns) show good agreement with more details in the HiRes image. (Color figure online)

INRs approximate the continuous spatio-temporal function of patient-specific PET signals. Such a representation requires fewer training examples than conventional DNNs, making it useful when data availability is limited. When compared

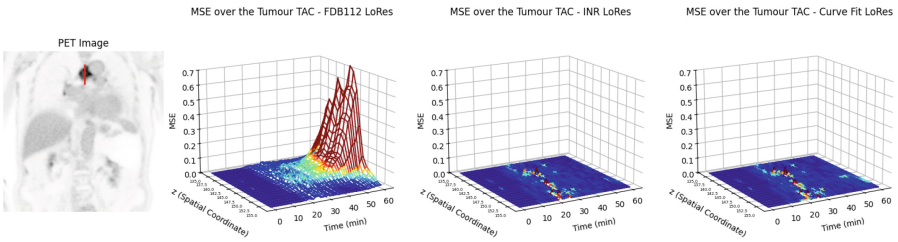


Fig. 4. MSE along a segment crossing a mediastinal lymph node metastasis (see PET slice, left). Plots show MSE as a function of space (left to bottom axis) and time (bottom to right axis). The left plot displays the 2D baseline; the centre plot shows the inr-LoRes-2D variant; the right plot shows the curve fit results. The baseline exhibits higher error in the lesion centre at late time points, whereas the INR yields lower MSE similar to CF, with errors instead aligning with the IDIF peak (not shown)

to CF, calculating dynamic parameters with INRs takes significantly shorter time and produces smoother images as spatial information is integrated regularizing the noise at a single voxel level. In addition, modelling kinetics through an INR allows more flexibility in handling anatomical and physiological variability across patients.

We integrated anatomical information from CT by incorporating HU values and, in a separate variant, 3D CT foundation model features. Both approaches provided comparable MSE but showed slightly faster convergence (not shown), suggesting that anatomical descriptors can guide the estimation of kinetic parameters. The proposed INRs consistently achieved lower MSE than a baseline DNN and CF, likely due to their ability to capture spatio-temporal dynamics via a continuous, data-efficient mapping.

Performance across INR variants was similar in accuracy, though convergence speed and memory demands differed. Foundation-model-based INRs converged faster but required more memory. Nonetheless, LoRes variants of our INR reduced the resource footprint, enabling training on standard hardware setups.

Comparisons with a baseline DNN and CF to a lesser extent proved that INRs show lower errors, especially in tumour and kidney regions, which often exhibit rapid or heterogeneous tracer kinetics. This highlights the robust parameter estimation of INRs when organ- or lesion-specific kinetic variability must be captured.

Accurate kinetic parameters (K_i, K_1, k_2, k_3, V_b) can reveal metabolic properties that aid in tumour assessment and therapy evaluation. The continuous form of INRs and their integration with anatomical descriptors can potentially reduce segmentation errors and support downstream tasks such as lesion classification. Further research could expand the scope by incorporating additional imaging modalities, testing on larger datasets, or refining network architectures and Fourier feature encodings. This future work may improve accuracy, promote generalisability, and further lower the hardware demands for advanced tracer kinetic modelling.

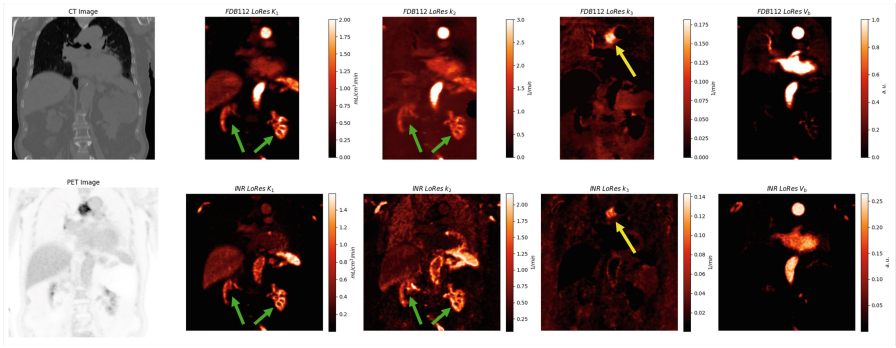


Fig. 5. Coronal slices of parametric images from the 2D baseline (upper row, second to last columns) and the inr-LoRes-2D variant. INR-derived maps exhibit greater detail and sharper edges. Boundaries of the kidneys (green arrow, prominent in K_1 and k_2) and tumour (yellow arrow, seen in k_3) are more clearly defined, highlighting tissue heterogeneity. The higher contrast in the tumour indicates greater specificity. (Color figure online)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Besson, F.L., et al.: 18 f-fdg pet and dce kinetic modeling and their correlations in primary nsclc: first voxel-wise correlative analysis of human simultaneous [18f] fdg pet-mri data. EJNMMI Res. **10**, 1–13 (2020)
2. Byra, M., Poon, C., Rachmadi, M.F., Schlachter, M., Skibbe, H.: Exploring the performance of implicit neural representations for brain image registration. Sci. Rep. **13**(1), 17334 (2023)
3. Cochet, A., et al.: Evaluation of breast tumor blood flow with dynamic first-pass 18f-fdg pet/ct: comparison with angiogenesis markers and prognostic factors. J. Nucl. Med. **53**(4), 512–520 (2012)
4. Cui, J., Gong, K., Guo, N., Kim, K., Liu, H., Li, Q.: Unsupervised pet logan parametric image estimation using conditional deep image prior. Med. Image Anal. **80**, 102519 (2022)
5. De Benetti, F., et al.: Self-supervised learning for physiologically-based pharmacokinetic modeling in dynamic pet. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 290–299. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-43907-0_28
6. Dimitrakopoulou-Strauss, A., Pan, L., Sachpekidis, C.: Kinetic modeling and parametric imaging with dynamic pet for oncological applications: general considerations, current clinical applications, and future perspectives. Eur. J. Nucl. Med. Mol. Imaging **48**, 21–39 (2021)
7. Huang, Z., et al.: Parametric image generation with the uexplorer total-body pet/ct system through deep learning. Eur. J. Nucl. Med. Mol. Imaging **49**(8), 2482–2492 (2022)

8. Kajáry, K., Lengyel, Z., Tőkés, A.M., Kulka, J., Dank, M., Tőkés, T.: Dynamic fdg-pet/ct in the initial staging of primary breast cancer: clinicopathological correlations. Pathol. Oncol. Res. **26**, 997–1006 (2020)
9. Liang, G., et al.: Combining deep learning with a kinetic model to predict dynamic pet images and generate parametric images. EJNMMI Phys. **10**(1), 67 (2023)
10. Molaei, A., et al.: Implicit neural representation in medical imaging: a comparative survey. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2381–2391 (2023)
11. Pai, S., et al.: Foundation model for cancer imaging biomarkers. Nat. Mach. Intell. **6**(3), 354–367 (2024)
12. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol. 20. Curran Associates, Inc. (2007)
13. Reed, A.W., et al.: Dynamic ct reconstruction from limited views with implicit neural representations and parametric motion fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2258–2268 (2021)
14. Sari, H., Mingels, C., et al.: First results on kinetic modelling and parametric imaging of dynamic 18 f-fdg datasets from a long axial fov pet scanner in oncological patients. Eur. J. Nucl. Med. Molec. Imaging 1–13 (2022)
15. Shao, W., Chen, Y., Li, N., Yang, Z., Meng, X., Xie, Z.: Fast and robust estimation of kinetic parameters in dynamic pet imaging using neural network-based discretization method (2023)
16. Shen, L., Pauly, J., Xing, L.: Nerp: implicit neural representation learning with prior embedding for sparsely sampled image reconstruction. IEEE Trans. Neural Netw. Learn. Syst. **35**(1), 770–782 (2022)
17. Shi, J., Zhu, J., Pelt, D.M., Batenburg, K.J., Blaschko, M.B.: Implicit neural representations for robust joint sparse-view ct reconstruction. arXiv preprint arXiv:2405.02509 (2024)
18. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Adv. Neural. Inf. Process. Syst. **33**, 7462–7473 (2020)
19. Strauss, L.G., et al.: Assessment of quantitative FDG pet data in primary colorectal tumours: which parameters are important with respect to tumour detection? Eur. J. Nucl. Med. Mol. Imaging **34**, 868–877 (2007)
20. Tancik, M., et al.: Fourier features let networks learn high frequency functions in low dimensional domains. Adv. Neural. Inf. Process. Syst. **33**, 7537–7547 (2020)
21. Wang, G., et al.: Total-body pet multiparametric imaging of cancer using a voxel-wise strategy of compartmental modeling. J. Nucl. Med. **63**(8), 1274–1281 (2022)
22. Wumener, X., et al.: Dynamic fdg-pet imaging for differentiating metastatic from non-metastatic lymph nodes of lung cancer. Front. Oncol. **12**, 1005924 (2022)
23. Ye, Y., Liu, H., Wang, L.: Hybrid kinetics embedding framework for dynamic pet reconstruction. arXiv preprint arXiv:2403.07364 (2024)

Q-Space Guided Collaborative Attention Translation Network for Flexible Diffusion-Weighted Images Synthesis

Pengli Zhu[1], Yingji Fu[1], Nanguang Chen[3], and Anqi Qiu[1,2,4(✉)]

[1] Department of Health Technology and Informatics, Hong Kong Polytechnic University, Hong Kong, China
[2] Mental Health Research Center, The Hong Kong Polytechnic University, Hong Kong, China
[3] Department of Biomedical Engineering, National University of Singapore, Singapore, Singapore
[4] Department of Biomedical Engineering, The Johns Hopkins University, Baltimore, USA
an-qi.qiu@polyu.edu.hk

Abstract. This study, we propose a novel *Q*-space Guided Collaborative Attention Translation Networks (**Q-CATN**) for multi-shell, high-angular resolution DWI (MS-HARDI) synthesis from *flexible q*-space sampling, leveraging the commonly acquired structural MRI data. Q-CATN employs a *collaborative attention mechanism* to effectively extract complementary information from multiple modalities and *dynamically* adjust its internal representations based on flexible *q*-space information, eliminating the need for fixed sampling schemes. Additionally, we introduce a range of task-specific constraints to preserve *anatomical fidelity* in DWI, enabling Q-CATN to accurately learn the intrinsic relationships between directional DWI signal distributions and *q*-space. Extensive experiments on the Human Connectome Project (HCP) dataset demonstrate that Q-CATN outperforms existing methods, including 1D-qDL, 2D-qDL, MESC-SD, and QGAN, in estimating parameter maps and fiber tracts both quantitatively and qualitatively, while preserving fine-grained details. Notably, its ability to accommodate *flexible q -space sampling* highlights its potential as a promising toolkit for clinical and research applications. Our code is available at https://github.com/Idea89560041/Q-CATN.

Keywords: DWI Synthesis · Conditional Generative Model · Collaborative Attention Translation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-032-04947-6_48.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 501–511, 2026.
https://doi.org/10.1007/978-3-032-04947-6_48

1 Introduction

Diffusion-weighted imaging (DWI) is a key non-invasive method for evaluating brain microstructure and connectivity, providing critical insights into development, aging, and neurodegenerative diseases. Advanced models like neurite orientation dispersion and density imaging (NODDI) [24] and diffusion kurtosis imaging (DKI) [10] offer superior tissue microstructure analysis over conventional diffusion tensor imaging (DTI). However, these methods require extensive q-space sampling, longer acquisition times, and complex computational processes, increasing susceptibility to motion artifacts, eddy current distortions, and physiological noise, which can affect their quantitative precision.

Recent advances in deep learning have shown promise in medical image synthesis, with several studies [1,13,20] estimating parameter maps from DWI using limited gradient directions. Q-space deep learning (qDL) [6] pioneered the direct mapping of sparsely sampled q-space DWI signals to microstructural parameters via a multilayer perceptron. Subsequent enhancements integrated 2D spatial information [5] and 3D sparse spatial patch representations with modified LSTM networks [21]. However, these methods are limited to generating fixed parameter maps, restricting their applicability to variably sampled DWI data. Thus, a more flexible DWI synthesis approach is needed to broaden its practical utility.

According to the principles of DWI, its generation involves a complex nonlinear relationship in q-space [12], making conditional generative adversarial networks (cGANs) [9,11] well-suited for DWI synthesis. Recent studies have shown their effectiveness in producing high-fidelity medical images, such as translating structural/functional MRI into DWI [16] and generating DWI-derived scalar maps [7]. Complementary modalities like T1- and T2-weighted MRI have also been shown to improve DWI synthesis [2,14]. However, existing methods are typically optimized for fixed q-space sampling aligned with their training datasets, limiting their applicability in clinical settings where heterogeneous sampling is common and site-specific data is often insufficient for training models. This highlights the need for a flexible DWI synthesis approach not constrained by predefined sampling strategies. While q-space cGANs such as Q-GAN [14] and aqDL [26] offer promising solutions by enabling DWI generation at arbitrary q-space points, they often rely on oversimplified input representations or fail to effectively capture inter-modality correlations, limiting their ability to fully utilize complementary information and constraining their generative performance.

To overcome the above limitations, we propose a novel q-space guided collaborative attention translation networks (Q-CATN), for multi-shell, high-angular resolution DWI (MS-HARDI) synthesis with flexible q-space sampling using commonly acquired structural MRI. The main features are outlined as follows:

1. The proposed Q-CATN framework supports MS-HARDI synthesis by incorporating *flexible* q-space conditional information, overcoming the limitations of fixed sampling strategies commonly seen in existing approaches;
2. By introducing a *collaborative attention mechanism*, Q-CATN effectively extracts compatible information from single modality and complementary

information across multi-modal inputs (e.g., b0, T1- and T2-weighted images), enhancing synthesis accuracy and robustness;
3. The proposed framework enables the generation of *densely* sampled q-space data, facilitating the reconstruction of various diffusion models, which significantly benefit downstream applications.

2 Methodology

2.1 Overall Architecture

Figure 1 presents the Q-CATN framework, comprising a single-modal attention (SMA) encoder, a multi-modal attention fusion (MMAF) module, a q-space embedding module, a SMA decoder, and a conditional discriminator. Q-CATN takes $\mathbf{x}_{b0}$, $\mathbf{x}_{t1}$, and $\mathbf{x}_{t2}$ as input, representing b0, T1-, and T2-weighted images, respectively. These are processed by SMA encoders to extract latent-space features $(\mathbf{z}_{b0}, \mathbf{z}_{t1}, \mathbf{z}_{t2})$. The MMAF module integrates these features into unified representation $\mathbf{z}$, conditioned on q-space coordinates $\vec{\boldsymbol{q}} = (g_x, g_y, g_z, b)$, which define the b-vector and b-value. The q-space embedding module plays a crucial role in transforming $\mathbf{z}$ into $\mathbf{z}_{bn}$ using $\vec{\boldsymbol{q}}$, facilitating flexible q-space-aware representation learning. The SMA decoder reconstructs $\mathbf{z}_{bn}$ into the image domain as $\mathbf{x}_{bn}$ under ground-truth supervision. Meanwhile, a discriminator is trained to distinguish synthesized from real DWIs. Further details are provided below.

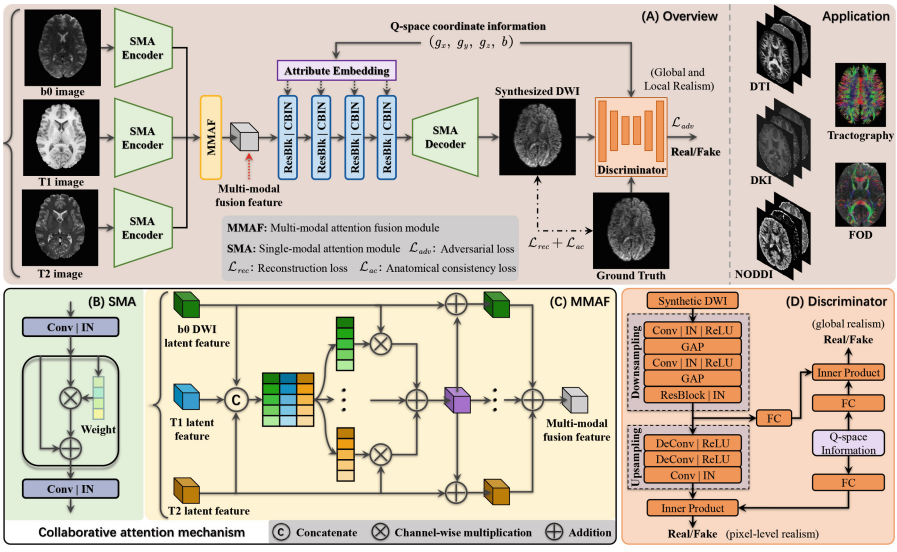


Fig. 1. Overview of Q-CATN. Panel (A) illustrates the high-level structure of the model. Panel (B) details the architecture of the single-modal attention module. Panel (C) shows the multi-modal attention fusion mechanism. Panel (D) outlines the structure of the conditional discriminator.

2.2 Collaborative Attention Mechanism

To enable the generation of precise and realistic DWI outputs, the proposed Q-CATN model employs a collaborative attention mechanism to effectively extract and integrate compatible-complementary information from multiple modalities. This architecture comprises two key components:

SMA Encoder. To improve multi-modal MRI synthesis, it is essential for each modality to provide *sufficient and complementary information*. We introduce a SMA encoder to extract more compatible features $\mathbf{z}_n$ across three encoder branches, computed as: $\mathbf{z}_n = G_n^{enc}(\mathbf{x}_n) + G_n^{enc}(\mathbf{x}_n) \otimes \delta\left(G_n^{enc}(\mathbf{x}_n)\right)$, where $G_n^{enc}(\cdot)$ denotes feature maps from the n-th SMA encoder, $\otimes$ represents channel-wise multiplication between feature maps and vector, and $\delta(\cdot)$ is an attention metric for DWI modality information. The SMA mechanism assigns a weight vector to each modality, emphasizing specific channels. These weighted features are combined with the input to yield an enhanced single-modal representation $\mathbf{z}_n$: $\{\mathbf{z}_{b0}, \mathbf{z}_{t1}, \mathbf{z}_{t2}\}$. Here, $G_n^{enc}(\cdot)$ employs strided convolutions for downsampling, while $\delta(\cdot)$ is implemented via global average pooling followed by two fully connected (FC) layers with ReLU and Sigmoid activations, respectively.

MMAF Module. To effectively integrate *DWI-specific information* from multiple modalities within the collaborative attention mechanism, we introduce the MMAF module, as depicted in Fig. 1(C). Specifically, the input modality features are concatenated and processed through FC layers to compute the attention matrix $\mathcal{A}$. This matrix is then used to weight the features via channel-wise multiplication. The resulting weighted features are combined to form a multi-modal feature, which is added to each input feature to produce modality-specific features $\mathbf{z}_n^{\text{att}}$. The multi-modal attention determines the weights $\mathcal{A}$ for the cross-modal feature channels as follows: $\mathcal{A} = G^m\left(\psi\left(\eta\left(\left(\mathbf{z}_{b0}, \mathbf{z}_{t1}, \mathbf{z}_{t2}\right)\omega_1\right)\omega_2\right)\right)$, where ω_1 and ω_2 are mapping matrices implemented by two FC layers, while η and ψ denote ReLU and Sigmoid activations, respectively. The function G^m applies softmax across each row of $\mathcal{A}$, ensuring each row represents modality weights for a specific pattern, and each column represents a modality's weights across all patterns. The matrix $\mathcal{A}$ is then used to extract correlated information from the multi-modal data, yielding the complementary feature $\mathbf{z}_n^{\text{att}}$ for the n-th modality: $\mathbf{z}_n^{\text{att}} = \mathbf{z}_n + \sum_{n=1}^{3}\left(\mathbf{z}_n \otimes \mathcal{A}_n\right)$, where $\mathcal{A}_n$ denotes the n-th column of $\mathcal{A}$. The final fused feature $\mathbf{z}$ is obtained by combining these modality-specific features.

2.3 *Q*-Space Embedding Module

To address the limitations imposed by *predefined sampling strategies*, we propose a *flexible* DWI synthesis approach enabling the generation of $\mathbf{z}_{bn}$ for any given variable $\vec{\boldsymbol{q}}$. Inspired by affine transformation parameters in normalization layers to encode attributes, we introduce central biasing instance normalization (CBIN) [23] into the residual blocks, which are dynamically modulated by the q-space

coordinates. The operation is defined as follows: $\mathrm{CBIN}(\mathbf{z}, \hat{\boldsymbol{q}}) = \frac{\mathbf{z}-\mu(\mathbf{z})}{\sigma(\mathbf{z})} + b_r(\hat{\boldsymbol{q}})$, where $\mathbf{z}$ denotes the feature map from the previous convolution, $\hat{\boldsymbol{q}}$ is the q-space embedding code, μ and σ are the instance mean and standard deviation, and b_r is the bias for the r-th feature map. Finally, the SMA decoder is used to precisely reconstruct $\mathbf{z}_{bn}$ back to the direction-specific image $\mathbf{x}_{bn}$.

2.4 *Q*-Space Conditional Discriminator

To improve the *realism* of the synthesized DWI, we employ a q-space conditional discriminator with two levels: its bottleneck layer assesses global image realism, while the output layer evaluates pixel-level fidelity, as illustrated in Fig. 1(D). Specifically, the discriminator processes $\mathbf{x}_{bn}$, extracting a global representation via its encoding path to evaluate global realism. Meanwhile, a decoder expands the output to match the input size, enabling per-pixel realism feedback. The q-space coordinates are integrated via conditional projection before the final layers of both branches. The final layer is defined as $f(\mathbf{x}_d, \vec{\boldsymbol{q}}) := (\vec{\boldsymbol{q}})^T V \gamma(\mathbf{x}_d) + \xi(\gamma(\mathbf{x}_d))$, where V denotes a learnable embedding of $\vec{\boldsymbol{q}}$, $\gamma(\mathbf{x}_d)$ represents the output prior to conditioning, and $\xi(\cdot)$ is a scalar function applied to $\gamma(\mathbf{x}_d)$.

2.5 Loss Function

Adversarial Loss. To enhance the realism of the synthetic DWI both *locally and globally*, we introduce adversarial losses for the encoder and decoder to achieve superior outcomes. The adversarial loss $\mathcal{L}_{adv}^{(*)}$ is defined as:

$$\mathcal{L}_{adv}^{(*)} = \mathbb{E}\left[\log\left(1 - D_{(*)}(\mathbf{y}_{bn}, \vec{\boldsymbol{q}})\right)\right] + \mathbb{E}\left[\log D_{(*)}(\mathbf{x}_{bn}, \vec{\boldsymbol{q}})\right], \tag{1}$$

where $(*)$ denotes either the encoder and decoder of the discriminator, respectively. $\mathbf{x}_{bn}$ is the synthesized DWI conditioning on the variable $\vec{\boldsymbol{q}}$, and $\mathbf{y}_{bn}$ represents the real tuple with $\vec{\boldsymbol{q}}$ sampled from the training data.

Reconstruction Loss. To better align low-frequency details and ensure consistency with the input, we further introduce a reconstruction loss $\mathcal{L}_{rec}$ as below:

$$\mathcal{L}_{rec} = \begin{cases} \mathbb{E}\left[\|\mathbf{x}_{bn} - \mathbf{x}_{bn}^*\|_1\right], & \text{if } b > 0 \\ \mathbb{E}\left[\|\mathbf{x}_{bn} - \mathbf{x}_{b0}\|_1\right], & \text{if } b = 0 \end{cases} \tag{2}$$

where $\mathbf{x}_{bn}^*$ is the reference DWI with targeted q-space coordinates $\vec{\boldsymbol{q}}$.

Anatomical Consistency Loss. To ensure the synthesized DWI accurately reflects the *underlying tissue microstructure*, we employ a spatially-correlative loss [25] to preserve image structure. The anatomical features of DWI are extracted using a VGG16 network, and self-similarity is computed as a map: $\mathcal{L}_{ac} = \|G_{ac}(\mathbf{x}_{bn}) - G_{ac}(\mathbf{x}_{bn}^*)\|_1$, where $G_{ac}(x_i) = (f_{x_i})^T(f_{x_*})$, $f_{x_i}^T$ represents the feature of a query point x_i, f_{x_*} denotes the features associated with patch, while $G_{ac}(x_i)$ captures the anatomical correlation between different query points.

Overall Loss Function. To summarize, the overall loss function for Q-CATN can be expressed as: $\mathcal{L} = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{ac}\mathcal{L}_{ac}$, where λ_{rec} and λ_{ac} are loss weights used to adjust the relative importance of each term.

2.6 Implementation

Dataset. We conducted experiments using preprocessed data from the Human Connectome Project (HCP) release[1]. Skull stripping was performed using brain masks, and T1- and T2-weighted images were resampled to b0 resolution, followed by image registration with DPABI [19]. The training set included 20 subjects (5400 DWIs), while 30 subjects were reserved for testing. Training data comprised 2D axial slices sampled from random directions, with DWI intensities normalized using the corresponding b0 and b-values scaled by their maximum.

Training and Inference. Q-CATN is trained using b0, T1, and T2 images as inputs, with diverse q-space coordinates integrated into latent features. The model generates the target DWI, while a discriminator enhances realism by distinguishing synthetic from real DWI. Optimization is performed using loss functions until convergence. During the inference, b0, T1-, T2-weighted images, and *flexible q-space conditions* are used to synthesize MS-HARDI.

Optimization. We implemented Q-CATN using Python 3.8 and PyTorch 1.13.0. During the training, we set the loss weights λ_{rec} and λ_{ac} to 100 and updated the discriminator every two generator updates. The model was trained on four NVIDIA A100 GPUs using the ADAM optimizer with a mini-batch size of 128. The initial learning rates for the generator and discriminator were set to 1×10^{-4} and 5×10^{-5}, respectively, and reduced by a factor of 0.95 after each epoch. The training process was conducted for a maximum of 300 epochs.

3 Experimental Results

To evaluate the performance of Q-CATN, we conducted comparative experiments with benchmark models. Specifically, 1D-qDL [6], 2D-qDL [5], and MESC-SD [21] were limited to generating specific parameter maps from undersampled DWI data (using 30 fixed-direction data by default), while QGAN [14] and Q-CATN employed the `sphere2cart` function from DIPY [4] to *flexibly simulate dense q-space sampling* (270 directional data) for downstream evaluation. Ground truth in all experiments was derived from the complete testing subject.

3.1 Qualitative Analysis

We conducted DWI synthesis experiments across varying b-values, as illustrated in Fig. 2. From the zoom-in perspective, Q-CATN preserves fine anatomical

[1] https://www.humanconnectome.org/.

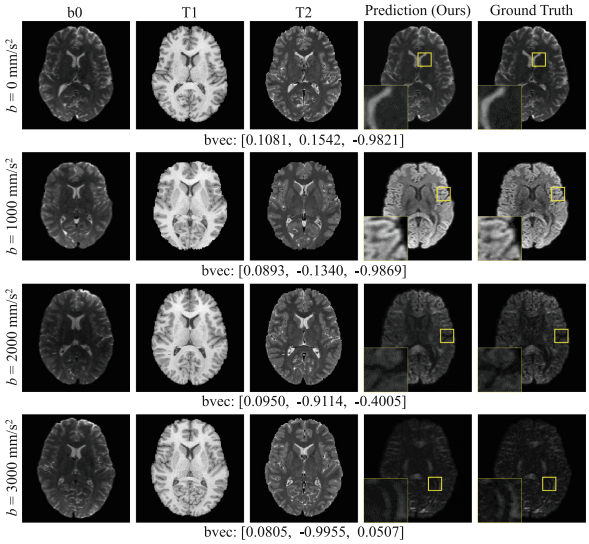


Fig. 2. DWI synthesis results under different b-values configurations. On the left, all potential input channels (b0, T1, T2) are shown, while the right side displays the predicted results, with a standard DWI slice serving as the reference.

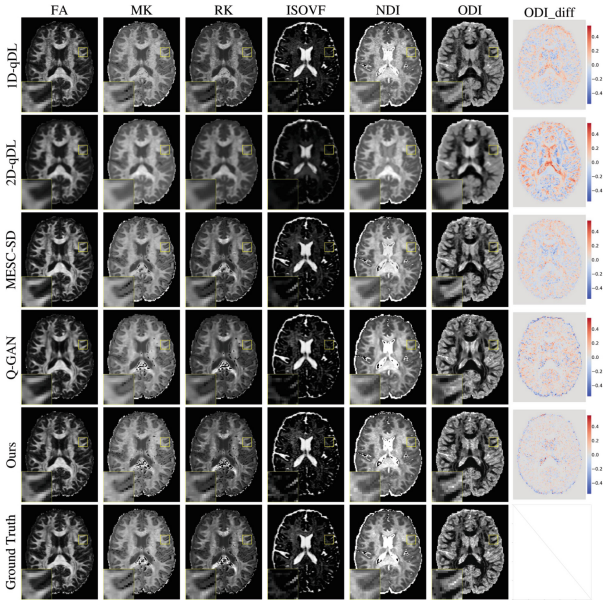


Fig. 3. Qualitative comparison of different methods for undersampling DWI parameter fitting. The first six columns show the various diffusion maps and the last column is the ODI error map, each row represents different comparison methods, with the reference map at the bottom.

structures, yielding results consistent with the ground-truth. We further evaluated against comparison methods for estimating parameter maps, performing diffusion imaging using DIPY [4] for DTI (using only $b = 1000\,s/mm^2$) and DKI, while fitting the NODDI model using AMICO [3]. As shown in Fig. 3, 1D-qDL fails to yield feasible results due to its reliance on predefined downsampling scheme and 2D-qDL exhibited notable inaccuracies in structural detail prediction. Although MESC-SD and Q-GAN demonstrated improved performance, anatomical details remained unclear upon closer inspection. In contrast, Q-CATN generated dense DWIs, producing parameter maps that closely align with the ground truth, surpassing all other methods.

3.2 Quantitative Analysis

To quantitatively assess the effectiveness in parameter fitting, we utilized several image quality metrics, including root mean square error (RMSE), multi-scale structural similarity index (MS-SSIM), peak signal-to-noise ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS). Except for RMSE and LPIPS, higher values denote better performance. Figure 4 summarizes the quantitative comparison of estimated diffusion maps, Q-CATN consistently surpasses all baseline models across four metrics, demonstrating its superior ability to adapt to proposed collaborative attention and flexible q-space sampling schemes.

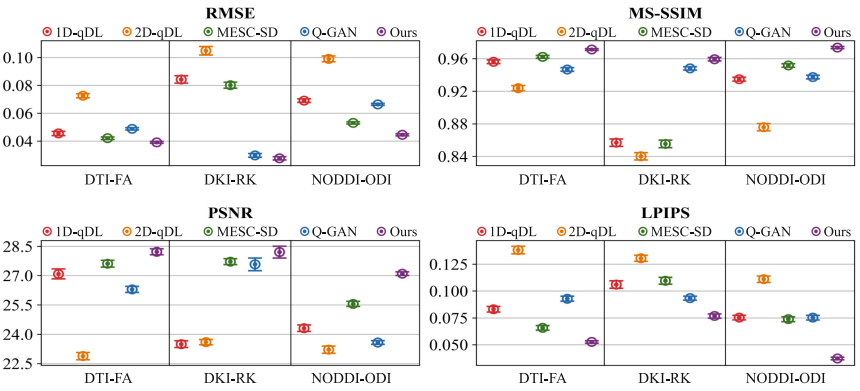


Fig. 4. Quantitative comparison of estimated parameter maps in different methods.

3.3 FOD and Tractography

To evaluate downstream performance beyond parameter fitting, we performed fiber orientation distribution (FOD) using MRtrix3 [17] and FSL [15], alongside

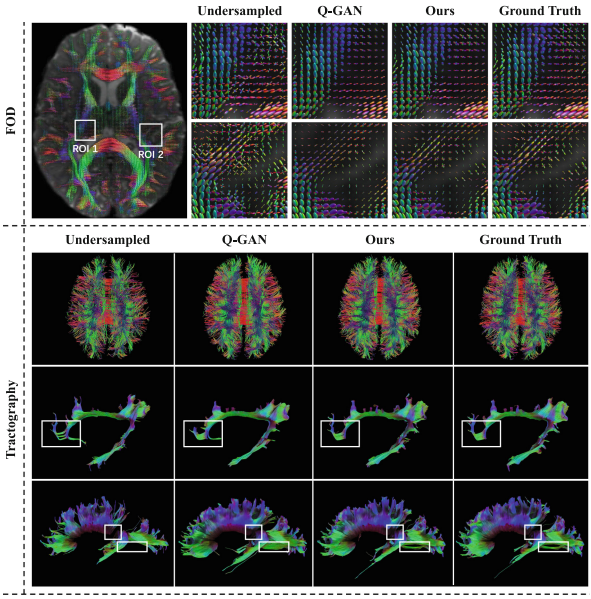


Fig. 5. Qualitative comparisons among different methods in fiber orientation distribution (FOD) and tractography. The top shows two magnified ROIs of FOD, and the bottom shows the tractography and two specific fiber tracts, i.e., Cingulum (L) and CorpusCallosum.

tractography using DSI-Studio [22]. Figure 5 shows that the fiber tracts generated from undersampled data exhibit the poorest quality, whereas those produced by Q-GAN better preserve continuity and integrity, and our results align more closely with the ground truth. These findings demonstrate that Q-CATN reliably and accurately represents underlying fiber structures in downstream analyses.

4 Conclusion

This study introduces a novel q-space guided collaborative attention translation network (Q-CATN) for MS-HARDI synthesis from flexible q-space sampling. Leveraging a collaborative attention mechanism, Q-CATN extracts complementary information from multiple modalities and modulates internal representations with flexible q-space conditions, addressing the limitations of fixed sampling strategies. Extensive experiments show that Q-CATN surpasses existing methods in synthesizing parameter maps and fiber tracts with fine-grained details, highlighting its potential for advancing diffusion applications. Notably, while recent diffusion models [8] have shown impressive capabilities in generating images, Q-CATN was chosen for its superior computational efficiency and faster inference times, crucial for real-time clinical applications.

Acknowledgements. This research/project is supported by the STI 2030 - Major Project (No. 2022ZD0209000), the National Research Foundation, Singapore, and the Agency for Science Technology and Research (A*STAR), Singapore, under its Prenatal/Early Childhood Grant (Grant No. H22P0M0007). Additional support is provided by the University Grants Committee (GRF: 15201124), and the Hong Kong Global STEM Scholar scheme.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, G., Hong, Y., Huynh, K.M., Yap, P.T.: Deep learning prediction of diffusion mri data with microstructure-sensitive loss functions. Med. Image Anal. **85**, 102742 (2023)
2. Cicimen, A.G., et al.: Image quality transfer of diffusion mri guided by high-resolution structural mri. arXiv preprint arXiv:2408.03216 (2024)
3. Daducci, A., Canales-Rodríguez, E.J., Zhang, H., Dyrby, T.B., Alexander, D.C., Thiran, J.P.: Accelerated microstructure imaging via convex optimization (amico) from diffusion mri data. Neuroimage **105**, 32–44 (2015)
4. Garyfallidis, E., et al.: Dipy, a library for the analysis of diffusion mri data. Front. Neuroinform. **8**, 8 (2014)
5. Gibbons, E.K., et al.: Simultaneous noddi and gfa parameter map generation from subsampled q-space imaging using deep learning. Magn. Reson. Med. **81**(4), 2399–2411 (2019)
6. Golkov, V., et al.: Q-space deep learning: twelve-fold shorter and model-free diffusion mri scans. IEEE Trans. Med. Imaging **35**(5), 1344–1351 (2016)
7. Gu, X., Knutsson, H., Nilsson, M., Eklund, A.: Generating diffusion MRI scalar maps from T1 weighted images using generative adversarial networks. In: Felsberg, M., Forssén, P.-E., Sintorn, I.-M., Unger, J. (eds.) SCIA 2019. LNCS, vol. 11482, pp. 489–498. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20205-7_40
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
10. Jensen, J.H., Helpern, J.A., Ramani, A., Lu, H., Kaczynski, K.: Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Magn. Reson. Med. Off. J. Int. Soc. Magn. Reson. Med. **53**(6), 1432–1440 (2005)
11. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
12. Mori, S.: Introduction to Diffusion Tensor Imaging. Elsevier, Amsterdam (2007)
13. Park, J., et al.: Diffnet: diffusion parameter mapping network generalized for input diffusion gradient schemes and b-value. IEEE Trans. Med. Imaging **41**(2), 491–499 (2021)

14. Ren, M., Kim, H., Dey, N., Gerig, G.: *Q*-space conditioned translation networks for directional synthesis of diffusion weighted images from multi-modal structural MRI. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12907, pp. 530–540. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87234-2_50
15. Smith, S.M., et al.: Advances in functional and structural mr image analysis and implementation as fsl. Neuroimage **23**, S208–S219 (2004)
16. Tian, Q., et al.: Deepdti: high-fidelity six-direction diffusion tensor imaging using deep learning. Neuroimage **219**, 117017 (2020)
17. Tournier, J.D., et al.: Mrtrix3: a fast, flexible and open software framework for medical image processing and visualisation. Neuroimage **202**, 116137 (2019)
18. Van Essen, D.C., et al.: The wu-minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
19. Yan, C.G., Wang, X.D., Zuo, X.N., Zang, Y.F.: Dpabi: data processing & analysis for (resting-state) brain imaging. Neuroinformatics **14**, 339–351 (2016)
20. Yang, J., et al.: Towards accurate microstructure estimation via 3d hybrid graph transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 25–34. Springer, Heidelberg (2023). https://doi.org/10.1007/978-3-031-43993-3_3
21. Ye, C., Li, Y., Zeng, X.: An improved deep network for tissue microstructure estimation with uncertainty quantification. Med. Image Anal. **61**, 101650 (2020)
22. Yeh, F.C., Verstynen, T.D., Wang, Y., Fernández-Miranda, J.C., Tseng, W.Y.I.: Deterministic diffusion fiber tracking improved by quantitative anisotropy. PLoS ONE **8**(11), e80713 (2013)
23. Yu, X., Ying, Z., Li, T., Liu, S., Li, G.: Multi-mapping image-to-image translation with central biasing normalization. arXiv preprint arXiv:1806.10050 (2018)
24. Zhang, H., Schneider, T., Wheeler-Kingshott, C.A., Alexander, D.C.: Noddi: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage **61**(4), 1000–1016 (2012)
25. Zheng, C., Cham, T.J., Cai, J.: The spatially-correlative loss for various image translation tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16407–16417 (2021)
26. Zong, F., et al.: Attention-based q-space deep learning generalized for accelerated diffusion magnetic resonance imaging. IEEE J. Biomed. Health Inf. (2024)

RDMR: Recursive Inference and Representation Disentanglement for Multimodal Large Deformation Registration

Yibo Hu[1], Ziqi Zhao[1], Qi Zhang[1], Lisa X. Xu[1,2], and Jianqi Sun[1,2(✉)]

[1] School of Biomedical Engineering, Shanghai JiaoTong University, Shanghai, China
[2] National Engineering Research Center of Advanced Magnetic Resonance Technologies for Diagnosis and Therapy (NERC-AMRT), Shanghai, China
milesun@sjtu.edu.cn

Abstract. Multimodal large deformation image registration is a challenging task in medical imaging, primarily due to significant modality differences and large tissue deformations. Current methods typically employ dual-branch multiscale pyramid registration networks. However, the dual-branch structure fails to explicitly enforce that the model learns modality-invariant image registration features. Furthermore, in the multiscale registration process, only the deformation field is propagated, which restricts the model's capacity to accommodate more complex deformations. To enhance the model's ability to learn features from different modalities, we propose a modality representation disentanglement method, incorporating Multi-layer Contrastive Loss(MCL) to enforce the learning of modality-invariant features. To address the challenge of complex large deformations, we introduce a Multi-Scale Feature fusion Registration module(MSFR), which integrates features and deformation fields from different scales during the registration process. To explore the registration potential of the trained model, we propose a Recursive Inference enhancement strategy that further improves registration performance. This model is referred to as RDMR. Based on experimental results from both private and public datasets, the RDMR model outperforms other SOTA models. Compared to the baseline registration model (Voxel Morph), the RDMR model achieved improvements of 1.4 and 4.5% points in the DSC metric, respectively. Our code is publicly available at:https://github.com/ybby2020/RDMR.

Keywords: Multimodal deformable registration · Modality-invariant contrastive · Recursive inference

1 Introduction

Aligning (registration) anatomical structures across multimodal medical images is a fundamental task in medical image analysis. This task has significant implications for clinical applications, including preoperative multimodal diagnosis,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 512–522, 2026.
https://doi.org/10.1007/978-3-032-04947-6_49

intraoperative image-guided surgical planning, and postoperative efficacy evaluation [1,2]. However, multimodal medical image registration faces significant challenges due to the inherent differences in imaging mechanisms across modalities. For example, MRI provides high-contrast images of soft tissues, whereas CT offers high-resolution images of bone structures. These modality-specific characteristics lead to significant discrepancies in intensity distribution profiles across imaging modalities. Furthermore, additional physiological factors [3] such as posture changes, respiratory-induced motion introduce complex spatial variations that frequently result in nonlinear, large-magnitude tissue deformations.

In recent years, deep learning-based deformable registration methods have achieved remarkable progress [4–10]. However, current studies predominantly focus on unimodal registration [12], with insufficient exploration of multimodal registration under large deformations. Current unimodal registration research addressing large deformations predominantly adopts dual-stream pyramid registration methods [13] [14] [15]. However, existing approaches exhibit some limitations that hinder their performance in multimodal large deformation registration tasks: (1) during the multiscale registration process, only the deformation field was transmitted, lacking interaction of image features, which leads to a decrease in registration accuracy in complex scenarios; (2) the current dual-branch design cannot ensure that the model learns the intrinsic, modality-invariant features of the image.

To address the challenges of large deformations and intensity discrepancies in multimodal deformable registration. We propose a novel Disentanglement-based Multimodal Registration model (DMR) and introduce an innovative recursive inference enhancement strategy to further improve model performance (RDMR). To the best of our knowledge, this is the first work to propose using recursive inference to enhance performance in multiscale multimodal registration. Extensive experiments on multi-center clinical datasets demonstrate that our model outperforms state-of-the-art(SOTA) approaches in both qualitative and quantitative evaluations. Specifically, our contributions can be summarized as follows:

(i) *Multi-scale Feature Fusion Registration:* We design a multiscale feature interaction module that fuses multiscale features(MSFR), breaking the limitations of conventional multiscale deformation field propagation and significantly enhancing the model's capability to perceive complex deformations.
(ii) *Disentanglement based on contrastive learning:* We propose a novel multilayer modality-invariant contrastive loss(MCLoss) to enforce feature distribution consistency across modalities at different encoder depths, thereby improving registration performance and model generalization.
(iii) *Low Computation Load Enhancement Mechanism:* We introduce a recursive inference strategy that boosts performance without retraining the network. By recursively adjusting the multi-scale modules of the trained model during inference, we can optimize its final structure with minimal computational overhead, further enhancing performance.

2 Method

The architecture of the proposed multimodal deformable registration model is shown in Fig. 1. During the training phase, multimodal data are fed into the dual-branch feature extraction module to obtain image features at different scales. Registration starts at the smallest scale, where features from different modalities at the same scale are input into the MSFR (Multi-Scale Feature Registration) module, and the process continues until the final deformation field is obtained. During the inference phase, further improvement of the registration results is achieved by recursively applying the MSFR module at each scale. The number of recursions is associated with parameters such as a, b, c, and d in Fig. 1.

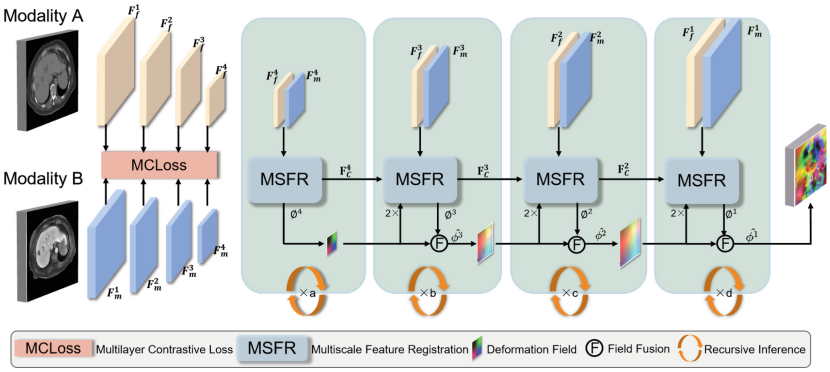


Fig. 1. Overview of the proposed RDMR architecture

2.1 Dual-Stream Encoding Branch

This module is used for image feature extraction, with the goal of extracting modality-invariant features to facilitate subsequent multiscale registration. We use convolutional neural network as the backbone to perform image feature extraction. Since the module is designed to encode images from different modalities, the weight parameters of the two branches are independent, with each branch encoding features for fixed and moving, respectively. The structure of both branch encoders is identical, and we describe one branch as an example. The encoder consists of four encoding blocks. The first block primarily increases the feature channels without changing the image feature dimensions. The following three encoding blocks have the same structure, as shown in Fig. 2(a). These blocks include convolutional downsampling and convolutional processes (instance normalization and activation functions applied after each convolution). After each encoding block, the number of channels in the features doubles, while the spatial dimensions are reduced by half. The output features at each layer are denoted as F_m^i, F_f^i for $i = 1, \ldots, 4$.

2.2 Multi-scale Feature Registration

The multi-scale deformable registration begins at the smallest scale, and the registration features and deformation field are propagated to the next scale for further refinement. The index i denotes the current scale, while i+1 refers to the previous (coarser) scale. Current multiscale pyramid registration models typically only propagate the deformation field backward. In this paper, we simultaneously propagate the image registration features and the deformation field. Specifically, we propose MSFR module to handle the input and output of both registration features and deformation fields.

The MSFR module(Fig. 2(b)) receives four inputs: the previous deformation field ϕ^{i+1} and registration features F_c^{i+1}, the current scale's moving image features F_m^i and fixed image features F_f^i. The registration features F_c^{i+1} are first processed by a transposed convolution to adjust dimensions to $\hat{F}_c^{i+1}$, while the moving image features F_m^i are warped by the deformation field to obtain $\hat{F}_m^i$. Then $\hat{F}_c^{i+1}$,F_f^i and $\hat{F}_m^i$ are concatenated and fed into a two-stage image feature extraction module, which consists of convolutional layers, normalization, and activation functions, ultimately outputting the registration features F_c^i. These registration features can be directly passed to the next scale, while also being input into the registration head to predict the deformation field(ϕ^i) at the current scale. The output deformation field at each scale is combined with the previously obtained deformation field. The registration head mainly consists of convolutional modules and a diffeomorphism layer [27]. The processing flow of MSFR can be represented by Eq. (1).

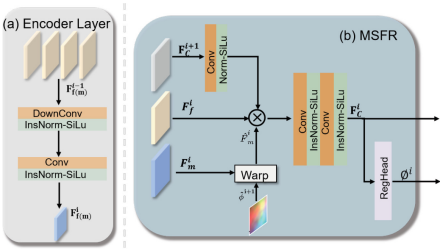


Fig. 2. (a) Details of the encoder block architecture. (b) Details of the MSFR architecture

$$\text{MSFR}\begin{cases}\hat{F}_c^{i+1} = ConvUP(F_c^{i+1}),\\ \hat{F}_m^{i} = warp(F_m^{i}, \hat{\phi}^{i+1}),\\ F = concat(F_f^{i}, \hat{F}_m^{i}, \hat{F}_c^{i+1})\\ F_c^{i} = CConv(F),\\ \phi^{i} = RegHead(F_c^{i}),\\ i = 3, 2, 1\end{cases} \tag{1}$$

where "CConv" denotes a two-layer convolution operation.

2.3 Multi-layer Modality-Invariant Contrastive Representation

The major challenge in multimodal registration is the intensity differences between images. In multiscale deformable registration research, even with dual-branch encoders, the model is not guaranteed to learn modality-invariant image features. To address this issue, we propose a multi-layer modality-invariant contrastive loss. We first introduce the design philosophy. *For the same patient, even with images acquired at different times and from different modalities, there tends to be a higher structural similarity.* In contrast, images from different patients or the same modality across patients exhibit noticeable structural differences. The fundamental idea behind contrastive learning [16,17] is to establish correlations between two signals, the 'query' and its 'positive' example, while contrasting them with other examples in the dataset, referred to as 'negatives.' Therefore, we propose treating images from the same patient across different modalities as positive pairs and images from different patients as negative pairs. This approach helps constrain the model to learn more essential or modality-independent feature information. Due to the presence of multiple output layers in the encoder branch, we propose computing the contrastive loss for the feature output at each layer. In summary, we define the contrastive loss [18] as follows:

$$\mathcal{L} = -\log\left[\frac{\exp(\boldsymbol{v}\cdot\boldsymbol{v}^{+}/\tau)}{\exp(\boldsymbol{v}\cdot\boldsymbol{v}^{+}/\tau)+\sum_{n=1}^{N}\exp(\boldsymbol{v}\cdot\boldsymbol{v}_n^{-}/\tau)}\right], \mathcal{L}_{\text{MCL}}(F_f, F_m) = \sum_{l=1}^{L} w_l \mathcal{L}^l \tag{2}$$

Here, τ represents a temperature parameter used to scale the distance between the query and other examples, with a default value of 0.5. $(\boldsymbol{v}, \boldsymbol{v}^{+})$ represents positive pairs from the same patient across different modalities. $(\boldsymbol{v}, \boldsymbol{v}_n^{-})$ represents the different patients across different modalities are referred to as negative pairs. l represents the layer index, and w_l represents the weight associated with a specific layer.

2.4 Recursive Inference Optimization

The multi-scale registration model can be viewed as an independent registration process at each individual scale. By recursively applying the model at each scale,

it mimics traditional iterative optimization methods in registration [19]. In this paper, we propose that recursive-based architectural refinements can effectively enhance the performance of pre-trained multi-scale image registration models, achieving significant improvements with minimal additional computational overhead. Equation 3 illustrates the computation process.

$$\phi = f_\phi^1(f_\phi^2(f_\phi^3(f_\phi^4(F_m^4, F_f^4, \phi^4) \times a, F_m^3, F_f^3) \times b, F_m^2, F_f^2) \times c, F_m^1, F_f^1) \times d \quad (3)$$

Here, f_ϕ^i represents the prediction function for the deformation and registration features at different scales, while F_m^i, F_f^i denote the input features at each scale, and a, b, c, d represent the number of recursions at the corresponding scales. The values of these four hyperparameters should be determined through inference testing.

Specially, the four hyperparameters for recursive inference determine the number of recursions at each scale. We begin with a single recursion per scale to identify which scale yields the greatest improvement in registration quality. We then increment the recursion that scales until no further gains. Unlike tuning learning rates or loss weights, which typically requires retraining, this strategy adjusts recursions post-training, greatly enhancing usability. Dataset-specific tuning is necessary but easy to perform.

2.5 Loss Function

This paper focuses on unsupervised multi-modal medical image deformable registration. To measure the similarity between multimodal images, we adopt the MIND metric [20]. Additionally, we apply a regularization function that imposes constraints on the gradients of the deformation fields. Finally, we utilize the proposed MCLoss. Consequently, the overall loss function consists of three components.

$$L(I_f, I_m, \phi) = \lambda_1 L_{MIND}(I_f, I_m \circ \phi) + \lambda_2 L_{\mathrm{MCL}}(F_f, F_m) + \sum_{p \in \Omega} \|\nabla \phi(p)\|^2 \quad (4)$$

where λ_1, λ_2 are the hyperparameters used to balance the contribution of loss functions. P represents the image voxel.

3 Experiments

3.1 Datasets and Comparison Methods

The first dataset is a proprietary dataset comprising 3D multimodal MR-CT liver-paired data(The data has undergone an ethical review). The dataset comprised 122 cases: hospital A(HA) included 40 cases, hospital B(HB) included 68 cases, and hospital C(HC) included 14 cases. For model training, we utilized all data from HA and 60 cases from HB, others were reserved for testing. These data were resampled to a voxel spacing of 1.75 × 1.75 × 2, then were

cropped to 160 × 160 × 64. The second dataset is the publicly multi-modal dataset AbdomenCT-1K(AMOS) [21]. We used 30 unpaired but annotated MR and CT, with 24 MR/CT volumes allocated for training and 6 for testing. Considering the combination of MR and CT, the actual training data consists of 576 pairs, with 36 pairs for testing. Additionally, we performed resampling and cropping on this dataset to ensure consistency in data dimensions.

To evaluate the performance of our proposed model, we compared it with several SOTA deformable registration algorithms, including VoxelMorph(VM) [11], TransMorph(TM) [22], VIT-V-Net [23], LKU-Net [25], RDP [26], and PIVIT [24].

We employed the Dice Similarity Coefficient (DSC) as an evaluation metric. Furthermore, to assess the plausibility of the deformation fields, we computed the percentage of voxels with a negative Jacobian determinant($(\%|J_\phi| \leq 0)$, which indicates local folding within the deformation field. In addition, we also compared the model's parameters and inference time.

3.2 Training Settings

All models are implemented using the pytorch. The model training was performed using an NVIDIA GeForce RTX 3090 GPU and an Intel i9-12900K CPU. We implement the model using Adam optimizer with a learning rate of $1e^{-4}$. The batch size is set as 4 and the networks are trained for 200 epochs. For the paired private datasets, λ_1 and λ_2 were set to 10 and 2, respectively. For the unpaired AMOS, we relax the weighting of the MCL, λ_1 and λ_2 were set to 10 and 0, respectively.

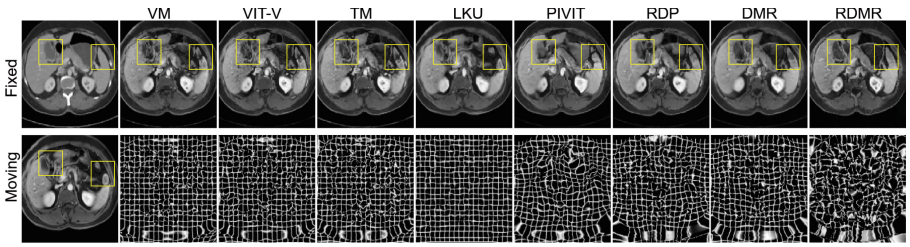


Fig. 3. The visualization of the results and the corresponding deformation fields

4 Results

Figure 3 shows the visualization results of different models. The content within the yellow boxes highlights that our proposed model exhibits a structure more similar to the fixed image than other models. The visualization of the deformation field (second row) shows that our proposed model is capable of producing a wider range of deformations, particularly in areas corresponding to

complex anatomical structures. Table 1 summarizes the comparative results of different models across three multimodal liver datasets. Our proposed DMR model demonstrates superior performance on all evaluation metrics compared to baseline methods. Notably, the RDMR with recursive inference achieves further improvements over the DMR. In terms of the DSC metric, RDMR achieves the best performance across all three datasets. Specifically, on the HB dataset, the DSC score improves by 1.4% relative to the baseline model VM, and on the AMOS dataset, the improvement is 4.5%. The proportion of negative Jacobian determinants about most models remain below 0.1%, indicating most models handle the image folding problem better.

Table 1. Comparison of results from different methods across three datasets

	HB		HC		AMOS			
Model	DSC↑	$\%\lvert J_\phi\rvert \leq 0 \downarrow$	DSC↑	$\%\lvert J_\phi\rvert \leq 0 \downarrow$	DSC↑	$\%\lvert J_\phi\rvert \leq 0 \downarrow$	T(S)	P(K)
VM	86.6±8.1	<0.03%	85.2±11.1	<0.15%	80.4±5.9	<0.15%	0.04	301.4
VIT-V	86.8±8.2	<0.02%	85.9±12.1	<0.08%	80.5±5.8	<0.16%	0.04	31519.8
TM	86.0±8.1	<0.05%	84.2±13.9	<0.10%	80.1±5.6	<0.16%	0.06	46771.2
LKU	84.3±9.1	<0.001%	78.3±17.4	<0.001%	79.0±5.8	<0.001%	0.03	2087.6
PIVIT	85.9±7.5	<0.001%	85.1±9.1	<0.01%	84.6±5.4	<0.08%	0.02	664.9
RDP	86.9±7.6	<0.001%	85.6±8.2	<0.001%	84.5±5.3	<0.03%	0.08	2240.8
DMR	87.7±8.2	<0.001%1	86.2±8.5	<0.001%	84.7±5.7	<0.05%	0.07	1904.2
RDMR	**88.0**±8.0	<0.001%	**87.0**±11.3	<0.001%	**84.9**±5.8	<0.04%	0.08	1904.2

The models tested on the HC dataset were trained on the HB dataset. From the metrics in Table 1, it is evident that each model performs worse on the HC than on the HB, as changes in data distribution can lead to a decline in the model's generalization ability. However, our proposed model still achieves the best performance on the HC dataset, demonstrating that the modality-invariant constraints enable the model to focus on the essential information in the images. This also indicates that our method helps mitigate the model generalization issues caused by multi-center data shift. From the structural perspective, models that employ a multi-scale registration strategy, such as PIVIT, RDP, DMR, and RDMR tend to achieve better results when addressing large deformation registration problems. The trend is particularly pronounced in the test results of the AMOS dataset. The last two columns of Table.1 presents each model's inference time and parameters. It is evident that all models exhibit fast inference times. Regarding the model parameter, our proposed model falls into the mid-range category. Thus, considering both performance and model complexity, our model achieves a better balance.

Ablation Study on Recursive Inference: Table 2 presents the model's performance based on different recursion depths. The choice of recursion follows

two rules: (1) select the setting yielding the highest DSC; (2) if DSC is equal, choose the configuration with lower computational cost. We first conducted a single recursive analysis across different scales, corresponding to columns 3–6 in Table 2. For the HB and HC datasets, the results improved when the configuration was set to (a,b,c,d) = 1121. We further tested the configuration 1131 and found that the performance declined. Therefore, the best model structure for the HB and HC datasets was 1121. Similarly, we analyzed the AMOS dataset and found that the performance improved with the configuration (a,b,c,d) = 1211, but declined when set to 1311. Thus, the most suitable model structure for the AMOS dataset was 1211. The experimental results indicate that, for the trained model, a simple inference test is sufficient to determine its final structure, further enhancing performance.

Table 2. shows the recursive registration results at different scales. The four-digit number represents the number of recursions at each scale, corresponding to the model's structural parameters a, b, c, and d

	1111	2111	1211	1121	1112	1131	1311
HB	87.7±8.2	87.1±9.4	87.2±8.8	**88.0**±8.0	87.0±8.3	87.9±7.9	–
HC	86.2±15.0	80.7±14.5	83.9±13.1	**87.0**±11.3	86.8±14.2	86.8±12.9	–
AMOS	84.7±5.7	82.7±6.4	**84.9**±5.8	84.6±5.8	84.8±6.1	–	84.8±6.0

5 Conclusion

We propose a novel multimodal deformable registration model RDMR, which effectively addresses the challenges of multi-modal registration, large deformations, and model generalizability. We propose a multiscale registration feature fusion mechanism and a multi-layer modality-invariant contrastive loss constraint to address the challenges of large deformations and multimodal registration, respectively. Additionally, we introduce a recursive inference strategy that further enhances the model's performance. This strategy is versatile and can be adopted by researchers in related fields. Our future research direction involves further optimizing the concept of recursive inference, allowing it to dynamically determine the appropriate number of recursions.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China Grant 62471297, in part by the Ministry of Science and Technology of the People's Republic of China under Grant 2023YFC2411401, in part sponsored by the Fundamental Research Funds for the Central Universities (No. YG2025ZD03).

Disclosure of Interests. The authors have no competing interests to declare relevant to this article 's content.

References

1. Zou, J., Gao, B., Song, Y., et al.: A review of deep learning-based deformable medical image registration. Front. Oncol. **12**, 1047215 (2022)
2. Xiao, H., Teng, X., Liu, C., et al.: A review of deep learning-based three-dimensional medical image registration methods. Quant. Imaging Med. Surg. **11**(12), 4895 (2021)
3. Mok, T.C.W., Chung, A.C.S.: Large deformation diffeomorphic image registration with laplacian pyramid networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 211–221. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_21
4. Zheng, J.Q., Wang, Z., Huang, B., et al.: Residual Aligner-based Network (RAN): motion-separable structure for coarse-to-fine discontinuous deformable registration. Med. Image Anal. **91**, 103038 (2024)
5. Song, X., et al.: Cross-modal attention for MRI and ultrasound volume registration. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 66–75. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_7
6. Shi, J., He, Y., Kong, Y., et al.: Xmorpher: full transformer for deformable medical image registration via cross attention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 217–226. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_21
7. Liu, L., Huang, Z., Liò, P., et al.: PC-SwinMorph: patch representation for unsupervised medical image registration and segmentation. arXiv preprint arXiv:2203.05684 (2022)
8. Guo, T., Wang, Y., Shu, S., et al.: Mambamorph: a mamba-based framework for medical mr-ct deformable registration. arXiv preprint arXiv:2401.13934 (2024)
9. Ma, T., Zhang, S., Li, J., et al.: IIRP-net: iterative inference residual pyramid network for enhanced image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11546–11555 (2024)
10. Tan, Z., Zhang, L., Lv, Y., et al.: GroupMorph: medical image registration via grouping network with contextual fusion. IEEE Trans. Med. Imaging **43**, 3807–3819 (2024)
11. Balakrishnan, G., Zhao, A., Sabuncu, M.R., et al.: Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans. Med. Imaging **38**(8), 1788–1800 (2019)
12. Mok, T.C.W., Li, Z., Bai, Y., et al.: Modality-agnostic structural image representation learning for deformable multi-modality medical image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11215–11225 (2024)
13. Pham, X.L., Luu, M.H., van Walsum, T., et al.: CMAN: cascaded multi-scale spatial channel attention-guided network for large 3D deformable registration of liver CT images. Med. Image Anal. **96**, 103212 (2024)
14. Kang, M., Hu, X., Huang, W., et al.: Dual-stream pyramid registration network. Med. Image Anal. **78**, 102379 (2022)
15. Hu, B., Zhou, S., Xiong, Z., et al.: Recursive decomposition network for deformable image registration. IEEE J. Biomed. Health Inform. **26**(10), 5130–5141 (2022)
16. Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y.: Contrastive learning for unpaired image-to-image translation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 319–345. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_19

17. Hu, Y., Zhang, S., Li, W., et al.: Unsupervised medical image synthesis based on multi-branch attention structure. Biomed. Signal Process. Control **104**, 107495 (2025)
18. Chen, T., Kornblith, S., Norouzi, M., et al.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
19. Zhao, S., Lau, T., Luo, J., et al.: Unsupervised 3D end-to-end medical image registration with volume tweening network. IEEE J. Biomed. Health Inform. **24**(5), 1394–1404 (2019)
20. Heinrich, M.P., Jenkinson, M., Bhushan, M., et al.: MIND: modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. **16**(7), 1423–1435 (2012)
21. Ma, J., Zhang, Y., Gu, S., et al.: Abdomenct-1k: is abdominal organ segmentation a solved problem? IEEE Trans. Pattern Anal. Mach. Intell. **44**(10), 6695–6714 (2021)
22. Chen, J., Frey, E.C., He, Y., et al.: Transmorph: transformer for unsupervised medical image registration. Med. Image Anal. **82**, 102615 (2022)
23. Chen, J., He, Y., Frey, E.C., et al.: Vit-v-net: vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468 (2021)
24. Ma, T., Dai, X., Zhang, S., et al.: PIViT: large deformation image registration with pyramid-iterative vision transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 602–612. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43999-5_57
25. Jia, X., Bartlett, J., Zhang, T., et al.: U-net vs transformer: is u-net outdated in medical image registration? In: International Workshop on Machine Learning in Medical Imaging, pp. 151–160. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-21014-3_16
26. Wang, H., Ni, D., Wang, Y.: Recursive deformable pyramid network for unsupervised medical image registration. IEEE Trans. Med. Imaging **43**(6), 2229–2240 (2024)
27. Dalca, A.V., Balakrishnan, G., Guttag, J., et al.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Med. Image Anal. **57**, 226–236 (2019)

Real Super-Resolution for Proximal Femur: Enhanced Computation of Structural Bone Metrics from Clinical CTs

Niklas C. Koser[1(✉)], Marten J. Finck[2], Felix N. von Brackel[3], Benjamin Ondruschka[4], Sören Pirk[2], and Claus-C. Glüer[1]

[1] i2Lab@Section Biomedical Imaging, Kiel University, University Hospital Schleswig-Holstein, Kiel, Germany
{niklas.koser,glueer}@rad.uni-kiel.de
[2] Visual Computing and Artificial Intelligence, Kiel University, Kiel, Germany
{mafi,sp}@informatik.uni-kiel.de
[3] Department of Osteology and Biomechanics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[4] Institute of Legal Medicine, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Abstract. Fracture risk due to osteoporosis is a highly prevalent disease with costs in the European Union alone of 56 billion p.a. Accurate assessment of the microarchitecture of the proximal femur (e.g., trabecular thickness, trabecular spacing, bone volume fraction) is essential for assessing bone strength and predicting fracture risk. High resolution (HR) CT provides the necessary spatial resolution. However, for best hip fracture risk assessment HR-CT imaging should be performed at the proximal femur but this would require an unacceptably high level of radiation dose. Therefore, we aimed to investigate whether deep learning based super-resolution (SR) models applied to low-resolution (LR) clinical CT images permit improved assessment of structural parameters.

In this study we adapted and optimized state-of-the-art model architectures to compare them in the context of CT-SR of the proximal femur. The dataset used consisted of pairs of clinical LR-CTs and HR-CTs of 50 individuals. This represents clinical reality and avoids bias of downsampling HR images to mimic LR images. Using automated preprocessing data is prepared for model training. We used three-stage template matching of point clouds to automatically extract the relevant regions of interest, from which metrics for bone microarchitecture were determined. We compared SRGAN, Real-ESRGAN+, LDM, and ResShift regarding improvement in structural assessment. We also tested whether 2.5D approaches –using multiple slices of the CT– are superior to 2D approaches. In terms of perceptual reconstruction, the ResShift 2.5D model outperforms the other SR models and achieves comparable results to the Real-ESRGAN+ architectures in the derivation of biomechanical properties.

N. C. Koser and M. J. Finck—Contributed equally to this work.
Code and data are available at: https://github.com/nkoser/Real-Super-Resolution.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 523–532, 2026.
https://doi.org/10.1007/978-3-032-04947-6_50

Keywords: CT Super-Resolution · Bone Microstructure · Clinical CT · Osteoporosis · Deep Learning

1 Introduction

Osteoporosis is a systemic skeletal disease characterized by reduced bone mass and microarchitectural structure of the bone [1,21]. The trabecular bone structure is degraded faster than the compact cortical bone, which is why bones with a high proportion of spongiosa, such as the proximal femur, have an increased risk of osteoporotic fractures [1]. The prevalence of osteoporosis increases significantly with age. In view of demographic trends and increasing life expectancy, it is predicted that the incidence of osteoporotic fractures will continue to rise substantially in the future [1,18,21]. This development significantly impairs the quality of life of those affected and also represents a considerable financial burden for healthcare systems – in 2019 alone, 4.3 million fractures in the European Union resulted in costs exceeding 56 billion euros [1,15,21]. To prevent these consequences, an efficient prognosis and early diagnosis are crucial, as timely preventive measures in line with guidelines yield high success rates. According to the current recommendations of the World Health Organization (WHO), the diagnosis of osteoporosis is primarily based on determining bone mineral density (BMD) using Dual-Energy X-Ray Absorptiometry (DXA) [1,21]. However, this standard metric only has a classification accuracy of around 65 % for assessing risk of hip fracture [11]. The scientific literature therefore recommends the integration of additional metrics that capture the microstructural bone properties. Parameters such as bone volume fraction (BV/TV), trabecular thickness (Tb.Th), trabecular spacing (Tb.Sp) and trabecular number (Tb.N) have been shown to be predictive of the mechanical strength of the bone and thus offer high diagnostic and prognostic potential for estimating fracture risk [1,18–20].

High-resolution peripheral quantitative computed tomography (HRpQCT), here denoted as HR-CT, facilitates the acquisition of detailed images essential for analyzing the bone microstructure. However, it is currently mainly used in scientific settings because of high costs, a time-consuming procedure, and a comparatively small field of view. It is primarily limited to peripheral skeletal regions and would pose a high radiation exposure for *in vivo* imaging [1,11,19,20]. To overcome these limitations, super-resolution (SR) methods could be applied to clinical CT images that are characterized by a lower spatial resolution (denoted as LR-CT). SR methods are used to increase the image resolution in order to achieve an approximation of the quality of HR images. Deep learning (DL) based approaches, which are capable of significantly improving image quality by training on extensive paired data sets of LR and HR images, are particularly promising. Recent studies show that these methods outperform common interpolation methods in terms of image sharpness, texture detail and structural accuracy [17]. A noteworthy class of algorithms are Generative Adversarial Networks (GANs), which generate realistic HR-images through competition between a generator and a discriminator network [16,17,23,24]. More recently, diffusion models have emerged as a promising class of algorithms due to their stable training dynamics and ability to provide highly detailed reconstructions [22].

In the field of DL-SR imaging of the proximal femur, only two approaches have been introduced so far – and both of these studies exhibit methodological and data-related limitations that hinder their applicability in clinical practice. Chan and Rajapakse [4] trained an SR3 model based on 26 HR-CT volumes with an isotropic voxel spacing of 240 µm. The corresponding LR-CT volumes were generated by bicubic downsampling to 720 µm. Despite improvements over methods like SRGAN, the significance of these results is limited by the small dataset size and the use of synthetically generated LR volumes. These downsampled images may still retain aspects of the HR data that are absent in truly independent LR acquisitions, making them less representative of real clinical imaging conditions. Frazer et al. [10] developed a SRGAN using a paired data set of 10 HR-CT volumes (60 µm isotropic) and 10 LR-CT volumes (0.3125 x 0.3125 x 0.25 mm) acquired, however, with a special clinical CT scanner that is not representative of commonly available CT scanners, as acknowledged by the authors. Therefore, the significance of these results is also limited. Besides the small and unbalanced sample size, the scaling applied during registration possibly leading to distortions in the metrics is a limitation. Although both studies are promising, they do not use current state-of-the-art (SOTA) SR models. Furthermore, they are based on inadequate datasets that are limited in both size and clinical relevance, as they still contain too much detail in the LR-CTs. There is also a lack of standardized evaluation methods, which makes it difficult to directly compare the results.

This paper aims to address these limitations by proposing the following contributions: (1) We introduce the largest paired dataset to date, consisting of 50 HR-CTs and corresponding LR-CTs, with the LR-CTs recorded under real clinical conditions. This dataset enables a realistic evaluation of SR methods in a medical context and should be made available to the research community; (2) We adapt and test various current SOTA SR models on 2D and 2.5D slices of the described dataset, also including the recently released ResShift [27] model that has not been evaluated on medical image data before; (3) We propose an automated evaluation pipeline based on template matching of surface point clouds. This method enables a standardized selection of regions of interest (ROIs), ensuring a reproducible evaluation of SR models based on established metrics.

2 Methodology

2.1 Individuals, Dataset and Data Preprocessing

Our dataset consists of paired CT images of 50 patients, including 31 male and 19 female individuals. The age of the patients ranges from 22 to 89 years, with an average age of 60.8 ± 17.5 years. Body height varies from 1.43 m to 2.00 m, with a mean of 1.73 m ± 0.11 m, while body weight ranges from 44 kg to 123 kg, with an average of 81.6 kg ± 19.4 kg. The LR-CTs were acquired post mortem *in situ* as whole-body scans with a Philips Incisive CT. Image acquisition was performed with anisotropic voxel spacing of 0.98 mm × 0.98 mm × 0.65 mm, peak kilo voltage output (kVp) of 120 kV, tube current of 113–130 mA, and an UB

filter. In the femur region, these CT scans exhibit a mean Signal-to-Noise Ratio (SNR) of 0.93 and a mean contrast ratio of 2.43. Subsequently, the HR-CTs were acquired using an HR-pQCT (Xtreme CT, SCANCO Medical AG, Brüttisellen, Switzerland). The left proximal femur was scanned *ex situ* with an isotropic voxel resolution of 82 µm, a kVp of 59.4 kV, 900 µA, and reconstructed with the Shepp-Logan kernel. This study was approved by the ethics committee of the Hamburg chamber of physicians (Reference Number: WF-057/21) and conducted in accordance with the ethical standards of the Declaration of Helsinki of 1964 and its later amendments.

In order to train neural networks for SR, semi-automatic preprocessing of our dataset is required to generate corresponding volume pairs. The following processing steps are performed for each pair of LR-CT and HR-CT. First, both DICOM volumes are loaded with the open source software 3D Slicer [8]. For initial adjustment, the LR-CT is roughly aligned with the HR-CT by manual translation and rotation. This is followed by precise registration using the BRAINS General Registration Module [14] with a rigid six degrees of freedom transformation model. The LR-CT serves as a moving volume and the HR-CT as a fixed volume. After registration, the LR-CT is resampled onto the isotropic voxel grid of the HR-CT using a specially developed 3D slicer module, which performs resampling with SimpleITK and a B-spline interpolation, and is subsequently cropped to match the spatial dimensions of the HR-CT. As only the LR-CT is transformed, the structural integrity of the HR-CT is preserved, ensuring that no relevant information is lost. This procedure is independent of the voxel spacing of the LR clinical CT scanner. Therefore, it is also transferable to other scanners. A segmentation mask is required for the further preprocessing steps of the volumes exported as DICOM. To automatically compute these masks, we use a 3D nnUNet [13] trained on a subset of the Total Segmentator dataset containing a total of 1,204 labelled CT images [26]. For this purpose, 102 CTs and their corresponding segmentation masks were first cropped to the relevant areas of the left and right femora. The resulting 204 volumes with an isotropic voxel spacing of 1 mm were then used as the basis for training. The trained nnUNet enables generating masks from downsampled LR-CTs of our dataset and upsamples them to the original resolution with voxel spacing of 82 µm. The value ranges were normalized to the range $[0, 1]$ and clipped. Finally, the mask is applied so that only the femur bone and no other structures (e.g., tissue or neighboring bones) is present in the volume.

2.2 Model Architectures

To enhance the resolution and reconstruction of anatomical structures in LR-CTs of the proximal femur, we adapted and compared four deep learning SR models (SRGAN [16], Real-ESRGAN+ [23], Latent Diffusion Model (LDM) [22] and ResShift [27]) both in 2D and 2.5D. The latter also includes four neighboring slices assuming that the contained structural information represents an advantage for the reconstruction of the considered slice.

SRGAN. Since Frazer et al. [10] were able to achieve promising results with an SRGAN, we aim to evaluate the capabilities of this model architecture trained on real LR-CTs of the proximal femur. Due to the voxel-to-voxel correspondence between the LR and HR-CTs, we removed the upsampling layer. Furthermore, we added four additional residual blocks into the generator, given that the dimensions of the input pixels exceed those of the original SRGAN. No modifications to the model were necessary for 2.5D, because the neighboring slices were integrated as additional channels.

Real-ESRGAN+. A strong limitation of SRGAN is its inability to reproduce finer structures in LR images [16]. To address these limitations Wang et al. developed ESRGAN [24] and Real-ESRGAN+ [23]. The latter approach not only focus on the reconstruction of fine details, but also introduces a process to eliminate undesirable artifacts that arise during the reconstruction process. Due to the larger input size compared to Real-ESRGAN+, we added three additional Residual-in-Residual Dense Blocks. Because of the low image sharpness of the LR-CTs and the different acquisition conditions compared to the HR-CT in our dataset, the default degradation process is omitted. No further adjustments were made to the original model regarding pixel unshuffling and the use of a unsharp masking filter to mitigate overshooting artifacts. The 2.5D input was processed in the same way as for the 2D input.

LDM. Due to the known challenges of training GANs [22], we use a LDM as another class of SR model algorithms. We deliberately decided against the SR3 model used by Chan et al. [4], as LDMs offers both higher efficiency and the ability to generate more realistic textures [22]. A LDM uses a variational autoencoder (VAE) to transform the HR image into a latent space with dimension that corresponds to those of the LR image. In our specific case, we apply pixel unshuffling to transform the LR-CT image to the dimension of the HR latent space $(6, 64, 64)$ without information loss. In the 2.5D case, the number of channels was increased to accommodate the requirements of pixel unshuffling.

ResShift. In contrast to LDMs, ResShift [27] is capable of generating visually plausible images within 15 iterations, facilitated by the formation of a Markov chain between HR and LR images. This is achieved by shifting the residual – the difference between the HR and LR images – which enhances the efficiency of transitions. ResShift demonstrates comparable or even superior performance compared to other SOTA SR methods across various applications [27]. Building on these findings, our study is the first to apply and evaluate ResShift on medical data, employing it as the fourth SR model in our experimental setup. To enable even faster training of ResShift in the latent space, we utilize a variational autoencoder (VAE) trained on HR-CTs. The VAE is also applied to LR-CTs to ensure alignment within the shared latent space. Additionally, a feature extractor is employed on the original LR-CTs to transmit LR features as conditioning inputs to the Swin-Unet [3]. In the 2.5D approach, the features of adjacent slices – extracted using the feature extractor adapted for five channels – are likewise supplied as conditioning inputs to the Swin-Unet.

3 Experiments and Results

3.1 Experimental Setup

Evaluation. Since the structural relationships are of primary importance in the reconstruction of the slices, the calculation of Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) is performed on the entire slices and are then averaged over the volume [25]. As SSIM and PSNR are sometimes unreliable for blurred images, we also record the Learned Perceptual Image Patch Similarity (LPIPS) [28] and the Gradient-SSIM (GSSIM) [5]. The average inference time (T/V) is also measured. To consistently and automatically extract structural bone metrics from the same anatomical region, independent of patient-specific variations and without manual ROI selection, we developed a multi-stage template matching approach based on surface point clouds. Bone volume, defined as the sum of bone voxels within the binary mask, was used to select three templates, which were uniformly sampled to ensure anatomical diversity. The three selected femur templates were segmented into anatomical regions (head, neck, trochanter, shaft) using 3D Slicer. The segmentation masks were converted into triangle meshes using Open3D [29], from which 10,000 labeled points were uniformly sampled as point clouds. The following procedure was then applied to all bone volume pairs (LR-CT/HR-CT or SR-CT/HR-CT): (1) The appropriate template is selected for both volumes and their segmentation mask. For each CT volume, a binary bone mask is then generated from a nnUNet trained on HR-CTs; (2) A surface point cloud is generated from the segmentation mask in the same way as the templates. We then apply a three-stage point cloud registration process. First, the point cloud centers are aligned by translation. This is followed in the second step by a rough registration using RANSAC [9]. Finally, a more precise fine adjustment is applied using ICP with scaling [2]. Then the labels of the template point cloud are transferred to the transformed point cloud so that a complete segmentation into the four anatomical regions is achieved; (3) The center of all associated points is calculated for the anatomical regions of the head, neck, and trochanter. 3D-ROIs with an edge length of 128 voxels are extracted around these reference points from the bone structure masks; (4) To evaluate the bone structure, BV/TV, Tb.Th, Tb.Sp, and Tb.N are calculated from the extracted ROIs using the Hildebrand algorithm [12].

Implementation Details. The dataset is divided into training, validation, and testing, corresponding to a 60/20/20 split. Due to the size of the volume data, 2D and 2.5D patches with a size of 512×512 are extracted from the axial slices. Sampling is centered around the voxel located inside the femur. Rotation and flipping are used to augment the data and ensure robustness.

All models are trained with Adam optimizer and a batch size of 64, with each model processing 10,000 samples per epoch. The NVIDIA H100 NVL GPU with 96 GB VRAM was used for this purpose. To train the SRGAN, the generator is first pre-trained with a learning rate of 1e–4 (multi-step learning rate scheduler) for 45 epochs in order to avoid overfitting the discriminator during

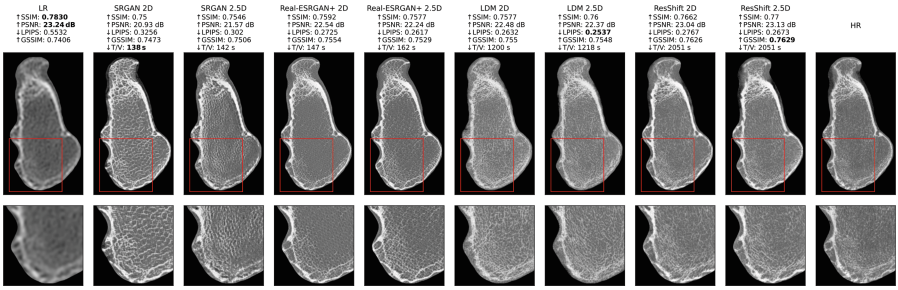


Fig. 1. Comparison of the tested SR models based on their reconstruction performance using an axial slice and an enlarged view (30 mm edge length)

the subsequent GAN training. The actual GAN training then takes place over 75 epochs, whereby the discriminator is optimized with a learning rate of 1e–6. The same training procedure is used for the Real-ESRGAN+. The generator is first pre-trained with a learning rate of 2e–4 over 20 epochs (constant learning rate scheduler), followed by adversarial GAN training over 60 epochs, whereby the discriminator is optimized with a learning rate of 1e–4. The LDM model was trained over 100 epochs with a learning rate of 1e–5 where forward diffusion was carried out with the DDPM method ($T = 1000$) and inference with the DDIM method ($T = 30$). The VAE was pre-trained over 75 epochs using a perceptual loss and a mean squared error (MSE) loss. A patch discriminator was then used for adversarial training (75 epochs, learning rate of 45e–6). The ResShift model was trained over 150 epochs with the same learning rate, using 8 diffusion steps and a Cosine scheduler. Our code will be shared publicly upon acceptance.

3.2 Comparative Analysis

The results used to assess the reconstruction performance of the SR models are presented in Fig. 1, exemplified by an axial slice from a test volume. Additionally, the average SSIM, PSNR, LPIPS and GSSIM values across all slices and volumes, along with the mean T/V, are reported. Qualitative analyses show that models based on GAN architectures generate symmetric recurrent structures as trabecular meshwork. SRGAN 2D in particular hallucinates microstructures that are not present in HR-CT. The LDM, on the other hand, is not able to correctly reproduce prominent structures such as the cortex and also shows only minor local differences in the trabecular meshwork. The most convincing model is the ResShift model, especially when trained on 2.5D data, as it can reconstruct both the cortex and the trabecular distribution well. For the other three models, the added value of 2.5D is marginal. The analysis of the quantitative metrics shows that SSIM and PSNR have the highest values for the LR-CT and thus do not match the qualitative results. The 2.5D LDM model achieves the best LPIPS value. The highest agreement with the visually assessed image quality is shown by the GSSIM, which is highest for ResShift 2.5D at 0.7629 and lowest for the

Table 1. Mean, standard deviation and Pearson correlation coefficient (r) of the bone microstructure metrics. Tb.Th and Tb.Sp are in mm, Tb.N in mm^{-1}

		SRGAN		Real-ESRGAN+		LDM		ResShift		
	LR	2D	2.5D	2D	2.5D	2D	2.5D	2D	2.5D	HR
BV/TV_h	.15±.2 r=.5	.54±.14 r=.35	.46±.15 r=.42	.43±.16 r=**.73**	.41±.14 r=.71	.62±.07 r=.29	.64±.08 r=.37	0.59±0.1 r=.64	.62±.1 r=**.73**	.56±.06
$Tb.Th_h$	.86±.53 r=.33	.6±.08 r=.64	.63±.1 r=.81	.51±.12 r=.77	.5±.1 r=.8	.54±.06 r=.82	.57±.07 r=.74	.63±.11 r=**.86**	.63±.11 r=**.86**	.58±.08
$Tb.Sp_h$	5.1±3.8 r=0	.63±.15 r=.13	.89±.49 r=−.08	.65±.14 r=.2	.65±.12 r=**.33**	.41±.06 r=−.22	.41±.06 r=−.1	.5±.09 r=−.06	.47±.07 r=.07	.64±.05
$Tb.N_h$	.21±.09 r=−.58	.82±.09 r=.22	.69±.15 r=−.15	.86±.06 r=.21	.88±.05 r=.47	1.1±.07 r=.74	1±.06 r=**.79**	.89±.09 r=.67	.92±.09 r=.62	.82±.06
BV/TV_n	.01±.01 r=.35	.31±.2 r=−.03	.25±.2 r=.21	.05±.04 r=.82	.06±.04 r=.77	.1±.07 r=.64	.12±.07 r=.83	.06±.05 r=.81	.06±.04 r=**.85**	.11±.1
$Tb.Th_n$	.79±1.18 r=.95	.46±.11 r=.85	.46±.14 r=.92	.4±.32 r=**.97**	.36±.2 r=**.97**	.32±.21 r=**.97**	.34±.19 r=**.97**	.36±.27 r=**.97**	.35±.29 r=**.97**	.35±.14
$Tb.Sp_n$	8.4±8.3 r=−.41	.99±.53 r=−.28	1.4±.95 r=.39	2.8±1.9 r=.72	2.1±1.3 r=**.93**	1.3±.78 r=−.2	1±.19 r=.51	2.3±.83 r=.49	2.2±1 r=.74	3.5±6
$Tb.N_n$	.07±.07 r=.37	.76±.24 r=−.27	.66±.29 r=.13	.41±.18 r=.68	.49±.17 r=**.75**	.72±.24 r=−.18	.75±.13 r=.51	.42±.15 r=.62	.44±.16 r=.74	.53±.25
BV/TV_t	0±.01 r=.27	.27±.2 r=−.24	.22±.19 r=.13	.03±.04 r=**.76**	.04±.03 r=.63	.05±.03 r=.08	.05±.04 r=.19	.02±.01 r=.06	.02±.02 r=.71	.11±.09
$Tb.Th_t$	.22±.24 r=.26	.4±.07 r=−.18	.37±.06 r=.34	.26±.04 r=**.83**	.27±.02 r=.75	.22±.02 r=.15	.23±.02 r=.58	.22±.03 r=.79	.21±.02 r=.79	.3±.05
$Tb.Sp_t$	6.9±8.3 r=−.19	1±.64 r=−.15	1.5±1.2 r=−.17	3.1±2.2 r=**.83**	1.9±.76 r=.6	1.3±.5 r=−.08	1.4±.41 r=−.21	2.8±1 r=.58	2.9±1.5 r=.48	1.6±.8
$Tb.N_t$	.05±.07 r=.26	.78±.23 r=−.18	.7±.3 r=−.04	.4±.2 r=**.66**	.51±.16 r=.6	.72±.19 r=.01	.66±.16 r=−.09	.37±.12 r=.58	.4±.19 r=.62	.6±.18

LR-CT at 0.7406. The GAN-based models have a significantly shorter inference time compared to the diffusion models. While the SRGAN 2D generates an SR volume in 138 s on average, the LDM 2D requires 1,200 s. The metrics of bone microstructure are compared for all models in Table 1.

The analysis of the correlation coefficients shows that the ResShift 2.5D and the Real-ESRGAN+ architectures in particular are superior to the other SR models examined in the reconstruction of bone microstructures. Both models, when applied in a 2.5D configuration, demonstrated statistically significant improvements over metrics derived from the LR-CT scans in 8 out of 12 evaluation metrics, as determined by Friedman's test followed by a Conover post-hoc analysis [6,7]. In contrast, the SRGAN achieved the lowest correlation with the HR-CTs. In most cases, all models provide a more accurate estimation of structural bone metrics.

4 Conclusion, Limitations and Future Work

In this paper we adapted and compared different SOTA SR model architectures – trained on a newly introduced dataset reflecting the clinical reality – for improving the calculation of bone microstructure metrics. Our analysis demonstrates, that ResShift 2.5D outperforms existing model architectures regarding perceptual reconstruction and shows similar results to the Real-ESRGAN+ when deriving structural bone metrics. Additionally, our introduced evaluation

pipeline enables standardized comparisons for other approaches. Although our results are promising, we also identified the following limitations: (1) The LR-CTs have limited resolution due to the real clinical setting leading to a loss of information and possible hallucination of bone microstructure; (2) Our method has only been validated with one clinical CT scanner; (3) Since the training is performed exclusively with axial slices, the reconstruction quality in the other two dimensions is less precise; (4) The quality of the results in both preprocessing and evaluation depends significantly on the accuracy of the binary masks. Given the current state, there are several avenues for future work, including the assessment with other metrics, the training on 3D patches and the validation through a suitable downstream task.

Acknowledgments. This work has been financially supported by the Federal Ministry of Education and Research, Germany (grant 01EC1908A). Furthermore, this work was supported by the IDIR-Project (Digital Implant Research), a cooperation financed by Kiel University, University Hospital Schleswig-Holstein and Helmholtz Zentrum Hereon.

Disclosure of Interest. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bartl, R. (ed.): Osteoporose: Prävention Diagnostik Therapie. Georg Thieme Verlag, Stuttgart, 01 edn. (2010)
2. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor Fusion IV: Control Paradigms and Data Structures, vol. 1611, pp. 586–606. Spie (1992)
3. Cao, H., et al.: Swin-unet: unet-like pure transformer for medical image segmentation. In: Computer Vision – ECCV 2022 Workshops, pp. 205–218. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9
4. Chan, T.J., Rajapakse, C.S.: A super-resolution diffusion model for recovering bone microstructure from ct images. Radiol. Artif. Intell. **5**(6), e220251 (2023)
5. Chen, G.H., Yang, C.L., Xie, S.L.: Gradient-based structural similarity for image quality assessment. In: 2006 International Conference on Image Processing, pp. 2929–2932 (2006)
6. Conover, W.J.: Practical Nonparametric Statistics. Wiley, Hoboken (1999)
7. Conover, W.J., Iman, R.L.: Multiple-comparisons procedures. informal report. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States) (1979)
8. Fedorov, A., et al.: 3d slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012)
9. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
10. Frazer, L.L., Louis, N., Zbijewski, W., Vaishnav, J., Clark, K., Nicolella, D.P.: Super-resolution of clinical ct: revealing microarchitecture in whole bone clinical ct image data. Bone **185**, 117115 (2024)

11. Galassi, A., Martín-Guerrero, J.D., Villamor, E., Monserrat, C., Rupérez, M.J.: Risk assessment of hip fracture based on machine learning. Appl. Bionics Biomech. **2020**, 8880786 (2020)
12. Hildebrand, T., Rüegsegger, P.: A new method for the model-independent assessment of thickness in three-dimensional images. J. Microsc. **185**(1), 67–75 (1997)
13. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
14. Johnson, H., Harris, G., Williams, K., et al.: Brainsfit: mutual information rigid registrations of whole-brain 3d images, using the insight toolkit. Insight J. **57**(1), 1–10 (2007)
15. Kanis, J.A., et al.: Scope 2021: a new scorecard for osteoporosis in europe. Arch. Osteoporos. **16**(1), 1–82 (2021)
16. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4681–4690 (2017)
17. Lepcha, D.C., Goyal, B., Dogra, A., Goyal, V.: Image super-resolution: a comprehensive review, recent trends, challenges and applications. Inf. Fusion **91**, 230–260 (2023)
18. Maquer, G., Musy, S.N., Wandel, J., Gross, T., Zysset, P.K.: Bone volume fraction and fabric anisotropy are better determinants of trabecular bone stiffness than other morphological variables. J. Bone Miner. Res. **30**(6), 1000–1008 (2015)
19. Mc Donnell, P., Mc Hugh, P.E., O' Mahoney, D.: Vertebral osteoporosis and trabecular bone quality. Ann. Biomed. Eng. **35**(2), 170–189 (2007)
20. Nishiyama, K.K., Shane, E.: Clinical imaging of bone microarchitecture with hr-pqct. Curr. Osteoporos. Rep. **11**(2), 147–155 (2013)
21. Porter, J.L., Varacallo, M.A.: Osteoporosis. In: StatPearls. StatPearls Publishing, Treasure Island (FL) (2025). http://www.ncbi.nlm.nih.gov/books/NBK441901/
22. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
23. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1905–1914 (2021)
24. Wang, X., et al.: Esrgan: enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
25. Wang, Z.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
26. Wasserthal, J., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiol. Artif. Intell. **5**(5) (2023)
27. Yue, Z., Wang, J., Loy, C.C.: Resshift: efficient diffusion model for image super-resolution by residual shifting. Adv. Neural Inf. Process. Syst. **36** (2024)
28. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
29. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: A modern library for 3D data processing (2018)

Region-Based Text-Consistent Augmentation for Multimodal Medical Segmentation

Kunyan Cai[1,5], Chenggang Yan[2], Min He[3], Liangqiong Qu[4], Shuai Wang[2,5(✉)], and Tao Tan[1(✉)]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macao, China
{p2317017,taotan}@mpu.edu.mo
[2] Hangzhou Dianzi University, Hangzhou, China
{cgyan,shuaiwang}@hdu.edu.cn
[3] Hangzhou Institute of Medicine, Chinese Academy of Sciences, Hangzhou, China
hemin@him.cas.cn
[4] Faculty of Science, The University of Hong Kong, Hong Kong, China
liangqqu@hku.hk
[5] Lishui Institute of Hangzhou Dianzi University, Lishui, China

Abstract. Medical image segmentation is crucial for various clinical applications, and deep learning has significantly advanced this field. To further enhance performance, recent research explores multimodal data integration, combining medical images and textual reports. However, a critical challenge lies in image data augmentation for multimodal medical data, specifically in maintaining text-image consistency. Traditional augmentation techniques, designed for unimodal images, can introduce mismatches between augmented images and text, hindering effective multimodal learning. To address this, we introduce **R**egion-**B**ased **T**ext-**C**onsistent **A**ugmentation (RBTCA), a novel framework for coherent multimodal augmentation. Our approach performs region-based image augmentation by first identifying image regions described in associated text reports and then extracting textual cues grounded in these regions. These cues are integrated into the image, and augmentation is subsequently performed on this modality-aware representation, ensuring inherent text-cue consistency. Notably, the RBTCA's plug-and-play design allows for straightforward integration into existing medical image analysis pipelines, enhancing its practical utility. We demonstrate the efficacy of our framework on the QaTa-Covid19 and our in-house Lung Tumor CT Segmentation (LTCT) datasets, achieving substantial gains, with a Dice coefficient improvement of up to 7.24% when integrated into baseline segmentation models. Our code will be released on https://github.com/KunyanCAI/RBTCA.

Keywords: Medical Image Segmentation · Data Augmentation · Multimodal Learning · Text-Image Consistency

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 533–543, 2026.
https://doi.org/10.1007/978-3-032-04947-6_51

1 Introduction

Medical image segmentation has become a critical component of modern healthcare, playing an increasingly important role in diagnosis, treatment planning, and disease monitoring across various modalities such as X-ray, CT, and MRI. The application of deep learning techniques has revolutionized this field, leading to remarkable advancements in automated medical image segmentation [2,3,10,13]. To further enhance the performance and clinical applicability of these segmentation systems, recent research has explored the integration of multimodal data [5,8], particularly combining medical images with associated textual reports like radiology reports and clinical notes [16]. These multimodal data sources offer complementary information and hold significant promise for achieving a more comprehensive and nuanced understanding of medical images, ultimately leading to improved segmentation performance and clinical utility [7,12,19,22].

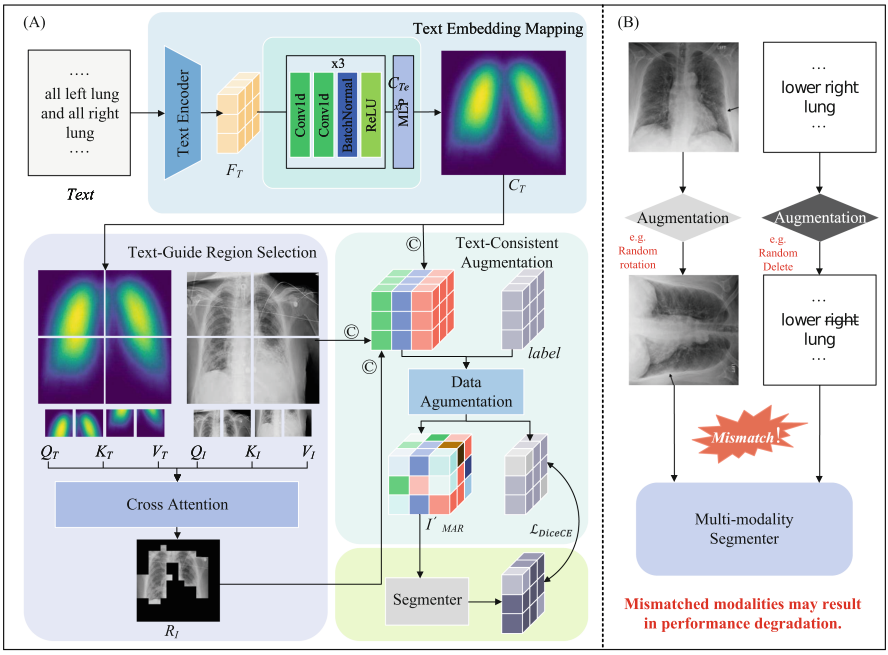


Fig. 1. Overview our proposed RBTCA framework and illustration of multimodal mismatch. (A): Visualized workflow of RBTCA. (B): Illustration of the semantic mismatch problem in multimodal data augmentation.

To effectively train deep learning models, especially in data-scarce domains like medical imaging, data augmentation is an indispensable technique. Data augmentation serves as a crucial strategy to synthetically increase the training dataset by applying various transformations to existing data, thereby mitigating

the challenges of limited annotated data, reducing overfitting, and enhancing the generalization capability of deep learning models [4,14,18,23]. This is particularly vital in medical image segmentation, where acquiring large, diverse, and expertly annotated datasets is often hindered by factors such as data scarcity, high annotation costs associated with expert radiologists, and stringent patient privacy regulations [6,24]. Traditional augmentation strategies, such as geometric transformations (e.g., rotation, flipping) and intensity adjustments (e.g., noise injection, contrast enhancement), have proven effective in unimodal settings by improving model generalization and reducing overfitting [14,23]. Similarly, in multimodal contexts involving paired image-text data (e.g., radiology reports with X-rays), basic text augmentation techniques like synonym replacement or word deletion are occasionally applied [11,15]. Yet, these methods operate in isolation, treating image and text modalities as independent streams. As illustrated in Fig. 1 (Part B), this isolation introduces semantic inconsistencies: for instance, flipping an X-ray image horizontally while retaining a textual description of "infections in the lower right lung" creates a mismatch between the augmented image and its report. Such discrepancies, quantified in Table 1, degrade the quality of multimodal training data, confuse models during joint representation learning, and ultimately limit segmentation accuracy.

Current augmentation frameworks fail to address a critical question: How can we ensure semantic alignment between augmented images and their associated textual descriptors in multimodal medical data? Most methods either focus solely on images, ignore modality interactions, or require extensive architectural modifications to segmentation networks. This limits their practicality, as medical imaging pipelines often rely on standardized, pre-trained models (e.g., *U-Net* [13]) that cannot be easily redesigned for modality-aware augmentation.

Table 1. Dice (%) for QaTa-Covid19 Segmentation in different augmentation modality.

Model	Modalities	Augmentation Modalities			
		None	Image	Text	Image+Text
SwinUnet	Image	71.08	74.01	–	–
LViT	Image+Text	76.98	76.70	76.60	76.89

To bridge this gap, we propose Region-Based Text-Consistent Augmentation (RBTCA), a novel, plug-and-play framework designed to harmonize multimodal augmentations while preserving semantic consistency. Unlike prior work, RBTCA operates as a lightweight preprocessing module, requiring no architectural changes to existing segmentation networks. Our approach leverages text-guided regional cues to align augmentations across modalities: (1) it identifies image regions (e.g., anatomical structures) corresponding to textual descriptions (e.g., "right lung"), (2) embeds these cues into the image to create a modality-aware representation (MAR), and (3) applies data augmentations to this modality-aware representation. This ensures that transformations, such as

geometric or intensity adjustments (e.g., flipping, noise injection), are consistently applied to both the image and its embedded textual cues, thus eliminating potential semantic mismatches. The key contributions of this work are threefold:

1) **A Plug-and-Play multimodal augmentation framework:** We propose RBTCA, a novel data augmentation strategy specifically designed for multimodal medical data, enabling seamless integration with diverse architectures (e.g., CNNs, Transformers) without structural modifications.
2) **Effortless multimodal consistency with standard augmentation:** RBTCA achieves multimodal consistency by directly leveraging off-the-shelf image augmentation methods, eliminating the need for modality-specific redesign.
3) **Open-source implementation and empirical validation**: RBTCA will be released as an open-source toolkit for adoption and reproducibility. Empirical validation across diverse architectures and datasets (QaTa-COV19, LTCT) demonstrates significant performance gains over baselines, validating its effectiveness and generalizability for multimodal medical image segmentation.

2 Methods

The RBTCA framework, visually detailed in Fig. 1 (Part A), achieves text-consistent data augmentation for multimodal medical image segmentation through a streamlined process. Initially, the framework maps a textual description T into a spatially aligned textual prompt $C_T \in \mathbb{R}^{1\times H\times W\times (D)}$. After that, with C_T's integration into the framework, it identifies a text-guided region of interest (ROI) probability map $R_I \in [0,1]^{1\times H\times W\times (D)}$ to precisely localize the image area semantically corresponding to the textual description. Following this, the framework integrates the textual prompt C_T and ROI R_I with the original medical image $I \in \mathbb{R}^{C\times H\times W\times (D)}$, generating a MAR $I_{MAR} \in \mathbb{R}^{C'\times H\times W\times (D)}$ that incorporates text-derived cues. Subsequently, image augmentation techniques $\mathcal{A}$ are applied to I_{MAR} and its label to produce augmented data $(I'_{MAR}, label') = (\mathcal{A}(I_{MAR}), \mathcal{A}(label))$, which is then used to train a segmentation model S_θ, learning a mapping $S_\theta(I'_{MAR}) \rightarrow pred'$ from augmented cue-enhanced images to predicted segmentation masks.

2.1 Text-to-Image Space Mapping and Region of Interest Selection

This stage aims to derive two key components from the input textual description T and the medical image I to a textual prompt in the image spatial space, C_T, and a text-guided ROI, R_I.

Textual Prompt Generation ($T \rightarrow C_T$). To generate the textual prompt C_T from the input textual description T, we first tokenize T using a pre-trained tokenizer, $Token(\cdot)$, converting the text into a sequence of tokens. Next, we utilize a pre-trained BERT-base-uncased model with its parameters unfrozen, $E_{BERT}(\cdot)$,

trained for our task, to encode the tokenized sequence into a contextualized text embedding, $\mathbf{e}_T \in \mathbb{R}^{L \times d_T}$. The text embedding can be computed as:

$$\mathbf{e}_T = E_{BERT}(Token(T)), \tag{1}$$

where d_T is the dimensionality of the BERT embedding, and L is the sequence length of the tokenized text. To map this text embedding into a spatially structured textual prompt, we employ an Embedding Mapping Block, $M_{EMB}(\cdot) : \mathbb{R}^{L \times d_T} \rightarrow \mathbb{R}^{1 \times H \times W \times (D)}$. Specifically, M_{EMB} comprises three sequential layers followed by a Multilayer Perceptron layer at the end. Each layer within M_{EMB} consists of: a 1-dimensional convolutional layer operating across the text embedding dimension (d_T), followed by another 1-dimensional convolutional layer operating across the sequence length dimension (L), Batch Normalization, and ReLU activation. We then reshape the output of M_{EMB} to obtain the textual prompt $C_T \in \mathbb{R}^{1 \times H \times W \times (D)}$:

$$C_T = M_{EMB}(\mathbf{e}_T). \tag{2}$$

Text-Guided Region of Interest Selection $((I, C_T) \rightarrow R_I)$. To extract the text-guided ROI R_I, we employ a Region of Interest Extraction Module, $M_{ROIE}(I, C_T)$, which takes both the medical image I and the textual prompt C_T as inputs. In the implementation of M_{ROIE}, both the medical image I and the textual prompt C_T are first patchified into patch sequences and then projected into Query (Q), Key (K), and Value (V) representations with learnable linear layers. Thus, for the image I and textual prompt C_T, we obtain query, key, and value representations as (Q_I, K_I, V_I) and $(Q_{C_T}, K_{C_T}, V_{C_T})$, respectively.

Subsequently, we utilize a cross-attention mechanism based on Agent attention [9]. Specifically, the average-pooled image query, $A = AvgPool(Q_I)$, acts as the "Agent" matrix. The region of interest R_I is then computed as:

$$R_I = M_{ROIE}(I, C_T) = \sigma(\sigma(AQ_{C_T}^T)\sigma(AK_{C_T}^T)V_{C_T}), \tag{3}$$

where σ denotes the sigmoid activation function. M_{ROIE} allows the textual prompt C_T to attend to spatially relevant locations in the image I, highlighting the region R_I that is semantically most relevant to the text description T. The output $R_I \in [0, 1]^{H \times W \times (D)}$ indicating the region of interest.

2.2 Text-Image Consistent Augmentation (TCA)

Following the previous stage, we proceed to create a MAR, I_{MAR}, by integrating the textual prompt C_T and the ROI R_I with the original medical image I. As C_T and R_I are spatially aligned with I, the cue-enhanced image I_{MAR} is readily generated through a simple channel-wise concatenation. Specifically, we concatenate the textual prompt C_T and the region of interest mask R_I as additional channels to the original medical image I.

$$I_{MAR} = I \oplus C_T \oplus R_I, \tag{4}$$

where $\oplus$ denotes channel-wise concatenation.

Consequently, we can directly apply standard image augmentation techniques to the cue-enhanced image I_{MAR} and its corresponding segmentation label, ensuring no conflict arises between the augmented image and its associated textual prompt. Specifically, we employ a set of common image augmentation transformations $\mathcal{A} = \{\text{Random Flip, Random Rotation, CutMix [21]}\}$. For each training sample $(I_{MAR}, label)$, a randomly selected subset $\mathcal{A}_{selected} \subseteq \mathcal{A}$ of these transformations is applied consistently to both I_{MAR} and $label$ to generate the augmented data pair $(I'_{MAR}, label')$. The augmentation process is formally represented as:

$$(I'_{MAR}, label') = \mathcal{A}_{selected}(I_{MAR}, label). \tag{5}$$

2.3 Segmentation Model Training and Loss Function

Our framework leverages two key components, the textual prompt C_T and ROI R_I, for modality-aware representation generation. To train these components, we employ Binary Cross-Entropy loss as auxiliary losses ($\mathcal{L}_{aux}$), applying it to both C_T and R_I with respect to the ground truth. This ensures effective learning of text-to-image space mapping and text-guided region identification.

For segmentation model training, we utilize the augmented modality-aware representation I'_{MAR} and their labels. We employ a standard composite loss function for medical image segmentation, $\mathcal{L}_{seg}$, which combines the Dice loss ($\mathcal{L}_{Dice}$) and the Cross-Entropy loss ($\mathcal{L}_{CE}$) with a weighting factor λ. The loss function can be represented as:

$$\mathcal{L}_{seg} = \lambda \mathcal{L}_{Dice} + (1 - \lambda)\mathcal{L}_{CE} \tag{6}$$

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \mathcal{L}_{aux}. \tag{7}$$

3 Experiments and Results

3.1 Experimental Setup

Datasets and Evaluation Metrics. To validate the effectiveness of RBTCA, we utilized two datasets: the QaTa-COV19 dataset and our in-house Lung Tumor CT Segmentation (LTCT) dataset. The QaTa-COV19 dataset, compiled by researchers from Qatar University and Tampere University, comprises 9258 chest X-ray radiographs from COVID-19 patients, featuring expert annotations of COVID-19 lesion regions and text annotations for multimodal research, as released by Li et al. in LViT [12]; these images have a resolution of 224×224 pixels and were used with the dataset splits from LViT. Additionally, our in-house LTCT dataset, consisting of 143 unenhanced CT scans from 95 patients, provides tumor region annotations and diagnostic reports. Originally at a 512×512 xy resolution with variable slice thicknesses, the LTCT dataset's CT images were preprocessed to achieve an isotropic spacing of $1\times1\times1\,\text{mm}^3$, axially cropped to 320 slices, windowed to [−700, 900] HU, and linearly normalized to [0, 1].

The LTCT dataset was randomly divided into training (120 scans) and testing (23 scans) sets. For quantitative evaluation, we employed the Dice Similarity Coefficient (Dice) and mean Intersection over Union (mIoU) metrics to assess segmentation performance.

Implementation Details. For image segmentation, we utilized six segmenters: unimodal *Unet* [13], *Swin-Unet* [1], *ConvNeXt* [17], and multimodal *LViT* [12], *ASDA* [20], *ReMamber* [19]. All models were trained in a fully supervised manner for 1000 epochs using the AdamW optimizer with a learning rate of 3e-4. To prevent overfitting, we employed early stopping, stopping training after 50 epochs without Dice improvement. The weighting factor λ in the loss function is set to 0.5 based on preliminary validation ($\lambda \in \{0.1,0.5,0.9\}$), and we found that $\lambda = 0.5$ balanced Dice and CE losses and yielded strong performance. We used batch sizes 64 for QaTa-Covid19. And for the LTCT dataset (3D CT volumes), we used a batch size of 1 during RBTCA modules' training, and 64 for segmenter training on xy-planar slices for segmenter training. To validate the effectiveness of RBTCA in 3D segmentation, we also employed *3D U-Net* [3] as the segmenter on the LTCT dataset with a batch size of 1.

3.2 Results of Segmentation Performance Evaluation

Comparison with Baseline Methods. To evaluate our module's effectiveness, we compared segmentation performance against baseline models. We first established baselines by assessing six segmenters on both datasets: unimodal *Unet*, *Swin-Unet*, *ConvNeXt*, and multimodal *LViT*, *ASDA*, *ReMamber*. Then, we evaluated performance with our module integrated into these baselines to demonstrate improvements. Detailed results are in Table 2.

The results show that integrating RBTCA significantly improved both unimodal and multimodal segmenters on QaTa-Covid19 Dataset. *SwinUnet* showed the largest gain, with Dice increasing by 7.24% to 78.32% and mIoU by 8.39%. *LViT* achieved the best performance, reaching a Dice of 81.19%. On the LTCT dataset, *Unet* with RBTCA showed a substantial Dice increase from 72.16% to 80.54%. Furthermore, using *3D U-Net* as a segmenter also benefited from RBTCA, with Dice improving from 64.67% to 71.08%. Some representative visualized results on the QaTa-Covid19 Dataset are shown in Fig. 2.

Ablation Study. To assess each component's contribution, we performed an ablation study. Table 2 shows segmentation performance with incremental addition of MAR and TCA. For unimodal models, both MAR and TCA improved performance. For example, *Unet* saw a 2.98% Dice increase with MAR and a further 1.53% with TCA. For multimodal models, MAR significantly improved performance (e.g., *LViT* by 4.21% Dice). However, TCA slightly decreased performance for multimodal models due to potential text-image feature mismatch. That could be explain as TCA's consistency fusion not aligned with original

Table 2. Results on the QaTa-Covid19 and Lung Tumor CT Dataset. MAR: Modality-Aware Representation; TCA: Text-Consistent Augmentation.

Backbone	Text	Modules		QaTa-Covid19		Lung Tumor CT	
		MAR	TCA	mIoU (%) ↑	Dice (%) ↑	mIoU (%) ↑	Dice (%) ↑
U-Net	×	×	×	75.56 ± 0.22	64.89 ± 0.24	72.16 ± 0.23	60.66 ± 0.21
	✓	✓	×	78.54 ± 0.18	67.86 ± 0.22	77.47 ± 0.19	66.23 ± 0.23
	✓	✓	✓	$\mathbf{80.09 \pm 0.18}$	$\mathbf{69.98 \pm 0.21}$	$\mathbf{80.54 \pm 0.23}$	$\mathbf{70.08 \pm 0.25}$
SwinUnet	×	×	×	71.08 ± 0.23	59.57 ± 0.25	66.25 ± 0.21	52.61 ± 0.21
	✓	✓	×	77.96 ± 0.19	67.15 ± 0.22	71.31 ± 0.22	57.34 ± 0.23
	✓	✓	✓	$\mathbf{78.32 \pm 0.20}$	$\mathbf{67.96 \pm 0.23}$	$\mathbf{72.82 \pm 0.19}$	$\mathbf{59.65 \pm 0.22}$
ConvNext	×	×	×	73.99 ± 0.21	62.70 ± 0.24	71.80 ± 0.24	58.84 ± 0.23
	✓	✓	×	77.96 ± 0.19	67.15 ± 0.22	72.01 ± 0.21	59.45 ± 0.22
	✓	✓	✓	$\mathbf{78.32 \pm 0.20}$	$\mathbf{67.96 \pm 0.23}$	$\mathbf{74.53 \pm 0.20}$	$\mathbf{62.03 \pm 0.23}$
LViT	✓	×	×	76.98 ± 0.21	66.43 ± 0.23	74.85 ± 0.17	63.73 ± 0.20
	✓	✓	×	$\mathbf{81.19 \pm 0.17}$	$\mathbf{71.23 \pm 0.20}$	$\mathbf{81.08 \pm 0.18}$	$\mathbf{71.54 \pm 0.21}$
	✓	✓	✓	80.23 ± 0.18	70.21 ± 0.21	80.11 ± 0.17	70.44 ± 0.23
ASDA	✓	×	×	77.61 ± 0.19	66.65 ± 0.21	74.39 ± 0.18	62.96 ± 0.22
	✓	✓	×	$\mathbf{78.15 \pm 0.18}$	66.92 ± 0.21	$\mathbf{75.36 \pm 0.17}$	$\mathbf{63.79 \pm 0.21}$
	✓	✓	✓	78.09 ± 0.19	$\mathbf{67.21 \pm 0.21}$	74.98 ± 0.18	62.87 ± 0.22
ReMamber	✓	×	×	79.12 ± 0.18	68.67 ± 0.21	79.57 ± 0.18	68.44 ± 0.21
	✓	✓	×	$\mathbf{81.03 \pm 0.17}$	$\mathbf{70.91 \pm 0.20}$	$\mathbf{80.33 \pm 0.19}$	$\mathbf{71.06 \pm 0.22}$
	✓	✓	✓	79.24 ± 0.19	69.99 ± 0.22	80.01 ± 0.20	70.81 ± 0.23
3D Unet	×	×	×	–	–	64.67 ± 0.24	52.28 ± 0.26
	✓	✓	×	-	-	65.40 ± 0.23	52.46 ± 0.24
	✓	✓	✓	–	–	$\mathbf{71.08 \pm 0.21}$	$\mathbf{62.21 \pm 0.23}$

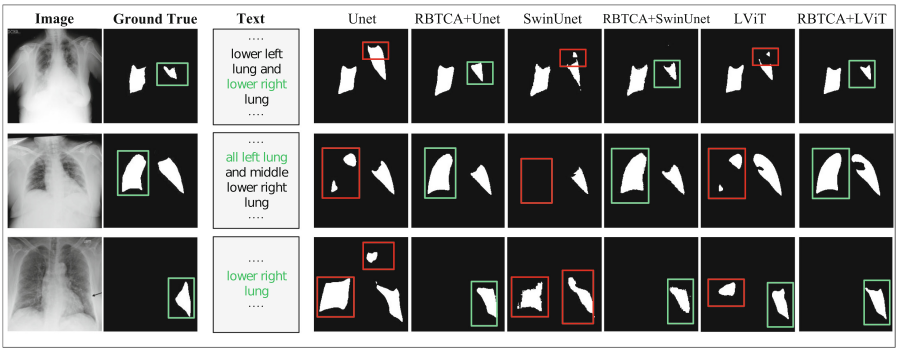


Fig. 2. Visualized result on the QaTa-Covid19 Dataset.

text, which by default inputs to the multimodal segmenter, negatively impacting performance. Thus, we recommend MAR + TCA for unimodal integration and primarily MAR for multimodal integration.

4 Conclusion

This work introduces RBTCA, a novel framework addressing multimodal medical image data augmentation. RBTCA, embeddable by design, achieves coherent augmentation via modality-aware representation and inherent text-image consistency. Key contributions are: (1) a plug-and-play framework for seamless integration, (2) broadly applicable multimodal consistency via standard image augmentation, and (3) significant, generalizable performance gains in segmentation. Extensive evaluations on diverse datasets validate RBTCA's effectiveness in enhancing segmentation accuracy. In summary, RBTCA presents a practical, effective approach to robust multimodal augmentation, contributing to improved deep learning model training and potentially enhanced clinical utility. Future research will explore RBTCA's versatility across medical imaging tasks and modalities.

Acknowledgments. This research was supported by Science and Technology Development Fund of Macao (0041/2023/RIB2), Macao Polytechnic University Grant (RP/FCA-15/2022), National Natural Science Foundation of China under Grant No. 62201323, Zhejiang Provincial Natural Science Foundation of China under Grant No. LDT23F01015F01 and No. LD25F020005, and Basic Research Program of Jiangsu No. BK20220266.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cao, H., et al.: Swin-unet: unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25066-8_9
2. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D u-net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
4. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
5. Duan, Y., et al.: 3MT-net: a multi-modal multi-task model for breast cancer and pathological subtype classification based on a multicenter study. IEEE J. Biomed. Health Inf. (2024)

6. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing **321**, 321–331 (2018)
7. Gao, Y., et al.: Improving neoadjuvant therapy response prediction by integrating longitudinal mammogram generation with cross-modal radiological reports: a vision-language alignment-guided model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 133–143. Springer, Cham (2024)
8. Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.: Deep learning-based image segmentation on multimodal medical imaging. IEEE Trans. Radiat. Plasma Med. Sci. **3**(2), 162–169 (2019)
9. Han, D., et al.: Agent attention: on the integration of softmax and linear attention. In: European Conference on Computer Vision, pp. 124–140. Springer, Cham (2024)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
11. Li, B., Hou, Y., Che, W.: Data augmentation approaches in natural language processing: a survey. Ai Open **3**, 71–90 (2022)
12. Li, Z., Li, Y., Li, Q., et al.: Lvit: language meets vision transformer in medical image segmentation. IEEE Trans. Med. Imaging **43**(1), 96–107 (2023)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019)
15. Shorten, C., Khoshgoftaar, T.M., Furht, B.: Text data augmentation for deep learning. J. Big Data **8**(1), 101 (2021)
16. Wang, S., et al.: Multi-scale context-guided deep network for automated lesion segmentation with endoscopy images of gastrointestinal tract. IEEE J. Biomed. Health Inform. **25**(2), 514–525 (2020)
17. Woo, S., et al.: Convnext v2: co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16133–16142 (2023)
18. Xiong, X., et al.: Distance guided generative adversarial network for explainable medical image classifications. Comput. Med. Imaging Graph. **118**, 102444 (2024)
19. Yang, Y., Ma, C., Yao, J., et al.: Remamber: referring image segmentation with mamba twister. In: European Conference on Computer Vision, pp. 108–126. Springer, Cham (2024)
20. Yue, P., et al.: Adaptive selection based referring image segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 1101–1110 (2024)
21. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)
22. Zhang, Z., et al.: UniMRISegNet: universal 3D network for various organs and cancers segmentation on multi-sequence MRI. IEEE J. Biomed. Health Inf. (2024)

23. Zhang, H., Cisse, M., Dauphin, Y.N., et al.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
24. Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V.: Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8543–8553 (2019)



ReSeg-UNet: A Reconstruction-Guided Optimization Framework for Enhanced Medical Image Segmentation

Lin Li[1], Dong Tang[1(✉)], Xiaowen Chu[2], Xiaofei Yang[1], and Fei Yu[3(✉)]

[1] School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China
tangdong@gzhu.edu.cn

[2] The Hongkong University of Science and Technology (Guangzhou), Duxue Road 1, Guangzhou, China

[3] School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China
yufei_hitcs@163.com

Abstract. Medical image segmentation is critical for accurate diagnosis; however, the task remains challenging due to the inherent ambiguities in low-contrast anatomical boundaries and the presence of extensive redundant features in the skip connections of segmentation models. To address these limitations, we propose ReSeg-UNet, a novel two-stage framework that synergizes image reconstruction with segmentation optimization. In the first stage, a composite reconstruction loss-combining Mean Squared Error (MSE) and L1 regularization is applied to a standard segmentation network, generating stable reconstruction weights that encode multi-scale feature representations. These weights explicitly capture both global anatomical context and local boundary details. In the second stage, a three-level cross-feature alignment mechanism is introduced: the encoder of the reconstruction model is aligned with the decoder of the segmentation model, the decoder of the former is aligned with the encoder of the latter, and the intermediate features of both models are also aligned. This strategy ensures multi-level feature consistency during downsampling, intermediate layers, and upsampling, effectively mitigating information loss in blurred regions. Extensive experiments on the Synapse (abdominal CT) and ACDC (cardiac MRI) datasets demonstrate significant improvements. Our code is available at https://github.com/Li-gzhu/ReSeg-UNet.git.

Keywords: Reconstruction-guided Segmentation · Medical Image Segmentation · Three-Level Cross Feature Alignment

1 Introduction

Medical image segmentation is essential for modern computer-aided diagnosis (CAD), providing pixel-level localization of anatomical structures and patholo-

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 544–554, 2026.
https://doi.org/10.1007/978-3-032-04947-6_52

gies [2,25,30]. Deep learning methods, like U-Net [23], drove significant advancements in this field due to their powerful feature extraction ability. However, most approaches are built with CNN, which may struggle to model long-range dependencies because of their fixed receptive fields [14,22]. This is a significant drawback in complex scenarios such as multi-organ CT segmentation or diffuse lesion delineation in MRI, where global anatomical relationships are crucial for accurate results. To address limitation, some researchers introduced Vision Transformers (ViTs) [7] to the medical image segmentation task, using self-attention mechanisms to capture global representation [10,12]. For instance, TransUNet [5] combines CNNs and ViTs for joint local-global feature learning in medical image classification. While Transformers excel at modeling distant voxel interactions, their quadratic computational complexity hinders deployment on high-resolution 3D medical volumes, particularly in resource-constrained clinical environments.

Recent advancements in state space models (SSMs) [9], such as Mamba [8], leverage the core design principle of utilizing SSMs to capture global dependencies while maintaining linear computational complexity [17,19,29,32,34], making them highly effective for processing large-scale data. However, in the field of medical image segmentation, where extremely high segmentation accuracy is required, their ability to capture fine-grained local features remains somewhat limited [20,28]. Additionally, numerous studies [21,26,27,33] have demonstrated that the skip connections in U-shaped architectures merely concatenate features from corresponding encoder and decoder layers, often introducing significant feature redundancy, which subsequently adversely affects segmentation performance.

In current medical image segmentation tasks, the inherent ambiguities in low-contrast anatomical boundaries and imaging artifacts (e.g., metal artifacts in CT or motion artifacts in MRI) often lead to inadequate capture of local features by traditional methods, resulting in semantic information loss. In the domain of natural image processing, existing studies have addressed this issue by introducing cross-level knowledge transfer mechanisms (e.g., feature distillation and spatial alignment) to mitigate semantic feature loss [4,6,16].

Inspired by advancements in natural image processing and the aforementioned limitations in medical image segmentation, we propose a two-stage optimization framework for medical image segmentation, termed ReSeg-UNet. Specifically, in the first stage, the framework performs an image reconstruction task by inputting the ground truth into a traditional segmentation network, generating reconstruction weights that encapsulate rich semantic information. In the second stage, the segmentation task leverages these stable weights from the first stage and employs a three-level cross-feature alignment mechanism (see Fig. 1) to guide and optimize the learning of local features in the segmentation model. Through this approach, the segmentation model is able to learn finer-grained semantic features while alleviating the issue of feature redundancy introduced by skip connections, thereby significantly improving segmentation performance. Moreover, the additional computational overhead introduced by our proposed optimization method in the segmentation task is minimal, as it is

confined solely to the calculation of the feature alignment loss. This results from our decoupled training strategy, where the first and second stages are trained independently, and the reconstruction weights in the second stage are kept frozen (i.e., no parameter updates are performed).

Contribution: 1) We innovatively propose a two-stage network framework that leverages image reconstruction to optimize medical image segmentation. 2) We design a composite reconstruction loss function to generate stable multi-scale feature weights, and furthermore, we propose a three-level cross-feature alignment mechanism to optimize local feature learning for the segmentation task. 3)The proposed method can be seamlessly integrated into existing U-shaped frameworks (e.g., U-Net, TransUNet, Swin-UNet, VM-UNet) with minimal additional cost, delivering performance gains effectively.

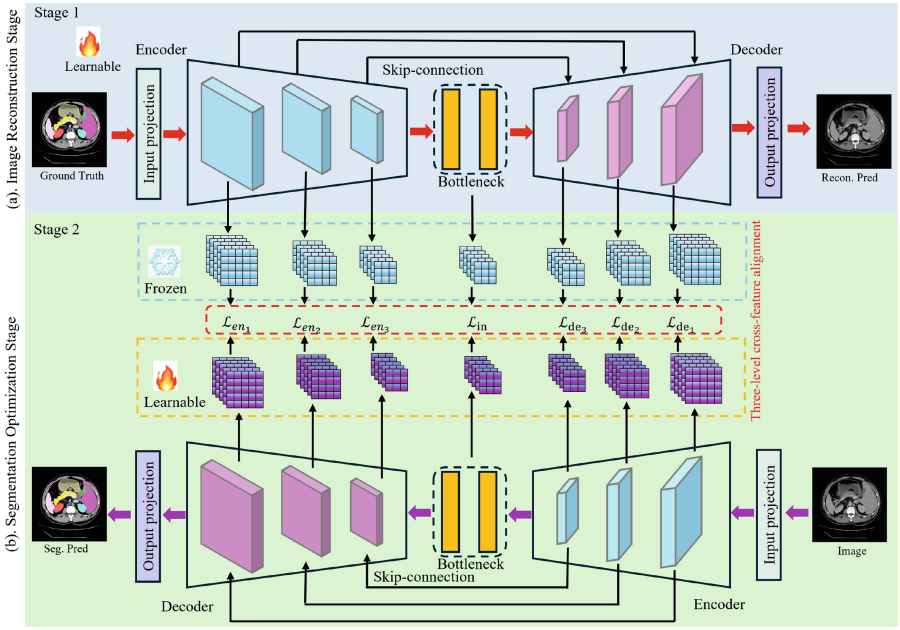


Fig. 1. Overview of ReSeg-UNet. (a) The ground truth is fed into a standard UNet segmentation network to perform image reconstruction, thereby obtaining stable reconstruction weights. (b) By utilizing the frozen reconstruction weights and adopting a three-level cross-feature alignment mechanism, the segmentation task is guided to learn more refined local features.

2 Method

2.1 Overview of ReSeg-UNet

ReSeg-UNet is a dual-stage optimization network tailored for medical image segmentation, as shown in Fig. 1. Current mainstream models, based on CNNs, Transformers, and Mamba, typically adopt a U-shaped structure comprising an encoder, bottleneck layer, and decoder. In the first stage, a standard segmentation network performs a reconstruction task using ground truth to generate stable, semantically rich weights. In the second stage, these weights guide the segmentation task through a three-level cross-feature alignment mechanism, enhancing local feature learning for robust and accurate segmentation. Below, we detail the ReSeg-UNet framework.

2.2 Stage 1: Image Reconstruction

Figure 1(a) illustrates the detailed architecture of the image reconstruction process. First, while keeping the model architecture unchanged, we perform the image reconstruction task by using the ground truth $\hat{x} \in \mathbb{R}^{H\times W\times C}$ (where H, W, and C denote the height, width, and number of channels of the image, respectively) as input, thereby obtaining stable reconstruction weights. To this end, we designed a composite loss function comprising the MSE loss and L1 regularization loss, which are employed to evaluate the global structural consistency and local detail restoration capabilities of the reconstructed image, respectively. The MSE loss is employed to quantify the pixel-level differences between the reconstructed image and the ground truth image, while the L1 loss is utilized to enhance the local detail restoration capability of the reconstructed image, particularly in preserving edge and texture features. The detailed formulas are as follows:

$$\mathcal{L}_{\text{MSE}}(x,\hat{x}) = \frac{1}{H\times W\times C}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}(x_{i,j,k}-\hat{x}_{i,j,k})^2 \tag{1}$$

$$\mathcal{L}_{\text{L1}}(x,\hat{x}) = \frac{1}{H\times W\times C}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}|x_{i,j,k}-\hat{x}_{i,j,k}| \tag{2}$$

where x represents the reconstructed predicted image and $\hat{x}$ denotes the ground truth image, the final reconstruction loss is formulated as the weighted sum of the MSE loss and the L1 loss, as shown in the following equation:

$$\mathcal{L}_{\text{Recon}}(x,\hat{x}) = \alpha\cdot\mathcal{L}_{\text{MSE}}(x,\hat{x}) + \beta\cdot\mathcal{L}_{\text{L1}}(x,\hat{x}) \tag{3}$$

Here, α and β are weighting coefficients used to balance the contributions of the two loss terms, with $\alpha = 0.6$ and $\beta = 0.4$.

2.3 Stage 2: Segmentation Optimization

Figure 1(b) illustrates the detailed architecture of the three-level cross-feature alignment mechanism in the second stage. This stage leverages the image reconstruction weights trained in the first stage to guide the optimization of the segmentation task. The core objective of the second stage is to achieve feature alignment between the reconstruction task and the segmentation task, ensuring that the segmentation network can fully reuse the anatomical features learned in the first stage. Specifically, the three-level cross-feature alignment mechanism consists of the following three levels: Feature alignment between the reconstruction encoder and the segmentation decoder (Recon-E → Seg-D): For each level l, the alignment loss is defined as:

$$\mathcal{L}_{\text{Align}}^{\text{Recon-E}\rightarrow\text{Seg-D}}(l) = \frac{1}{H_l \times W_l \times C_l} \left\| F_{\text{Recon-E}}^{l} - F_{\text{Seg-D}}^{l} \right\|^2 \tag{4}$$

Here, $F_{\text{Recon-E}}^{l}$ denotes the feature at the l-th layer of the reconstruction task encoder, and $F_{\text{Seg-D}}^{l}$ represents the feature at the l-th layer of the segmentation task decoder. H_l, W_l, and C_l denote the spatial dimensions and the number of channels of the feature maps, respectively. The alignment loss between Recon-D and Seg-E is defined as:

$$\mathcal{L}_{\text{Align}}^{\text{Recon-D}\rightarrow\text{Seg-E}}(l) = \frac{1}{H_l \times W_l \times C_l} \left\| F_{\text{Recon-D}}^{l} - F_{\text{Seg-E}}^{l} \right\|^2 \tag{5}$$

Here, $F_{\text{Recon-D}}^{l}$ denotes the feature at the l-th layer of the reconstruction task decoder, and $F_{\text{Seg-E}}^{l}$ represents the feature at the l-th layer of the segmentation task encoder. For the bottleneck layer (intermediate features), the alignment loss is defined as:

$$\mathcal{L}_{\text{Align}}^{\text{in}} = \frac{1}{H_{\text{in}} \times W_{\text{in}} \times C_{\text{in}}} \left\| F_{\text{Recon-in}} - F_{\text{Seg-in}} \right\|^2 \tag{6}$$

Here, $F_{\text{Recon-in}}$ denotes the feature of the bottleneck layer in the reconstruction task, and $F_{\text{Seg-in}}$ represents the feature of the bottleneck layer in the segmentation task. The total feature alignment loss is the average of the aforementioned three-level alignment losses, formulated as:

$$\mathcal{L}_{\text{Align}} = \frac{1}{3} \left(\sum_{l=1}^{3} \left(\mathcal{L}_{\text{Align}}^{\text{Recon-E}\rightarrow\text{Seg-D}}(l) + \mathcal{L}_{\text{Align}}^{\text{Recon-D}\rightarrow\text{Seg-E}}(l) \right) + \mathcal{L}_{\text{Align}}^{\text{in}} \right) \tag{7}$$

Optimization Objective Function: The total loss function combines $\mathcal{L}_{\text{Align}}$ with the original segmentation loss function $\mathcal{L}_{\text{Seg}}$, which employs both cross-entropy and Dice loss. The formula is as follows:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Seg}} + \lambda \cdot \mathcal{L}_{\text{Align}} \tag{8}$$

Here, λ is the weighting coefficient for the feature alignment loss, which controls the contribution of feature alignment to the segmentation task. The value of λ is set to 0.035.

Table 1. Comparison with SOTA methods on Synapse dataset. Δ denotes the improvement gain (%) by comparing with the original method.

Model	DSC(%)	HD	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
U-Net [23](MICCAI'15)	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
U-Net++ [31] (MICCAI'18)	76.91	36.93	88.19	68.89	81.76	75.27	93.01	58.20	83.44	70.52
TransUNet [5](arxiv'21)	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
UCTransNet [27](AAAI'22)	78.23	26.75	84.25	64.65	82.35	77.65	94.36	58.18	84.74	79.66
Swin-UNet [3](ECCV'22)	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
HiFormer [12](WACV'23)	80.69	19.14	87.03	68.61	84.23	78.37	94.07	60.77	90.44	82.03
VM-UNet [24](arxiv'24)	81.08	19.21	86.40	69.41	86.16	82.76	94.17	58.80	89.51	81.40
H2Former [11](TMI'23)	82.16	18.79	87.03	69.78	86.38	83.49	94.71	64.49	**90.69**	80.73
MISSFormer [13](TMI'23)	81.96	18.20	87.71	63.86	87.97	82.80	94.68	60.90	90.60	82.27
Swin-UMamba [18] (MICCAI'24)	82.57	16.21	88.40	70.41	87.16	**83.76**	95.17	**64.80**	89.51	81.40
U-Net (our)	78.86	27.68	88.60	71.02	82.24	70.21	94.78	62.23	88.60	73.20
Δ	+2.01	-12.02	-0.47	+1.30	+4.47	+1.61	+1.35	+8.25	+1.93	-2.38
TransUNet (our)	79.65	26.59	89.20	64.21	81.23	76.58	94.60	62.38	87.33	81.67
Δ	+2.17	-5.10	+1.97	+1.08	-0.64	-0.44	+0.52	+6.52	+2.25	+6.05
Swin-UNet (our)	80.58	18.09	86.36	67.51	87.96	79.26	95.26	60.35	89.66	78.28
Δ	+1.45	-3.46	+0.89	+0.98	+4.68	-0.35	+0.97	+3.77	-1.00	+1.68
VM-UNet (our)	**83.22**	**15.73**	**89.62**	**71.21**	**88.02**	81.36	**95.76**	63.36	90.61	**85.82**
Δ	+2.14	-3.48	+3.22	+1.80	+1.86	-1.40	+1.59	+4.56	+1.10	+4.42

3 Experiments and Results

3.1 Dataset

Synapse Dataset [15]**:** The Synapse dataset comprises 30 scans of eight abdominal organs: left and right kidneys, aorta, spleen, gallbladder, liver, stomach, and pancreas. It contains 3,779 clinically enhanced abdominal CT images in axial view. The dataset is divided into 18 samples for training and 12 for testing. We report Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD95) as evaluation metrics for this dataset.

ACDC Dataset [1]**:** The ACDC dataset contains 100 cardiac MRI scans acquired from diverse clinical patients, with pixel-level annotations for three cardiac substructures: left ventricle (LV), right ventricle (RV), and myocardium (MYO). Following the standard experimental protocol of baseline models, the dataset was partitioned into 70 cases (1,930 axial slices) for training, 10 cases for validation, and 20 cases for testing. We evaluate our method using the DSC as the evaluation metric.

3.2 Experiment Settings

We evaluated the effectiveness of our proposed loss function on U-Net [23], TransUNet [5], Swin-UNet [3], and VM-UNet [24] using both the Synapse and ACDC datasets. The training configurations (i.e., batch size, optimizer, learning rate, etc.) were kept consistent with those of the baseline models. For all experiments, the input image size was set to 224×224, and the data augmentation and preprocessing steps were identical to those used in the baseline models. Training was conducted on an Nvidia GTX 4090 GPU with 24 GB of memory. In line

Table 2. Comparison of different methods in ACDC dataset.

Model	DSC(%)	RV	Myo	LV
U-Net [23](MICCAI'15)	89.57	85.81	87.47	95.42
TransUNet [5](arxiv'21)	89.71	86.67	87.27	95.18
Swin-UNet [3](ECCV'22)	90.00	88.55	85.62	95.73
HiFormer [12](TMI'23)	90.82	88.55	88.44	94.47
MISSFormer [13](TMI'23)	91.19	89.85	88.38	95.34
VM-UNet [24](arxiv'24)	90.56	88.77	87.89	95.02
Swin-UMamba [18](MICCAI'24)	91.53	90.01	88.98	95.60
U-Net (our)	91.21	88.61	**89.73**	95.29
Δ	+1.64	+2.80	+2.26	-0.13
TransUNet (our)	91.83	**90.60**	89.25	95.64
Δ	+2.12	+3.93	+1.98	+0.54
Swin-UNet (our)	91.24	89.46	88.59	**95.74**
Δ	+1.24	+0.91	+2.97	+0.01
VM-UNet (our)	**91.85**	90.30	89.58	95.67
Δ	+1.29	+1.53	+1.69	+0.65

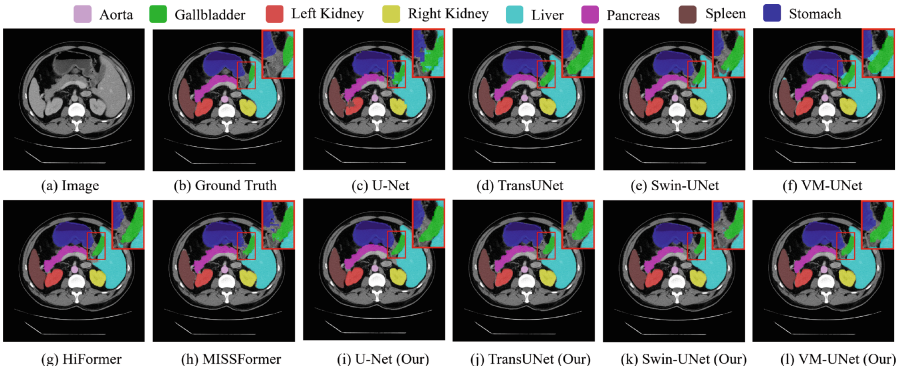


Fig. 2. Result visualization on Synapse dataset.

with the literature, TransUNet, Swin-UNet, and VM-UNet utilized pre-trained weights from ImageNet, while U-Net was trained from scratch.

3.3 Results

The experimental results are presented in Table 1 for the Synapse dataset and Table 2 for the ACDC dataset. From these results, it is evident that our proposed optimization method for medical image segmentation baseline models is highly effective, delivering significant performance improvements. Specifically, on the

Synapse dataset, U-Net, TransUNet, Swin-UNet, and VM-UNet achieved average Dice Similarity Coefficient (DSC) improvements of 2.01%, 2.17%, 1.45%, and 2.14%, respectively. Moreover, the Hausdorff Distance (HD) was reduced by 12.02 mm, 5.10 mm, 3.46 mm, and 3.48 mm, respectively. For the segmentation accuracy of the eight organs, most organs showed significant improvements, especially the pancreas, where the four models achieved improvements of 8.25%, 6.25%, 3.77%, and 4.65%, respectively. On the ACDC dataset, the average DSC improved by 1.64%, 2.02%, 1.24%, and 1.29%, respectively. Further comparisons reveal that the optimized VM-UNet outperformed Swin-UMamba on the Synapse dataset, with an average DSC increase of 0.65% and an HD reduction of 0.48mm. On the ACDC dataset, the optimized VM-UNet also achieved a 0.32% higher average DSC than Swin-UMamba.

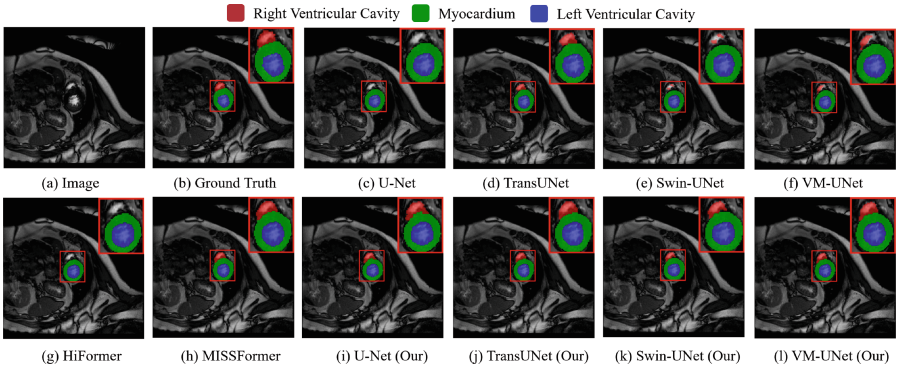


Fig. 3. Result visualization on ACDC dataset.

The optimization effects are more evident in the segmentation visualizations. In the Synapse dataset (see Fig. 2), the optimized model achieves higher segmentation accuracy than the baseline models, with outputs closer to the ground truth. In the ACDC dataset (see Fig. 3), the baseline model performs poorly in right ventricle segmentation, while the optimized model significantly improves edge precision. This validates the effectiveness of our method in enhancing local feature extraction for baseline models.

3.4 Ablation Studies

To further investigate our proposed method, we conducted a series of ablation studies on four optimized baseline models using the Synapse dataset. The results and analysis are as follows: In the ablation experiments on different loss functions (see Fig. 4(a)), the segmentation performance of the baseline models significantly improved after incorporating the three-level cross-feature alignment, achieving optimal performance when all three losses were utilized. In the ablation experiment on the feature alignment balancing factor λ (see Fig. 4(b)), we

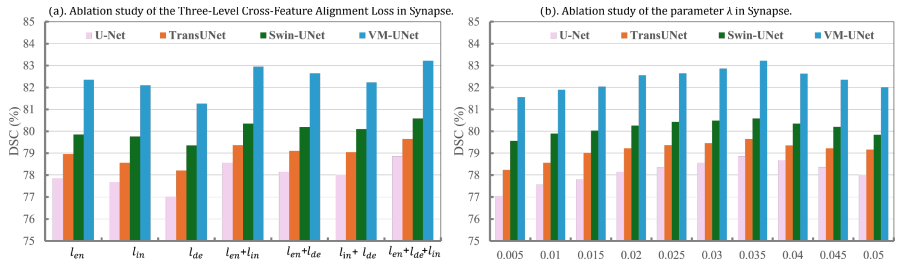


Fig. 4. Ablation study on the number of feature alignment loss functions and λ is conducted based on the Synapse dataset.

observed that the model reached its optimal performance when $\lambda = 0.035$. This demonstrates that proper feature guidance can effectively enhance segmentation performance.

4 Conclusion

In this paper, we introduce ReSeg-UNet, a dual-stage optimization framework for medical image segmentation that integrates image reconstruction with segmentation. By incorporating a composite reconstruction loss and a three-level cross-feature alignment mechanism, our approach effectively enhances the performance of current mainstream 2D medical image segmentation models based on CNNs, Transformers, and Mamba architectures. Specifically, it improves the segmentation capabilities for low-contrast boundaries and small anatomical structures, such as the pancreas. Extensive experiments on the Synapse (abdominal CT) and ACDC (cardiac MRI) datasets demonstrate that ReSeg-UNet achieves significant performance improvements over baseline models and outperforms existing state-of-the-art methods. Future work will focus on extending this framework to 3D and multi-modal segmentation tasks.

Acknowledgments. This work was supported in part by the Doctoral Research Start-up Foundation of Liaoning University of Technology (Grant No. XB2025019)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Medical Imaging **37**(11), 2514–2525 (2018)
2. Bongratz, F., Rickmann, A., Wachinger, C.: Neural deformation fields for template-based reconstruction of cortical surfaces from MRI. Med. Image Anal. **93**, 103093 (2024)

3. Cao, H., et al.: Swin-unet: unet-like pure transformer for medical image segmentation. In: ECCV Workshops (3). Lecture Notes in Computer Science, vol. 13803, pp. 205–218. Springer, Heidelberg (2022)
4. Chen, D., et al.: Cross-layer distillation with semantic calibration. In: AAAI, pp. 7028–7036. AAAI Press (2021)
5. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. CoRR arxiv:2102.04306 (2021)
6. Chung, I., Park, S., Kim, J., Kwak, N.: Feature-map-level online adversarial knowledge distillation. In: ICML. Proceedings of Machine Learning Research, vol. 119, pp. 2006–2015. PMLR (2020)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR. OpenReview.net (2021)
8. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. CoRR arxiv:2312.00752 (2023)
9. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: ICLR. OpenReview.net (2022)
10. Hatamizadeh, A., et al.: UNETR: transformers for 3d medical image segmentation. In: WACV, pp. 1748–1758. IEEE (2022)
11. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: an Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation. IEEE Trans. Med. Imaging **42**(9), 2763–2775 (2023)
12. Heidari, M., et al.: Hiformer: hierarchical multi-scale representations using transformers for medical image segmentation. In: WACV, pp. 6191–6201. IEEE (2023)
13. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: an effective transformer for 2d medical image segmentation. IEEE Trans. Med. Imaging **42**(5), 1484–1494 (2023)
14. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
15. Landman, B., Xu, Z., Iglesias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proceedings of MICCAI, p. 12 (2015)
16. Li, L.: Self-regulated feature learning via teacher-free feature distillation. In: ECCV (26). Lecture Notes in Computer Science, vol. 13686, pp. 347–363. Springer, Heidelberg (2022)
17. Liao, W., Zhu, Y., Wang, X., Pan, C., Wang, Y., Ma, L.: Lightm-unet: mamba assists in lightweight unet for medical image segmentation. CoRR arxiv:2403.05246 (2024)
18. Liu, J., et al.: Swin-umamba: mamba-based unet with imagenet-based pretraining. In: MICCAI (9). Lecture Notes in Computer Science, vol. 15009, pp. 615–625. Springer, Heidelberg (2024)
19. Liu, Y., et al.: Vmamba: visual state space model. CoRR arxiv:2401.10166 (2024)
20. Ma, J., Li, F., Wang, B.: U-mamba: enhancing long-range dependency for biomedical image segmentation. CoRR arxiv:2401.04722 (2024)
21. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. CoRR arxiv:1804.03999(2018)
22. Rahman, M.M., Munir, M., Marculescu, R.: EMCAD: efficient multi-scale convolutional attention decoding for medical image segmentation. In: CVPR, pp. 11769–11779. IEEE (2024)

23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Ruan, J., Xiang, S.: Vm-unet: vision mamba unet for medical image segmentation. CoRR arxiv:2402.02491 (2024)
25. Shaker, A.M., et al.: UNETR++: delving into efficient and accurate 3d medical image segmentation. IEEE Trans. Med. Imaging **43**(9), 3377–3390 (2024)
26. Sun, G., et al.: Da-transunet: integrating spatial and channel dual attention with transformer u-net for medical image segmentation. CoRR arxiv:2310.12570 (2023)
27. Wang, H., Cao, P., Wang, J., Zaïane, O.R.: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: AAAI, pp. 2441–2449. AAAI Press (2022)
28. Wang, Z., Zheng, J., Zhang, Y., Cui, G., Li, L.: Mamba-unet: unet-like pure visual mamba for medical image segmentation. CoRR arxiv:2402.05079 (2024)
29. Wu, R., Liu, Y., Liang, P., Chang, Q.: Ultralight vm-unet: parallel vision mamba significantly reduces parameters for skin lesion segmentation. CoRR arxiv:2403.20035 (2024)
30. Xie, Y., He, X., Yang, B., Lyu, F., Liu, S.: Cam-guided translation for unpaired weakly-supervised medical image segmentation. In: ICME, pp. 1–6. IEEE (2024)
31. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
32. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: efficient visual representation learning with bidirectional state space model. In: ICML. OpenReview.net (2024)
33. Zhu, W., et al.: Selfreg-unet: self-regularized unet for medical image segmentation. In: MICCAI (8). Lecture Notes in Computer Science, vol. 15008, pp. 601–611. Springer, Cham (2024)
34. Zhu, Y., Zhang, D., Lin, Y., Feng, Y., Tang, J.: Merging context clustering with visual state space models for medical image segmentation. IEEE Trans. Med. Imaging (2025). https://doi.org/10.1109/TMI.2025.3525673



Reverse Imaging for Wide-Spectrum Generalization of Cardiac MRI Segmentation

Yidong Zhao[1], Peter Kellman[2], Hui Xue[3], Tongyun Yang[1], Yi Zhang[1], Yuchi Han[4], Orlando Simonetti[4], and Qian Tao[1(✉)]

[1] Department of Imaging Physics, Delft University of Technology, Delft, The Netherlands
q.tao@tudelft.nl
[2] National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, MD, USA
[3] Microsoft Research, Redmond, USA
[4] Cardiovascular Division, The Ohio State University Wexner Medical Center, Columbus, OH, USA

Abstract. Pretrained segmentation models for cardiac magnetic resonance imaging (MRI) struggle to generalize across different imaging sequences due to significant variations in image contrast. These variations arise from changes in imaging protocols, yet the same fundamental spin properties, including proton density, T_1, and T_2 values, govern all acquired images. With this core principle, we introduce *Reverse Imaging*, a novel physics-driven method for cardiac MRI data augmentation and domain adaptation to fundamentally solve the generalization problem. Our method reversely infers the underlying spin properties from observed cardiac MRI images, by solving ill-posed nonlinear inverse problems regularized by the prior distribution of spin properties. We acquire this "spin prior" by learning a generative diffusion model from the multiparametric SAturation-recovery single-SHot acquisition sequence (mSASHA) dataset, which offers joint cardiac T_1 and T_2 maps. Our method enables approximate but meaningful spin-property estimates from MR images, which provide an interpretable "latent variable" that lead to highly flexible image synthesis of arbitrary novel sequences. We show that Reverse Imaging enables highly accurate segmentation across vastly different image contrasts and imaging protocols, realizing wide-spectrum generalization of cardiac MRI segmentation.

Keywords: CMR Segmentation · Reverse Imaging · MR Physics

1 Introduction

Segmentation is crucial for evaluating cardiac biomarkers from cardiac magnetic resonance imaging (MRI), for example, the ejection fraction from balanced

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 555–565, 2026.
https://doi.org/10.1007/978-3-032-04947-6_53

steady-state free-precession (bSSFP) cine imaging [3,4]. Learning-based methods have become the standard practice for cardiac MRI segmentation [3,4,15,17], mostly trained on widely accessible bSSFP cine images with manual annotations [3,4,17]. However, models trained on bSSFP cine are prone to fail in test images with heterogeneous contrasts [4,17]. This can be partially mitigated by data augmentation (e.g. blur, gamma correction, etc.) [15], which effectively generalizes pretrained models to bSSFP cine images cross vendors and centers [4,6,27].

However, a much stronger contrast change comes from different MRI sequences, which cannot be tackled by common image augmentation techniques. For example, an alternative to bSSFP is the gradient echo sequence (GRE), which is less affected by field inhomogeneities [23] and more suitable for patients with implanted devices [10]. However, GRE has much poorer contrasts and easily fails the models trained on bSSFP. An even stronger contrast change occurs in quantitative imaging, such as T_1 mapping by the modified Look-Locker Inversion Recovery Sequence (MOLLI) [18]. In MOLLI, the image readouts have substantial contrast variations during the spin relaxation process, alternating between bright and dark blood, with low myocardium-blood contrast. Consequently, segmentation models trained on bSSFP images (source domain) can fail completely on MOLLI images (target domain) even with extensive data augmentation.

To tackle the cross-sequence generalization challenge, previous works have focused on disentangling "*content*" and "*style*" [19,20]. BayeSeg learns to remove the style and extracts the boundary for segmentation [11,12]. Domain adaptation techniques strive to learn content and style embeddings [5,28]. The learned "content" is supposed to be shared between sequences, while cross-sequence translation is achieved by interchanging the learned "style" [5,9,28]. However, such strategies require data from new sequences in order to learn the disentangling, and, importantly, the extracted content or style is barely interpretable. In this study, we argue that the fundamental but overlooked "content" is the underlying *spin properties*, including the voxel-wise proton density (PD), T_1, and T_2 [13] of blood and tissue. Given this content, the style is fully governed by the MRI signal models of new pulse sequences. This motivates us to infer the spin properties from the observed image, for more interpretable and physics-grounded image translation to tackle the cross-sequence domain generalization challenge.

Unlike quantitative MRI that directly maps spin properties [7,16,18], retrieving the underlying spin properties from a qualitative image such as bSSFP is ill-posed because infinitely many solutions may explain the same image. This challenge parallels that of accelerated MRI, where partial measurements are insufficient for image reconstruction, yet the inverse problem can be solved by incorporating a regularizer that defines the prior distribution of plausible images [21]. Recent MRI reconstruction works use generative diffusion models [14] as *a priori* regularizers, allowing high-quality reconstruction from partial measurements [8,26]. Following the same spirit, we introduce a diffusion-based generative prior on spin properties, called "spin prior", to regularize the ill-posed inverse problem of spin property estimation. Specifically, we lever-

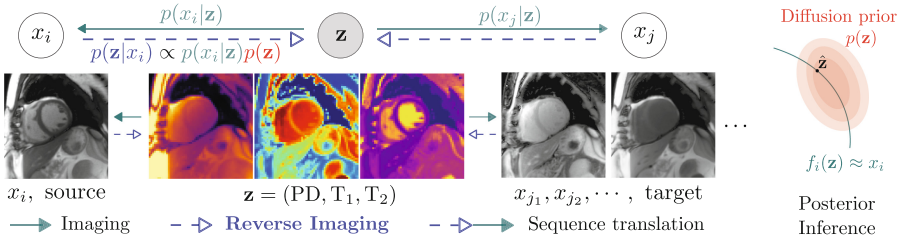


Fig. 1. Images acquired by sequences i and j from a subject share the same underlying spin properties $\mathbf{z}$. Reverse Imaging (dashed arrows) estimates $\hat{\mathbf{z}}$ from the prior $p(\mathbf{z})$ that explains the observed image x_i. Images of an arbitrary sequence can be generated by MR physics from $\hat{\mathbf{z}}$ (solid arrows). In this process, the only unknown is the prior $p(\mathbf{z})$, which we learn from the mSASHA data by a diffusion model (red).

age the multiparametric SAturation-recovery single-SHot acquisition (mSASHA) datasets, which perform joint cardiac T_1 and T_2 mapping, to learn the spin prior. We term the process of inferring spin properties from observed images as *Reverse Imaging*. Reverse Imaging uniquely enables physics-grounded data augmentation and domain adaptation without the need for target domain data. We will show that Reverse Imaging leads to high-quality zero-shot generalization to a wide spectrum of MRI sequences. We make the following contributions:

- We propose an interpretable cross-sequence translation framework for cardiac MRI, which explicitly formulates the translation as applying MRI physics forward models to estimated tissue spin properties (i.e. PD, T_1 and T_2).
- We introduce Reverse Imaging, a novel approach that reversely infers spin properties from observed MR images using a physics-guided reverse diffusion process. The diffusion model serves as a generative prior that regularizes the estimation of the underlying spin properties.
- With Reverse Imaging, we achieve high-quality zero-shot generalization of cardiac MRI segmentation, from bSSFP to a wide spectrum of unseen contrasts and imaging protocols.

2 Method

2.1 Physics-Based Cross-Sequence Translation

Physics-Prior Decomposition Let x_i and x_j be images acquired using sequences i and j, on which we aim to perform cross-sequence translation. The direct translation $p(x_j|x_i)$ from x_i to x_j is intractable. To solve this, we propose a decomposition of $p(x_j|x_i)$, by introducing an underlying physics $\mathbf{z} = (\mathrm{PD}, \mathrm{T}_1, \mathrm{T}_2)$ as a "latent variable" :

$$p(x_j|x_i) = \int_{\mathbf{z}} p(x_j, \mathbf{z}|x_i) d\mathbf{z} = \int_{\mathbf{z}} p(x_j|\mathbf{z}, x_i) p(\mathbf{z}|x_i) d\mathbf{z} = \int_{\mathbf{z}} p(x_j|\mathbf{z}) p(\mathbf{z}|x_i) d\mathbf{z}. \quad (1)$$

A probability graph model is shown in Fig. 1. Reverse Imaging can be formulated as posterior inference $p(\mathbf{z}|x_i)$, expressed as:

$$\log p(\mathbf{z}|x_i) \propto \log p(x_i|\mathbf{z}) + \log p(\mathbf{z}) + \text{Const.} \tag{2}$$

As a result of (1) and (2), the inference $p(x_i|x_j)$ relies on two likelihoods $p(x_i|\mathbf{z})$ and $p(x_j|\mathbf{z})$, which are directly accessible from the imaging physics forward model. The only missing piece for inference, therefore, is the prior distribution on the spin properties $p(\mathbf{z})$. Moreover, we can interchange x_i and x_j in (1) once the prior $p(\mathbf{z})$ is known, allowing dual-direction translation of both $p(x_j|x_i)$ and $p(x_i|x_j)$.

Likelihoods from MR Imaging Physics The log-likelihood of observing image x_i from spin properties $\mathbf{z}$ under sequence i is:

$$\log p(x_i|\mathbf{z}) \propto \|f_i(\mathbf{z}) - x_i\|_2^2, \tag{3}$$

where the signal equation (i.e. forward model) f_i can be expressed by the Bloch equation of MR imaging [13]. Below, we list the signal equations for the sequences involved in this work, namely, bSSFP, MOLLI T_1 mapping, and GRE. We denote the flip angle (FA) as ω, the repetition and echo times as and TE. The bSSFP imaging equation [24] with short and TE is:

$$f_{\text{SS}}(\mathbf{z}) = \frac{\text{PD}\sin(\omega)}{1+\cos(\omega)+[1-\cos(\omega)]\text{T}_1/\text{T}_2}. \tag{4}$$

For T_1 mapping, the phase-insensitive readout [25] at inversion time t_{INV} is:

$$f_{\text{MOLLI}}(\mathbf{z}) = \left| f_{\text{SS}}(\mathbf{z}) \left(1 - \text{INV}\cdot\exp(-\frac{t_{\text{INV}}}{\text{T}_1^*})\right)\right|, \tag{5}$$

where the apparent T_1 is $\text{T}_1^* = \left(\frac{1}{\text{T}_1}\cos^2(\frac{\omega}{2}) + \frac{1}{\text{T}_2}\sin^2(\frac{\omega}{2})\right)^{-1}$ and the inversion factor $\text{INV} = 1 + \frac{\sin(\frac{\omega}{2})}{\sin(\omega)}\left(1+\cos(\omega)+[1-\cos(\omega)]\frac{\text{T}_1}{\text{T}_2}\right)$. For GRE cine images from patients with implanted devices, let $\text{E}_1 = \exp(-\frac{}{\text{T}_1})$ and $\text{E}_2 = \exp(-\frac{\text{TE}}{\text{T}_2})$, the GRE signal model [13] is described as

$$f_{\text{GRE}}(\mathbf{z}) = \text{PD}\sin(\omega)\frac{1-\text{E}_1}{1-\cos(\omega)\text{E}_1}\text{E}_2. \tag{6}$$

2.2 Generative Prior on Spin Properties from mSASHA

To learn the spin prior $p(\mathbf{z})$, we leverage an advanced quantitative MRI sequence mSASHA, which enables joint T_1 and T_2 mapping [7]. mSASHA consists of a train of saturation recovery pulses (SR) with SR time TS, followed by a T_2-prep pulse with duration TD and echo time TE. Its signal equation is:

$$f_{\text{mSASHA}}(\mathbf{z}) = A(\mathbf{z})\left\{1 - \left[1 - \left(1 - e^{-\frac{\text{TS}}{\text{TE}}}\right)e^{-\frac{\text{TE}}{\text{T}_2}}\right]e^{-\frac{\text{TD}}{\text{T}_1}}\right\}, \tag{7}$$

where the unknowns $A(\mathbf{z}), \mathrm{T}_1, \mathrm{T}_2$ can be extracted from multiple acquisitions of f_{mSASHA} with varying TS, TD and TE by least-square fitting [7]. A encodes PD and sequence parameters, and the approximation PD $\approx A$ can be used [1,2].

To model the prior of spin properties $\mathbf{z} = (\mathrm{PD}, \mathrm{T}_1, \mathrm{T}_2)$, we propose to train a denoising diffusion probabilistic model (DDPM) [14] in the $\mathbf{z}$ space. DDPM adds noise to the real $\mathbf{z}$'s from mSASHA in the forward process $\mathbf{z}_t = \sqrt{1-\beta_t}\mathbf{z}_{t-1} + \sqrt{\beta_t}\epsilon$ for T steps, where β_t is the noise scheduler at t and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The learned model can generate data by starting with a Gaussian distribution $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ and reversing the diffusion process:

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t + (1-\alpha_t)\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t)\right) + \sigma_t \epsilon, \tag{8}$$

where $\alpha_t = 1 - \beta_t$ and $\sigma_t^2 = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$. The diffusion model learns a denoising backbone parameterized by θ, $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t) \approx -\frac{1}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t, t)$, $\overline{\alpha}_t = \prod_{r=1}^{t} \alpha_r$, as the generative prior for $\mathbf{z}$.

2.3 Reverse Imaging and Cross-Sequence Synthesis

Reverse Imaging searches for $\mathbf{z}$ in the prior space that best explains the observation x_i. Following conditional generation principles [8], we replace $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t)$ in (8) with $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|x_i)$:

$$\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|x_i) \propto \nabla_{\mathbf{z}_t} \log p(x_i|\mathbf{z}_t) + \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t). \tag{9}$$

We approximate the time-step-dependent likelihood $\log p(x_i|\mathbf{z}_t)$ by: $\log p(x_i|\mathbf{z}_t) \approx \log p(x_i|\tilde{\mathbf{z}}_0(\mathbf{z}_t))$, where $\tilde{\mathbf{z}}_0(\mathbf{z}_t) \approx (\mathbf{z}_t - \sqrt{1-\overline{\alpha}_t}\epsilon_\theta(\mathbf{z}_t))/\sqrt{\overline{\alpha}_t}$ [8]. Starting from $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$, the recursive process for reverse the imaging sequence x_i is

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t, t)\right) + \sigma_t \epsilon - \xi \frac{\partial}{\partial \mathbf{z}_t} \left\| f_i\left(\tilde{\mathbf{z}}_0(\mathbf{z}_t)\right) - x_i \right\|_2^2, \tag{10}$$

where ξ is a step size parameter. The guided reverse diffusion process in (10) yields the estimated spin properties $\hat{\mathbf{z}} = \mathbf{z}_0$ in the prior distribution that reconstructs x_i. After Reverse Imaging, we can synthesize images of novel sequences using the imaging physics equations f_j described in (4)–(6). Synthesizing other imaging sequences is also possible by sequence-specific Bloch equations. In addition, we can also use the image physics f_j of x_j to guide generation by replacing f_i, x_i with f_j, x_j in (10) and translating from target to source (T2S).

3 Experiments

Datasets. We used the ACDC dataset [3] training split with bSSFP cine images of 100 subjects as the source domain dataset. The mSASHA dataset for learning the spin prior consists of 63 subjects with 186 short-axis slices. For evaluation of

cross-sequence generalizability, we included two datasets imaged with different sequences: 1) MOLLI T_1 mapping data: 49 subjects with 131 short-axis slices in total. Each slice contains 11 baseline images with different contrasts. 2) cine MRI of 25 subjects with implanted devices: 16 subjects with the bSSFP sequence (Dev-bSSPF) and 9 with GRE sequence (Dev-GRE). Both the MOLLI and the Device datasets have significantly different contrasts than those of the ACDC.

Experimental Settings. The DDPM is pretrained on ACDC images for $36,000$ steps and then fine-tuned with the real spin properties of from mSASHA for $6,000$ steps on the resolution of 128×128. We employ the hugging-face `diffuser` for the implementation of DDPM [22]. The estimated spin properties were resized to their original resolution after Reverse Imaging. We perform $T = 1000$ forward steps and 200 reverse steps of diffusion in (10) for Reverse Imaging, with the step size $\xi = 400$. As the exact FA for ACDC scans is unknown, we approximate it with $\omega = 45^\circ$ as typically used. We open source the estimated spin properties for the ACDC dataset and code[1]. For evaluating the generalization performance, we include the following methods:

- **Baseline** We train an nnUNet [15] with the ACDC data as the baseline;
- **BayeSeg** We compare our model against BayeSeg [11,12] which extracts the boundary and contour information from images for generalization;
- **RI-T2S** We translate the MOLLI and GRE data into bSSFP cine by Reverse Imaging (RI) and perform segmentation on the translated (T2S) images with the baseline model;
- **RI-Aug.** We perform Reverse Imaging on the ACDC cine images and extend the nnUNet framework by synthesizing MOLLI and GRE images through the estimated spin properties for physics-based augmentation in (5) and (6).

4 Results and Discussion

4.1 Reverse Imaging

We first show an example of Reverse Imaging a bSSFP cine in Fig. 2-(a). The estimated spin properties $\hat{\mathbf{z}}$ accurately reconstruct the bSSFP image, indicating a high likelihood. The predicted PD has low contrast between tissues, while in T_1 and T_2 estimation, blood has a high T_1 and T_2, myocardium has high T_1 but low T_2, and fat has low T_1 but a high T_2. These match the reference relative magnitude of the spin properties [29]. Based on estimated spin properties, images of various imaging protocols can be generated with large contrast variations, as shown in Fig. 2-(b). Reverse Imaging does not provide precise spin property estimation, but we note that they are not used for quantitative evaluation purposes w.r.t. T_1 or T_2, instead, they serve to synthesize novel contrasts that are essential for wide-spectrum generalization. Reverse Imaging also enables the T2S translation from MOLLI to bSSFP cine (Fig. 2-(c)). Practically, our reverse imaging results of cine images can be used for the prior training although the mSASHA dataset is not widely available.

[1] https://github.com/Ido-zh/cmr_reverse.git

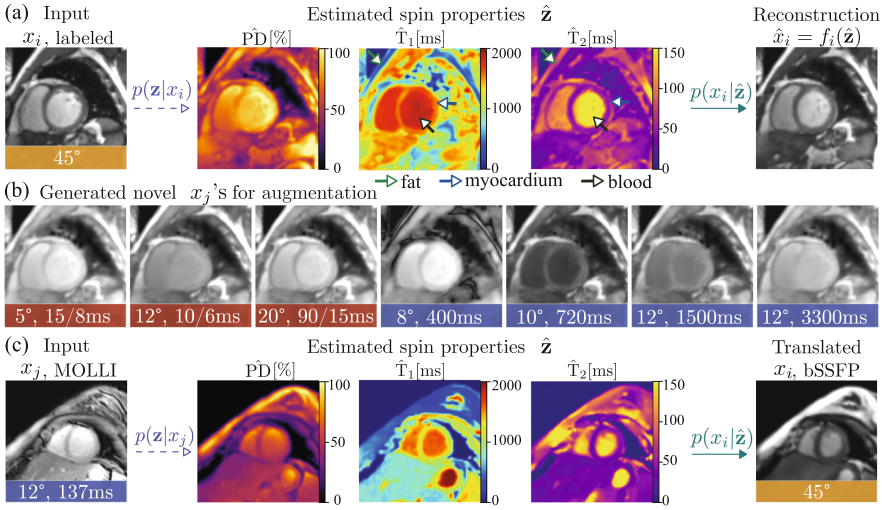


Fig. 2. (a) Reverse Imaging a bSSFP image x_i (yellow). The estimated $\hat{\mathbf{z}}$ accurately reconstructs $\hat{x}_i \approx x_i$. (b) Generating GRE (red) and MOLLI (blue) images with $\hat{\mathbf{z}}$. The FA in degrees and /TE or t_{INV} in milliseconds are given for each image. (c) Reverse Imaging a MOLLI readout for target-to-source (T2S) translation. (Color figure online)

4.2 Segmentation Performance

We now demonstrate that Reverse Imaging can significantly improve the generalization of segmentation models. In Table 1, we list the segmentation accuracy measured by the Dice score of the left ventricle (LV), myocardium (MYO) and right ventricle (RV) in the challenging MOLLI and Device datasets. To further evaluate the robustness against contrast variation, we also report the Dice for each baseline image of MOLLI in Fig. 3, as MOLLI readouts show strong contrast variation during inversion recovery. Qualitative examples of segmentation results are shown in Fig. 4.

Table 1. Segmentation accuracy on novel sequences measured by Dice score [%]. Statistical significance ($p < 0.05$) is indicated by *.

Methods	MOLLI			Device		
	RV	MYO	LV	RV	MYO	LV
Baseline [15]	24.0 ± 12.9	47.4 ± 19.5	39.9 ± 22.0	68.9 ± 25.2	86.9 ± 8.6	91.9 ± 7.6
BayeSeg [12]	51.5 ± 26.9	41.4 ± 19.9	57.0 ± 24.1	53.6 ± 21.0	62.6 ± 12.6	72.6 ± 19.1
RI-T2S	63.6 ± 36.5	69.8 ± 20.1	82.7 ± 21.0	66.4 ± 24.2	82.3 ± 15.2	88.6 ± 12.1
RI-Aug.	$\mathbf{87.0 \pm 17.8}^*$	$\mathbf{86.5 \pm 9.8}^*$	$\mathbf{91.6 \pm 9.3}^*$	$\mathbf{87.4 \pm 9.5}^*$	$\mathbf{93.1 \pm 4.0}^*$	$\mathbf{96.0 \pm 3.5}^*$

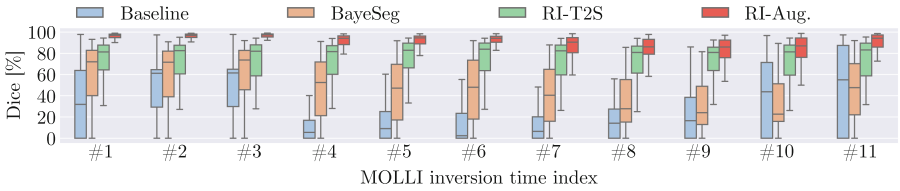


Fig. 3. Segmentation accuracy on each MOLLI readout during inversion recovery.

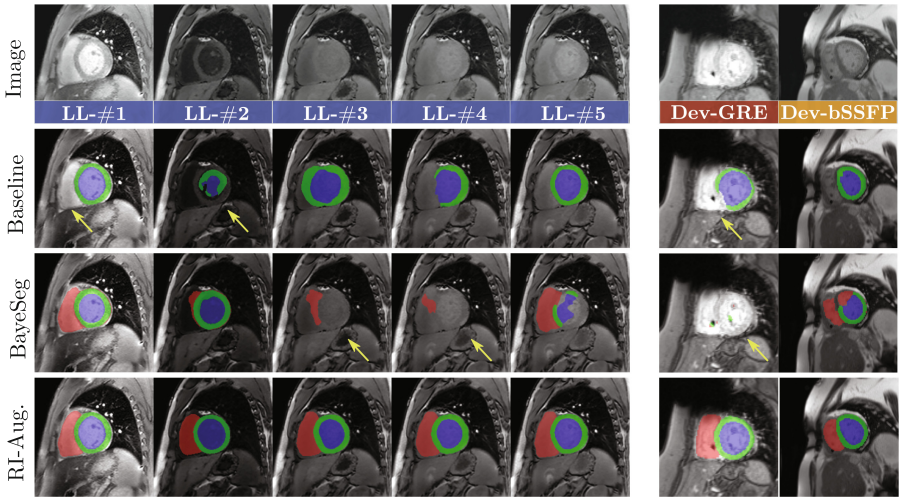


Fig. 4. Segmentation of 5 MOLLI baseline images (LL-#1-#5) and 2 device cine images, all of which have different contrast than bSSFP cine images used for training.

The baseline nnUNet generalizes poorly in the MOLLI images (Table 1), especially when the blood-myocardium contrast is inverted (see #4-#7 in Fig. 3 and LL-#2 in Fig. 4). BayeSeg can improve performance on dark blood images, yet it can still fail on low-contrast images (e.g. Figure 4 LL-#3-#5). The device images have bright blood but lower contrast, and the baseline nnUNet works more robustly than BayeSeg with higher Dice scores (Table 1). Performing the T2S transform (RI-T2S) significantly improved the segmentation accuracy for MOLLI, but a performance drop is observed in the device dataset. This is because the short TE ($\approx$ 1.5 ms) in GRE minimizes T_2 weighting (cf. (6)). In this case, Reverse Imaging is limited by the little amount of information carried by the observed image. However, this can be solved by using Reverse Imaging for augmentation instead (RI-Aug.), which achieves the highest segmentation accuracy in both datasets, significantly better than all other methods (Table 1, $p < 0.05$). With RI-Aug., the trained model consistently achieves highly accurate segmentation on images with various contrasts, including low-contrast, bright and dark blood, and GRE images (Fig. 4).

5 Conclusion

We introduced Reverse Imaging, a novel physics-grounded approach for cross-sequence generalization in cardiac MRI segmentation. Combining imaging physics and a generative diffusion prior, our method estimates underlying spin properties from observed images. Reverse Imaging provides an interpretable and physically grounded solution for domain adaptation, addressing the challenge of contrast variations across imaging protocols. Our experiments demonstrate that Reverse Imaging significantly boosts segmentation robustness, enabling zero-shot generalization to unseen sequences without requiring target-domain data. This approach offers a new paradigm for improving cardiac MRI segmentation generalizability, with potential applications to broader medical imaging tasks where domain shifts pose a significant challenge.

Acknowledgments. We acknowledge the financial support from the TU Delft AI Initiative and the Dutch Research Council (NWO).

Disclosure of Interests. We the authors hereby declare no competing interests in this paper.

References

1. Akçakaya, M., Basha, T.A., Weingärtner, S., Roujol, S., Berg, S., Nezafat, R.: Improved quantitative myocardial T2 mapping: impact of the fitting model. Magn. Reson. Med. **74**(1), 93–105 (2015)
2. Akçakaya, M., Weingärtner, S., Basha, T.A., Roujol, S., Bellm, S., Nezafat, R.: Joint myocardial T1 and T2 mapping using a combination of saturation recovery and t2-preparation. Magn. Reson. Med. **76**(3), 888–896 (2016)
3. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018)
4. Campello, V.M., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. IEEE Trans. Med. Imaging **40**(12), 3543–3554 (2021)
5. Chen, C., et al.: Unsupervised multi-modal style transfer for cardiac MR segmentation. In: Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges: 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 13 October 2019, Revised Selected Papers 10, pp. 209–219. Springer (2020)
6. Chen, C., Qin, C., Ouyang, C., Li, Z., Wang, S., Qiu, H., Chen, L., Tarroni, G., Bai, W., Rueckert, D.: Enhancing MR image segmentation with realistic adversarial data augmentation. Med. Image Anal. **82**, 102597 (2022)
7. Chow, K., et al.: Improved accuracy and precision with three-parameter simultaneous myocardial T1 and T2 mapping using multiparametric SASHA. Magn. Reson. Med. **87**(6), 2775–2791 (2022)
8. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. arXiv preprint arXiv:2209.14687 (2022)

9. Cui, Z., et al.: Structure-driven unsupervised domain adaptation for cross-modality cardiac segmentation. IEEE Trans. Med. Imaging **40**(12), 3604–3616 (2021)
10. Ferreira, P.F., Gatehouse, P.D., Mohiaddin, R.H., Firmin, D.N.: Cardiovascular magnetic resonance artefacts. J. Cardiovasc. Magn. Reson. **15**(1), 41 (2013)
11. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: Joint modeling of image and label statistics for enhancing model generalizability of medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 360–369. Springer (2022)
12. Gao, S., Zhou, H., Gao, Y., Zhuang, X.: BayeSeg: Bayesian modeling for medical image segmentation with interpretable generalizability. Med. Image Anal. **89**, 102889 (2023)
13. Hashemi, R.H., Bradley, W.G., Lisanti, C.J.: MRI: the Basics: The Basics. Lippincott Williams & Wilkins (2012)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural. Inf. Process. Syst. **33**, 6840–6851 (2020)
15. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
16. Kellman, P., Hansen, M.S.: T1-mapping in the heart: accuracy and precision. J. Cardiovasc. Magn. Reson. **16**(1), 2 (2014)
17. Martín-Isla, C., et al.: Deep learning segmentation of the right ventricle in cardiac MRI: the M&Ms challenge. IEEE J. Biomed. Health Inform. **27**(7), 3302–3313 (2023)
18. Messroghli, D.R., Radjenovic, A., Kozerke, S., Higgins, D.M., Sivananthan, M.U., Ridgway, J.P.: Modified look-locker inversion recovery (MOLLI) for high-resolution t1 mapping of the heart. Magnetic Reson. Med. Official J. Int. Soc. Magnetic Reson. Med. **52**(1), 141–146 (2004)
19. Ouyang, C., et al.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Trans. Med. Imaging **42**(4), 1095–1106 (2022)
20. Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., Rueckert, D.: Data efficient unsupervised domain adaptation for cross-modality image segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 669–677. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_74
21. Pauly, J.M.: Compressed sensing MRI. Sig. Process. Mag. IEEE **25**(2), 72–82 (2008)
22. von Platen, P., et al.: Diffusers: state-of-the-art diffusion models. https://github.com/huggingface/diffusers (2022)
23. Ridgway, J.P.: Cardiovascular magnetic resonance physics for clinicians: part I. J. Cardiovasc. Magn. Reson. **12**(1), 71 (2010)
24. Scheffler, K., Lehnhardt, S.: Principles and applications of balanced SSFP techniques. Eur. Radiol. **13**(11), 2409–2418 (2003). https://doi.org/10.1007/s00330-003-1957-x
25. Schmitt, P., et al.: Inversion recovery TrueFISP: quantification of t1, t2, and spin density. Magn. Reson. Med. Official J. Int. Soc. Magn. Reson. Med. **51**(4), 661–667 (2004)
26. Song, Y., Shen, L., Xing, L., Ermon, S.: Solving inverse problems in medical imaging with score-based generative models. arXiv preprint arXiv:2111.08005 (2021)
27. Tao, Q., et al.: Deep learning-based method for fully automatic quantification of left ventricle function from cine MR images: a multivendor, multicenter study. Radiology **290**(1), 81–88 (2019)

28. Xie, Q., et al.: Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. IEEE Trans. Med. Imaging **43**(1), 4–14 (2022)
29. Xu, Z., et al.: Reference ranges of myocardial t1 and t2 mapping in healthy chinese adults: a multicenter 3t cardiovascular magnetic resonance study. J. Cardiovasc. Magn. Reson. **25**(1), 64 (2023)

Self is the Best Learner: CT-Free Ultra-low-Dose PET Organ Segmentation via Collaborating Denoising and Segmentation Learning

Zanting Ye[1], Xiaolong Niu[1], Xu Han[2], Xuanbin Wu[1], Wantong Lu[1], Yijun Lu[1], Hao Sun[1], Yanchao Huang[3], Hubing Wu[3], and Lijun Lu[1,4,5,6(✉)]

[1] School of Biomedical Engineering, Southern Medical University, Guangzhou, China
[2] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3] Nanfang PET Center, Nanfang Hospital Southern Medical University, Guangzhou, China
[4] Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China
[5] Guangdong Province Engineering Laboratory for Medical Imaging and Diagnostic Technology, Southern Medical University, Guangzhou, China
[6] Pazhou Lab, Guangzhou, China
ljlubme@gmail.com

Abstract. Organ segmentation in Positron Emission Tomography (PET) plays a vital role in cancer quantification. Low-dose PET (LDPET) provides a safer alternative by reducing radiation exposure. However, the inherent noise and blurred boundaries make organ segmentation more challenging. Additionally, existing PET organ segmentation methods rely on co-registered Computed Tomography (CT) annotations, overlooking the problem of modality mismatch. In this study, we propose LDOS, a novel CT-free ultra-LDPET organ segmentation pipeline. Inspired by Masked Autoencoders (MAE), *we reinterpret LDPET as a naturally masked version of Full-Dose PET (FDPET)*. LDOS adopts a simple yet effective architecture: a shared encoder extracts generalized features, while task-specific decoders independently refine outputs for denoising and segmentation. By integrating CT-derived organ annotations into the denoising process, LDOS improves anatomical boundary recognition and alleviates the PET/CT misalignments. Experiments demonstrate that LDOS achieves state-of-the-art performance with mean Dice scores of 73.11% (^{18}F-FDG) and 73.97% (^{68}Ga-FAPI) across 18 organs in 5% dose PET. Our code will be available at https://github.com/yezanting/LDOS.

Keywords: Organs segmentation · Ultra-LDPET · MAE · Denoising

Z. Ye, X. Niu and X. Han—These authors contributed equally to this work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 566–576, 2026.
https://doi.org/10.1007/978-3-032-04947-6_54

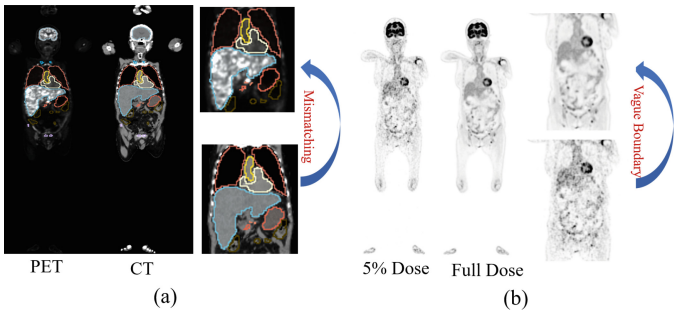


Fig. 1. (a) Misalignments between PET and CT caused by respiratory motion (e.g., liver displacement). (b) Noise and blurred boundary artifacts in LDPET.

1 Introduction

Positron emission tomography (PET) is a powerful molecular imaging modality that visualizes radiotracer distribution to reveal physiological processes. Low-dose PET (LDPET) reduces radiation exposure and has demonstrated diagnostic equivalence to Full-Dose PET (FDPET) [4,14,25,26]. However, its clinical adoption remains limited by the lack of robust quantitative tools in high-noise settings. Accurate tracer uptake measurements in non-tumoral organs are critical for tumor quantification within Volumes of Interest (VOIs) [1,15]. LDPET organ segmentation offers a promising approach for tracer uptake assessment and kinetic measurements [23,27], yet it remains relatively understudied.

In this study, we summarize the challenges of LDPET organ segmentation: (1) **Data scarcity**: Clinical FDPET dominance limits LDPET dataset availability. (2) **Annotation complexity**: Low soft-tissue contrast in PET, worsened in LDPET, makes organ annotation challenging. Existing PET organ segmentation methods rely on co-registered Computed Tomography (CT) annotations, overlooking the issue of modality mismatch (Fig. 1a). (3) **Noise-induced ambiguity**: Blurred anatomical boundaries hinder accurate segmentation (Fig. 1b).

To address these challenges, we propose LDOS, a novel CT-free LDPET organ segmentation pipeline via collaborating denoising and segmentation learning. Inspired by MAE [8], we reinterpret LDPET as a naturally masked version of FDPET, where the denoising process inherently recovers organ-level semantics. Recent studies have demonstrated that incorporating organ annotations into LDPET denoising can improve FDPET reconstruction, highlighting the shared semantic priors between denoising and segmentation tasks [5,11,24].

Unlike traditional MAE-based methods that depend on large-scale pre-training, LDOS employs self-denoising to reinforce organ-level semantic learning, reducing the risks of overfitting and bias-learning in small-scale PET datasets [16,18,21]. From a technical perspective, LDOS uses a simple yet effective training pipeline. A shared encoder extracts generalized features for both tasks, while task-specific decoders independently refine outputs for denoising

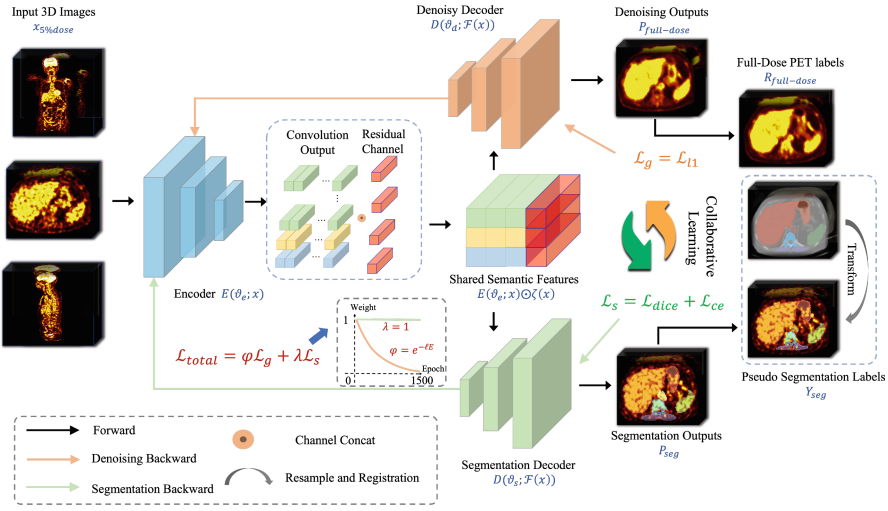


Fig. 2. Overview of LDOS. LDOS employs the shared encoder and independent decoders for denoising and segmentation tasks.

(supervised by FDPET) and segmentation (supervised by CT annotations during training only). By incorporating CT-derived organ annotations into the denoising process, LDOS improves anatomical boundary recognition and mitigates PET/CT misalignments.

2 Related Work

PET Organ Segmentation. Existing PET segmentation methods primarily focus on organs with high tracer uptake (e.g., tumors) in FDPET, while organ segmentation in LDPET remains largely unexplored [2,17,20]. Furthermore, PET organ segmentation methods rely on co-registered CT annotations, ignoring modality mismatch [19,22]. To address this gap, we propose LDOS, which operates directly on noisy ultra-LDPET data. By incorporating CT-derived organ annotations into the denoising process, LDOS improve anatomical boundary recognition and mitigates PET/CT misalignments.

MAE in Medical Image Analysis. MAE excels in natural images with large-scale pretraining (e.g., ImageNet [6]). However, medical imaging lacks comparable datasets, making pretraining on limited data prone to overfitting and bias [16,18,21]. Inspired by MAE, where masking removes random patches, we reinterpret LDPET as positional masking of FDPET, where low-dose acquisition inherently "masks" high-frequency details. Instead of resource-intensive pretraining, LDOS integrates denoising and segmentation into a single-stage collaborative learning process, avoiding small-data pitfalls and ensuring anatomical consistency.

3 Main Methodology

LDOS employs a simple yet effective training pipeline, which improves segmentation accuracy by jointly reconstructing FDPET signals. As shown in Fig. 2, it employs a shared encoder E_θ and two task-specific decoders D_ϕ and D_ψ.

3.1 Model Architecture

Shared Encoder. LDOS uses the nnU-Net backbone [9] with the ResEncL configuration [10]. Inspired by MAE, LDOS employs the LDPET denoising process to learn semantic latent representations. MAE uses the patch-level discretized images as inputs [7]. In the medical imaging context, large-scale patch masking leads to excessive information loss. The degradation in LDPET physically stems from lower photon counts and consequent statistical noise. LDOS approximates this effect as a pixel-level masking process, which facilitates the retention of core information:

$$x_i = \psi(\Lambda r_i, k) \tag{1}$$

where x_i is the LDPET image, r_i is the FDPET image, and Λr represents the pixel-level discretization. The encoder processes the input x_i to extract semantic latent features:

$$F(x_i) = E(\vartheta_e; x_i) \odot \zeta(x_i) \tag{2}$$

where $F(x_i)$ represents shared semantic features, E is the encoder, ϑ_e denotes encoder parameters, $\odot$ signifies dimensionality stacking, and ζ is the residual connection. Unlike standard MAE approaches that focus on visible pixels, our method models all pixels, providing comprehensive semantic representation.

Segmentation and Denoising Decoders. The encoder output feeds into both segmentation and denoising decoders, which share a consistent architecture. To enhance feature learning, a deep supervision strategy is incorporated [9]:

$$P^z_{full-dose} = D_d(\vartheta^z_d; F(x_i)) \tag{3}$$

$$P^z_{seg} = \frac{\exp(D^c_s(\vartheta^z_s; F(x_i)))}{\sum_C \exp(D^c_s(\vartheta^z_s; F(x_i))} \tag{4}$$

where $P^z_{full-dose}$ is the FDPET prediction at scale z, D is the decoder, and ϑ^z_d denotes the decoder parameters. Similarly, P^z_{seg} refers to the segmentation prediction for the c-th class, with D^c_s being the decoder for class c and C the total number of segmentation classes.

3.2 Training via Collaborative Denoising and Segmentation

To address PET/CT misalignments [2,12], misalignment data augmentation [19] is applied during training. This strategy mitigates label inconsistencies and

improves coordination between segmentation and denoising. The denoising decoder is supervised by the loss function L_g:

$$L_g = \sum_{Z} \omega^z \left| P^z_{full-dose} - r \right| \tag{5}$$

where Z is the total number of scales, ω^z is the weight of scale z, and r is the FDPET image. For segmentation, a combination of Dice and cross-entropy losses is used:

$$L_{ce} = \sum_{Z} \sum_{C} \upsilon^z \left(-y_c \log P^z_{seg,c} - (1 - y_c) \log(1 - P^z_{seg,c}) \right) \tag{6}$$

$$L_{dice} = \sum_{Z} \upsilon^z \left(1 - \frac{2 \sum_{C} (y_c \cdot \rho(soft\max(P^z_{seg,c})))}{\sum_{C} y_c + \rho(soft\max(P^z_{seg,c}))} \right) \tag{7}$$

$$L_s = L_{ce} + L_{dice} \tag{8}$$

where L_{ce} is the cross-entropy loss, C is the total number of classes, υ^z is the weight at scale z, y_c is the ground truth for class c, ρ is one-hot embedding, and $P^z_{seg,c}$ is the segmentation prediction. The total loss function combines denoising and segmentation losses:

$$L_{total} = \varphi L_g + \lambda L_s \tag{9}$$

where $\varphi = e^{-\ell E}$; $\ell = 0.002$, $\lambda = 1$, and E denotes the number of epochs. Initially, denoising is emphasized to capture semantic features, and as training progresses, the focus shifts to refine segmentation accuracy.

Table 1. Performance of the 18 organs segmentation results on the ^{18}F-FDG and ^{68}Ga-FAPI datasets (5-fold cross validation). Dice, IoU, HD95, and ASD are used to evaluate LDOS, and '±' represents 'mean ± standard deviation'. The green and yellow highlights mean the highest and lowest segmentation results.

Organs	^{18}F-FDG Dataset				^{68}Ga-FAPI Dataset			
	Dice (%)	IoU (%)	HD95 (mm)	ASD (mm)	Dice (%)	IoU (%)	HD95 (mm)	ASD (mm)
Spleen	76.28 ± 2.31	58.99 ± 3.50	13.56 ± 4.32	3.77 ± 1.16	69.37 ± 5.61	54.90 ± 5.85	17.67 ± 4.32	4.95 ± 0.88
Colon	62.22 ± 3.47	53.33 ± 4.94	32.35 ± 8.07	9.42 ± 2.54	44.76 ± 4.79	29.68 ± 1.61	42.99 ± 7.33	8.58 ± 2.14
Urinary Bladder	72.75 ± 7.64	58.61 ± 11.99	9.23 ± 5.87	3.65 ± 3.16	79.94 ± 2.99	66.15 ± 6.51	14.74 ± 8.06	3.50 ± 1.44
Sacrum	69.20 ± 4.18	57.84 ± 5.03	12.29 ± 2.72	3.92 ± 0.65	71.05 ± 4.97	53.55 ± 3.86	10.20 ± 2.56	2.92 ± 0.21
Vertebrae	77.38 ± 4.85	65.87 ± 2.19	10.06 ± 3.31	2.84 ± 0.96	72.10 ± 2.22	54.92 ± 0.78	11.06 ± 6.49	2.40 ± 0.13
Heart	81.06 ± 4.30	72.59 ± 5.35	12.11 ± 3.56	4.10 ± 2.07	82.51 ± 1.57	69.92 ± 1.59	10.97 ± 0.86	3.71 ± 0.46
Aorta	76.38 ± 3.18	59.63 ± 2.04	12.20 ± 2.84	4.34 ± 0.75	72.32 ± 4.66	58.12 ± 2.10	10.44 ± 1.85	2.91 ± 0.20
Clavicula	63.34 ± 5.09	54.53 ± 4.38	15.78 ± 7.34	5.57 ± 2.76	59.08 ± 5.49	40.95 ± 3.75	10.39 ± 3.09	2.85 ± 0.10
Femur	77.05 ± 2.46	66.71 ± 3.94	10.88 ± 4.34	3.38 ± 1.94	88.58 ± 1.67	79.91 ± 2.05	5.74 ± 1.40	2.53 ± 0.95
Hip	76.50 ± 4.00	65.70 ± 3.78	10.54 ± 6.84	3.77 ± 1.64	76.86 ± 2.08	62.75 ± 2.60	8.67 ± 2.00	2.70 ± 0.19
Autochthon	75.92 ± 2.06	63.69 ± 2.75	8.28 ± 4.34	3.16 ± 1.55	85.05 ± 3.70	74.08 ± 3.22	5.56 ± 0.61	2.07 ± 0.12
Kidney	75.61 ± 4.79	64.35 ± 3.97	14.35 ± 5.06	4.78 ± 1.62	75.81 ± 3.62	62.07 ± 3.47	11.03 ± 0.58	3.30 ± 0.28
Brain	94.63 ± 3.65	91.12 ± 2.39	4.58 ± 1.79	1.87 ± 0.68	95.64 ± 0.87	91.65 ± 0.70	4.57 ± 0.23	1.86 ± 0.15
Liver	86.52 ± 2.46	78.78 ± 3.57	14.82 ± 5.33	5.14 ± 2.25	85.19 ± 1.97	72.30 ± 3.48	19.80 ± 5.50	5.56 ± 1.38
Stomach	63.90 ± 8.16	52.76 ± 7.49	23.24 ± 12.56	7.59 ± 2.01	69.76 ± 8.80	57.94 ± 8.86	29.92 ± 11.75	7.50 ± 2.44
Pancreas	52.61 ± 2.33	40.64 ± 5.94	18.80 ± 7.66	5.45 ± 2.29	54.58 ± 5.50	30.58 ± 6.94	21.38 ± 1.94	4.67 ± 0.42
Lung	89.33 ± 0.80	84.72 ± 2.01	12.10 ± 2.48	4.21 ± 1.50	91.88 ± 1.74	84.55 ± 2.03	7.81 ± 0.70	2.57 ± 0.25
Esophagus	50.78 ± 2.14	37.89 ± 6.45	19.49 ± 9.56	7.32 ± 5.32	56.96 ± 2.44	41.51 ± 4.13	9.00 ± 1.74	2.33 ± 0.17
Mean	73.11 ± 1.02	62.65 ± 3.07	14.15 ± 2.04	4.68 ± 1.01	73.97 ± 2.26	60.31 ± 2.72	14.00 ± 2.17	3.72 ± 0.46

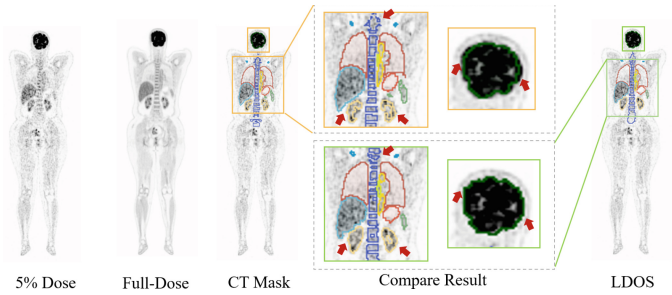


Fig. 3. A sample of head and abdominal organs segmentation on the ^{18}F-FDG. For clarity, segmentation results for some organs are not displayed.

4 Experiment and Results

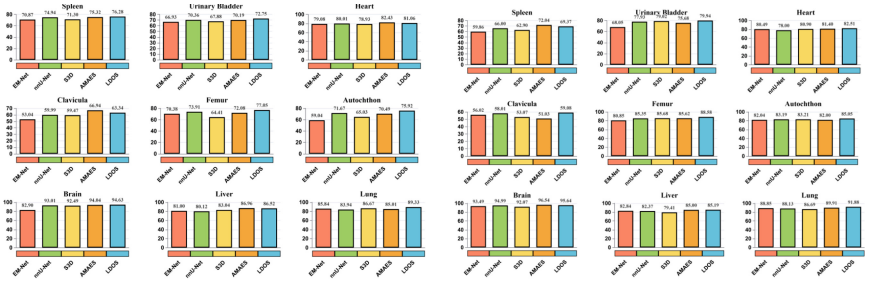
4.1 Datasets and Implementation Details

Excluding significant misalignments, we used 52 ^{18}F-FDG and 60 ^{68}Ga-FAPI tracer scans acquired from Nanfang Hospital Southern Medical University on a UIH uEXPLORER (Total-Body) PET/CT to validate LDOS. The sampling times were set to 300 s (100% dose) for FDPET and 15 s (5% dose) for ultra-LDPET. Each dataset included 18 segmented organs and was split into training (80%), validation (10%), and testing (10%) subsets. Five-fold cross-validation was performed separately per tracer. Notably, direct organ annotation on PET images was challenging due to constraints in the clinical workflow. Consequently, we utilized annotations from corresponding CT scans as the gold standard, which were then manually refined. Once trained, our model performs organ segmentation on PET images alone, without reliance on the CT modality.

LDOS was implemented based on the nnU-Net and trained from scratch. We trained 500 epoches. The patch size was $192 \times 192 \times 192$, and the batch size was 2. Training was conducted on an NVIDIA GeForce RTX 4090 GPU with 24 GB memory.

Table 2. Performance of the 18 organs segmentation results on real LDPET and FDPET images. Dice (%) is used as the metric.

Metric	^{18}F-FDG			^{68}Ga-FAPI		
	FDPET	LDPET	LDOS	FDPET	LDPET	LDOS
Dice	76.03 ± 3.61	68.03 ± 2.35	**73.11 ± 1.02**	74.65 ± 3.09	65.73 ± 2.79	**73.97 ± 2.26**
IoU	62.98 ± 4.01	50.77 ± 7.39	**62.65 ± 3.07**	63.68 ± 2.45	48.70 ± 8.31	**60.31 ± 2.72**
HD95	15.71 ± 3.50	21.05 ± 3.84	**14.15 ± 2.04**	11.03 ± 2.64	23.94 ± 5.33	**14.00 ± 2.17**
ASD	3.96 ± 1.67	6.88 ± 1.78	**4.68 ± 1.01**	3.12 ± 1.06	7.06 ± 2.00	**3.72 ± 0.46**



(a) 18F-FDG Results (b) 68Ga-FAPI Results

Fig. 4. Performance comparison of selected organs across different methods on the ^{18}F-FDG (a) and ^{68}Ga-FAPI (b) datasets. Dice (%) is used as the metric.

4.2 Multiorgan Segmentation Results

The segmentation performance of LDOS was evaluated using multiple metrics, including Dice Similarity Coefficient (Dice), Intersection over Union (IoU), 95th percentile Hausdorff distance (HD95), and Average Surface Distance (ASD). A 5-fold cross-validation was conducted on both datasets. The quantitative segmentation results are shown in Table 1. The average Dice scores across the 18 organs were 73.11%±1.02% and 73.97%±2.26% for the ^{18}F-FDG and ^{68}Ga-FAPI datasets, respectively. Additionally, the average HD95 were 14.15 mm ± 2.04 mm and 14.00 mm ± 2.17 mm, demonstrating the robust segmentation capability of LDOS at ultra-LPPET (5% dose).

LDOS was compared to the segmentation results of real LDPET and FDPET. For a fair comparison, the same nnU-Net architecture was employed, excluding the denoising decoder loss. As shown in Table 2 and Fig. 4, LDOS achieved notable improvements in Dice scores, with 5.08% and 8.24% higher accuracy than LDPET on the ^{18}F-FDG and ^{68}Ga-FAPI datasets, respectively. Moreover, LDOS achieved segmentation accuracy comparable to or surpassing FDPET for certain organs.

PET organ annotations for model training are derived from corresponding CT images, as direct organ labeling on PET images is impractical. However, achieving precise alignment of training is challenging due to inherent spatial shifts. As illustrated in Fig. 3, LDOS mitigates the misalignments by semantic feature learning of the self-denoising process, providing a novel solution to improve ultra-LDPET organ segmentation.

4.3 Comparisons with Previous Results

As shown in Table 3, Fig. 4 and 5, LDOS outperformed the state-of-the-art methods. Pre-training methods based on MAE [16,21] did not yield significant improvements. LDOS uses semantic features extracted during the denoising process to achieve superior results. These findings are consistent with prior

Table 3. Performance comparison of different methods on the ^{18}F-FDG and ^{68}Ga-FAPI datasets. Dice (%) is used as the metric. The best performance is highlighted in bold, and the suboptimal performance is underlined.

Method	Pre-train Data	^{18}F-FDG	^{68}Ga-FAPI
Baseline (ResEncU)	-	68.03 ± 2.35	65.73 ± 2.79
EM-Net [3]	-	65.45 ± 5.11	62.77 ± 3.46
nnU-Net [9]	-	67.58 ± 7.30	63.96 ± 4.15
S3D (nnU-Net) [21]	AutopetIII	68.86 ± 3.61	64.86 ± 1.98
	Our Datasets	70.35 ± 1.78	70.87 ± 2.35
	AutopetIII and Our Datasets	70.03 ± 0.96	$\underline{72.76 \pm 2.36}$
AMAES (U-Net B) [16]	-	67.39 ± 5.00	64.97 ± 6.27
	BRAINS-45K	66.92 ± 3.76	64.68 ± 2.99
	AutopetIII	68.45 ± 4.37	64.56 ± 3.89
	AutopetIII and Our Datasets	$\underline{71.01 \pm 1.65}$	69.33 ± 1.24
LDOS	-	$\mathbf{73.11 \pm 1.02}$	$\mathbf{73.97 \pm 2.26}$

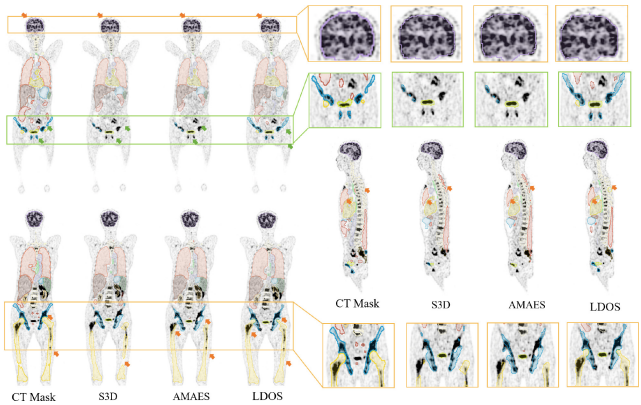


Fig. 5. Visualization of segmentation results on the ^{18}F-FDG dataset.

studies [13,16,18,21], which indicate that pre-training on small-scale datasets provide limited advantages for medical imaging tasks.

4.4 Ablation Experiment

To evaluate the contribution of the proposed pipeline, we compared its against LDPET and fake FDPET. As shown in Table 4, LDOS improved Dice scores on both the ^{18}F-FDG and ^{68}Ga-FAPI datasets. Unlike fake FDPET, which employed sequential denoising and segmentation, LDOS's collaborative design achieved superior results without complex denoising model training. LDOS reduces deployment costs while enhancing usability. Additionally, results from the LDPET-LDOS/ow (omits the L1 loss weight attenuation strategy) indicates that focusing on denoising semantic information during the early training stages

Table 4. Ablation studies of LDOS. LDPET-LDOS/ow means LDOS without the L1 loss weight attenuation strategy, and Fake-FDPET means that denoising and segmentation are performed sequentially. The best performance is highlighted in bold.

Method	Dataset	Dice (%)	IoU (%)	HD95 (mm)	ASD (mm)
LDPET	^{18}F-FDG	68.03 ± 2.35	50.77 ± 7.39	21.05 ± 3.84	6.88 ± 1.78
Fake-FDPET	–	**7**3.75 ± 3.02	61.46 ± 2.96	**1**3.76 ± 3.05	5.35 ± 1.84
LDPET-LDOS/ow	–	72.01 ± 4.00	61.01 ± 2.66	15.02 ± 2.59	5.72 ± 1.55
LDPET-LDOS	–	73.11 ± 1.02	**6**2.65 ± 3.07	14.15 ± 2.04	**4**.68 ± 1.01
LDPET	^{68}Ga-FAPI	65.73 ± 2.79	48.70 ± 8.31	23.94 ± 5.33	7.06 ± 2.00
Fake-FDPET	–	73.45 ± 3.01	**6**1.05 ± 4.33	16.54 ± 3.20	4.07 ± 1.28
LDPET-LDOS/ow	–	71.09 ± 3.71	57.97 ± 3.15	17.39 ± 2.69	3.88 ± 1.09
LDPET-LDOS	–	**7**3.97 ± 2.26	60.31 ± 2.72	**1**4.00 ± 2.17	**3**.72 ± 0.46

and prioritizing segmentation optimization in the later stages positively impacts the model's overall performance.

5 Conclusion

Ultra-LDPET organ segmentation is critical for reliable tracer uptake assessments and kinetic measurement, but remains largely unexplored. In this study, we propose LDOS, a CT-free ultra-LDPET organ segmentation pipeline. LDOS reinterprets LDPET as a naturally masked version of FDPET and extracts semantic information through a self-denoising process. By incorporating CT-derived organ annotations into the denoising process, LDOS improves anatomical boundary recognition and mitigates PET/CT misalignments, providing a novel solution for LDPET quantification analysis.

Acknowledgments. This research was funded by the National Natural Science Foundation of China (62371221, 12326616 and 62201245), National High end Foreign Experts Recruitment Plan (G2023030025L)and the Science and Technology Program of Guangdong Province 2022A0505050039.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aoki, J., et al.: Fdg pet of primary benign and malignant bone tumors: standardized uptake value in 52 lesions. Radiology **219**(3), 774–777 (2001)
2. Bao, N., et al.: Ct-less whole-body bone segmentation of pet images using a multimodal deep learning network. IEEE J. Biomed. Health Inf. (2024)

3. Chang, A., Zeng, J., Huang, R., Ni, D.: Em-net: efficient channel and frequency learning with mamba for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 266–275. Springer (2024)
4. Chen, W., et al.: Evaluation of pediatric malignancies using total-body pet/ct with half-dose [18f]-fdg. Eur. J. Nucl. Med. Mol. Imaging **49**(12), 4145–4155 (2022)
5. Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M., Chen, Z.: Deep learning based synthesis of mri, ct and pet: review and analysis. Med. Image Anal. **92**, 103046 (2024)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
7. Dosovitskiy, A.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods **18**(2), 203–211 (2021)
10. Isensee, F., et al.: nnu-net revisited: a call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 488–498. Springer (2024)
11. Jiang, C., Pan, Y., Cui, Z., Nie, D., Shen, D.: Semi-supervised standard-dose pet image generation via region-adaptive normalization and structural consistency constraint. IEEE Trans. Med. Imaging **42**(10), 2974–2987 (2023)
12. Kovacs, B., et al.: Addressing image misalignments in multi-parametric prostate mri for enhanced computer-aided diagnosis of prostate cancer. Sci. Rep. **13**(1), 19805 (2023)
13. Li, W., Yuille, A., Zhou, Z.: How well do supervised models transfer to 3d image segmentation. In: The Twelfth International Conference on Learning Representations, vol. 1 (2024)
14. Liu, G., et al.: Short-time total-body dynamic pet imaging performance in quantifying the kinetic metrics of 18f-fdg in healthy volunteers. Eur. J. Nucl. Med. Mol. Imaging **49**(8), 2493–2503 (2022)
15. Mansouri, Z., et al.: Development and validation of survival prognostic models for head and neck cancer patients using machine learning and dosiomics and ct radiomics features: a multicentric study. Radiat. Oncol. **19**(1), 12 (2024)
16. Munk, A., Ambsdorf, J., Llambias, S., Nielsen, M.: Amaes: augmented masked autoencoder pretraining on public brain mri data for 3d-native segmentation. arXiv preprint arXiv:2408.00640 (2024)
17. Oh, K.T., Lee, S., Lee, H., Yun, M., Yoo, S.K.: Semantic segmentation of white matter in fdg-pet using generative adversarial network. J. Digit. Imaging **33**(4), 816–825 (2020)
18. Rokuss, M., et al.: From fdg to psma: a hitchhiker's guide to multitracer, multicenter lesion segmentation in pet/ct imaging. arXiv preprint arXiv:2409.09478 (2024)
19. Salimi, Y., Mansouri, Z., Shiri, I., Mainta, I., Zaidi, H.: Deep learning-powered ct-less multitracer organ segmentation from pet images: a solution for unreliable ct segmentation in pet/ct imaging. Clin. Nucl. Med. (2025). https://doi.org/10.1097/RLU.0000000000005685

20. Taghanaki, S.A., et al.: Combo loss: Handling input and output imbalance in multi-organ segmentation. Comput. Med. Imaging Graph. **75**, 24–33 (2019)
21. Wald, T., et al.: Revisiting mae pre-training for 3d medical image segmentation. arXiv preprint arXiv:2410.23132 (2024)
22. Wang, H., et al.: Robust and generalizable artificial intelligence for multi-organ segmentation in ultra-low-dose total-body pet imaging: a multi-center and cross-tracer study. Eur. J. Nuclear Med. Molecular Imaging 1–15 (2025)
23. Zaidi, H., Karakatsanis, N.: Towards enhanced pet quantification in clinical oncology. Br. J. Radiol. **91**(1081), 20170508 (2017)
24. Zhang, J., Cui, Z., Jiang, C., Guo, S., Gao, F., Shen, D.: Hierarchical organ-aware total-body standard-dose pet reconstruction from low-dose pet and ct images. IEEE Trans. Neural Netw. Learn. Syst. (2023)
25. Zhou, B., Tsai, Y.J., Chen, X., Duncan, J.S., Liu, C.: Mdpet: a unified motion correction and denoising adversarial network for low-dose gated pet. IEEE Trans. Med. Imaging **40**(11), 3154–3164 (2021)
26. Zhou, L., Schaefferkoetter, J.D., Tham, I.W., Huang, G., Yan, J.: Supervised learning with cyclegan for low-dose fdg pet image denoising. Med. Image Anal. **65**, 101770 (2020)
27. Ziai, P., et al.: Role of optimal quantification of fdg pet imaging in the clinical practice of radiology. Radiographics **36**(2), 481–496 (2016)



Semi-Supervised Deformation-Free Image-to-Image Translation for Realistic CT Synthesis from CBCT

Ji Yong Han[1], Su Yang[2], Sujeong Kim[1], Sunjung Kim[3], Sang-Heon Lim[1], Heejin Yun[1], Dahee Kim[2], and Won-Jin Yi[1,2,4(✉)]

[1] Interdisciplinary Program in Bioengineering, Graduate School of Engineering, Seoul National University, Seoul, Republic of Korea
wjyi@snu.ac.kr

[2] Applied Bioengineering, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea

[3] Imaging R&D Center, Osstem Implant Co., Ltd., Seoul, Republic of Korea

[4] Oral and Maxillofacial Radiology and Dental Research Institute, School of Dentistry, Seoul National University, Seoul, Republic of Korea

Abstract. Cone-Beam Computed Tomography (CBCT) is widely used for diagnostics and treatment planning in oral and maxillofacial field due to its low radiation dose and high spatial resolution. Still, its clinical utility is limited by low contrast and incorrect Hounsfield Unit (HU) values. In contrast, multi-detector CT (CT) provides high contrast and reliable HU measurements, with a higher radiation dose. In this work, we present a novel two-stage framework for unpaired CBCT-to-CT synthesis that ensures the exact preservation of anatomical structure, maintains high resolution, and achieves accurate HU value. In the first stage, we generate pseudo-paired CT images. In the second stage, we utilize a UNet++ generator enhanced with Interpolation and Convolution Upsampling (ICUP), Edge-Conditioned Skip Connections (ECSC), and a dual discriminator strategy for a semi-supervised approach. Consequently, we generate realistic CT images using pseudo-paired CT images. Extensive quantitative and qualitative evaluations demonstrate that our method outperforms existing unpaired translation techniques, producing realistic CT images that closely match CT images in both HU accuracy and exactly preserve anatomical structure of the CBCT. The code is available at https://github.com/HANJIYONG/Semi-Supervised-Deformation-Free-I2I.

Keywords: Computed Tomography · Unpaired Image-to-Image Translation · Semi-supervised Learning

1 Introduction

Cone-Beam Computed Tomography (CBCT) is widely used for diagnostics and treatment planning in the oral and maxillofacial field [1]. CBCT has the

J. Y. Han and S. Yang—These authors contributed equally.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 577–586, 2026.
https://doi.org/10.1007/978-3-032-04947-6_55

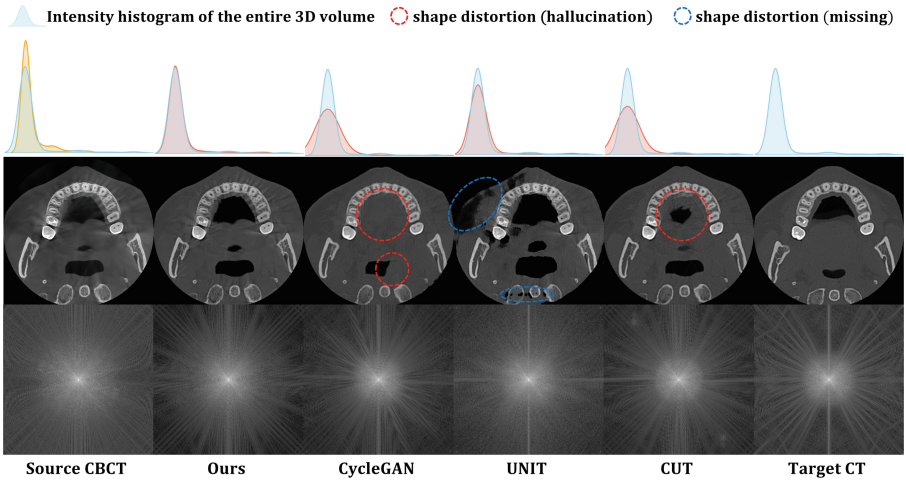


Fig. 1. Illustration of shape inconsistency and resolution degradation in unpaired CBCT-to-CT translation. Our method preserves structural integrity, aligns HU intensities with CT, and maintains fine spatial details. The Fourier-domain magnitude maps visualize spatial frequency characteristics: CBCT shows a broad, circular energy distribution indicating high-resolution content, while the target CT and other methods exhibit more centralized, low-frequency energy patterns. Our method retains a wider radial frequency distribution, highlighting better preservation of anatomical detail.

advantages of lower radiation exposure and higher spatial resolution compared with multi-detector CT (CT), and it provides three-dimensional information on anatomical structures. However, its reduced contrast resulting from lower radiation dose limits use in diagnostics and bone density measurements [2,3]. In contrast, CT provides smoother images with enhanced contrast while providing consistent Hounsfield Unit (HU) measurements [4]. Recent advances in AI-based solutions have demonstrated that image-to-image translation (I2I) methods can selectively preserve the benefits of both modalities [5–8]. These approaches enhance image quality with high spatial resolution and uniform HU scale. However, in clinical settings, obtaining paired CBCT and CT data is challenging due to variations in patient posture, breathing, and motion during image acquisition. Unpaired I2I methods have been explored to address these challenges. CycleGAN [9] enforces cycle consistency on unpaired datasets to map images between domains. Similarly, contrastive unpaired translation (CUT) [10] employs a contrastive learning framework to align corresponding patches between input and output images without an explicit cycle consistency constraint. UNIT [11] assumes a shared latent space between the two domains, enabling unsupervised translation via coupled generative adversarial networks. Although these methods have shown promising performance, they may cause shape distortions, which are critical in medical imaging. Moreover, since CT typically exhibits lower spatial resolution than CBCT, the high-resolution advantage of CBCT is compro-

mised during translation (see Fig. 1). In this paper, we propose a novel semi-supervised I2I framework that addresses the above issues. First, we develop an Edge-conditioned unpaired I2I model, inspired by the CUT approach, to generate pseudo-paired CT images. Second, we perform a paired I2I using pseudo-paired data to synthesize realistic CT images without shape distortion. We validate our method against various unpaired I2I methods, demonstrating improvements in both image quality and fidelity. Our contributions can be summarized as follows:

- We present a novel semi-supervised framework for realistic CT synthesis, employing a pseudo-paired CT generation strategy followed by a paired CT generation stage that prevents structural distortion and preserves resolution
- We design an edge-conditioned generator that robustly guides CT synthesis to avoid structural distortion and ensure HU consistency. Moreover, we propose a multi-task loss that combines several loss functions to preserve structure, reduce resolution loss, and achieve accurate HU representation.
- We validate our method using structural integrity, resolution, and HU accuracy metrics, with comprehensive qualitative and quantitative analyses confirming its effectiveness.

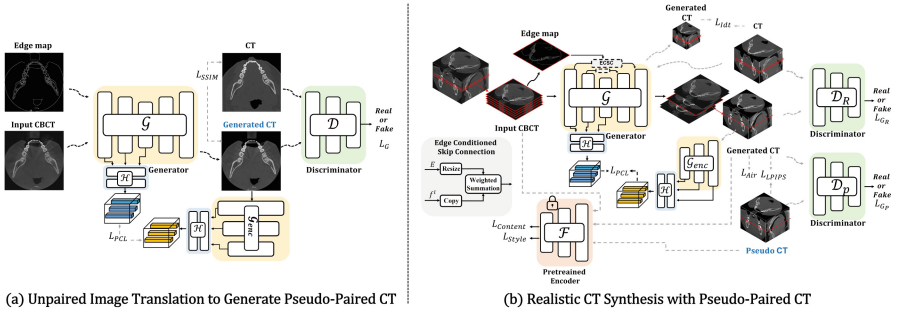
2 Method

2.1 Preliminaries: Contrastive Unpaired Translation

Contrastive Unpaired Translation (CUT) [10] is an unpaired I2I method that maximizes the mutual information between corresponding patches in the source and generated output. Unlike cycle consistency-based approaches, CUT eliminates the need for reverse mapping by using patch-wise contrastive learning to preserve local features during translation. In this process, a source image x is transformed by the generator $\mathcal{G}$ into a translated image $\hat{y} = \mathcal{G}(x)$ that aligns with the target domain. To maintain local feature consistency, CUT maximizes the mutual information between patches in x and the corresponding patches in $\hat{y}$ while distinguishing them from unrelated patches. Patches are embedded as vectors $z_k = \mathcal{H}(\mathcal{G}_{enc}(x)_k)$ and $w_k = \mathcal{H}(\mathcal{G}_{enc}(\hat{y})_k)$, where $\mathcal{H}$ is a multi-layer perceptron (MLP) and k indexes the spatial locations in $\mathcal{G}_{enc}(x)$. The patch-wise contrastive loss is defined as:

$$L_{PatchNCE} = \sum_{k \in \psi} \left[-\log \frac{exp(z_k^\top w_k / \tau)}{\sum_{i \neq k} exp(z_i^\top w_k / \tau)} \right] \tag{1}$$

where ψ is a set of spatial indices sampled from $\{1, 2, ..., S\}$ (with S being the total number of spatial locations) and τ is the temperature parameter controlling the sharpness of the distribution. In this formulation, w_k serves as anchor, (w_k, z_k) constitutes a positive pair, and (w_k, z_i) for $i \neq k$ forms a negative pair.



(a) Unpaired Image Translation to Generate Pseudo-Paired CT

(b) Realistic CT Synthesis with Pseudo-Paired CT

Fig. 2. Overview of the proposed two-stage framework. (a) Unpaired image-to-image translation with edge condition generates pseudo-CT images from CBCT data. (b) A paired translation stage generates realistic CT images, ensuring consistent HU value and preserving resolution.

2.2 Unpaired Image Translation to Generate Pseudo-Paired CT

CUT [10] preserves structural consistency by maximizing mutual information between local patches in the input and generated images, eliminating the cycle consistency loss required by CycleGAN [9] and reducing computational overhead. However, unpaired supervision can still lead to content deformation. To address this, we propose an edge-conditional generator that synthesizes CT while retaining essential content. Specifically, we extract a gradient edge map $x_{edge} \in \mathbb{R}^{H\times W\times 1}$, to capture structural boundaries, then concatenate it with the source CBCT image from a two-channel input $x \in \mathbb{R}^{H\times W\times 2}$. We also incorporate a multi-scale discriminator [12] that evaluates both global structure and local texture details, enhancing the quality and stability of the translation. This approach faithfully reconstructs fine textures while maintaining structural coherence, forming a crucial step in creating pseudo-paired datasets for higher-fidelity synthesis in the subsequent stage.

2.3 Realistic CT Synthesis with Pseudo-Paired CT

Semi-Supervised Framework. With the constructed pseudo-paired CT dataset, we perform a paired image-to-image translation. Since the pseudo dataset is not entirely accurate, real CT images are incorporated to better align the HU value distribution. To prevent structural deformation, we adopt a UNet++ [13] generator enhanced with Edge-Conditioned Skip Connections (ECSC) and an Interpolation and Convolution Upsampling (ICUP) strategy. ECSC fuses feature maps with the edge map through a learnable weighted summation, preserving resolution and preventing shape distortion. The ICUP method replaces transposed convolution-based upsampling, eliminating checkerboard artifacts and yielding smoother 3D reconstructions. All batch normalization layers are replaced with instance normalization for improved consistency

[14]. While the pseudo-paired dataset exhibits structural coherence and approximate HU alignment, it does not guarantee perfect HU accuracy. To address this, we employ a dual discriminator framework: D_R discriminates between synthetic CT and real unpaired CT, and D_P compares synthetic CT with pseudo-paired CT, thereby enhancing the reliability of the final translation.

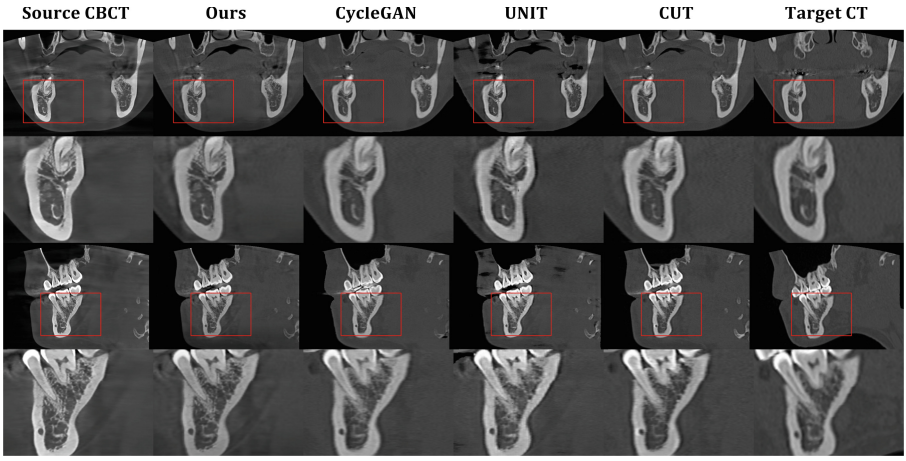


Fig. 3. Qualitative comparison with different methods for unpaired image-to-image translation. The first and third views represent the coronal and sagittal view, respectively. The second and fourth rows correspond to cropped images of red-boxed regions. (Color figure online)

Multi-task Loss. Our overall loss function is a weighted sum of components that enforce perceptual quality, structural consistency, and HU value alignment. We include LPIPS loss (L_{LPIPS}) [15], style loss (L_{Style}) [16] to guide HU values, patch-wise contrastive loss (L_{PCL}), content loss ($L_{Content}$) [16], identity loss (L_{Idt}) [9], and airway segmentation loss (L_{Air}) for structural consistency. Additionally, a dual adversarial loss ($L_{DualAdv}$) is introduced to enforce consistency in pseudo-paired data and to reduce HU differences with CT images. The LPIPS loss preserves perceptual similarity by comparing deep feature representations between the generated and pseudo-paired images:

$$L_{LPIPS} = \mathbb{E}_{X\sim P(X)}[\|\phi(G(X)) - \phi(X_p)\|_2^2] \tag{2}$$

where $\phi(\cdot)$ denotes feature embeddings extracted from a frozen pre-trained network. We adopt the VGG-16 [17] pretrained on ImageNet [18]. $G(X)$ is generated image, and X_p is the pseudo-paired reference. The Style loss enforces texture consistency by aligning the Gram matrices of feature maps:

$$L_{Style} = \sum_l \|\text{Gram}(\psi_l(G(X))) - \text{Gram}(\psi_l(X_p))\|_2^2 \tag{3}$$

where $\psi_l(\cdot)$ represents the feature maps at layer l of pre-trained network, and Gram$(\cdot)$ computes the corresponding Gram matrix. The Patch-wise Contrastive Loss L_{PCL} is applied as defined in Eq. (1) to enhance local feature consistency. The Content loss ensures structural fidelity by minimizing the difference between feature embeddings of the generated and pseudo-paired images:

$$L_{Content} = \sum_{l} \|\psi_l(G(X)) - \psi_l(X_p)\|_2^2 \tag{4}$$

The Identity loss encourages the generator to maintain the target domain's structure:

$$L_{Idt} = \mathbb{E}_{X \sim P(X_R)}[\|X - G(X)\|_2^2] \tag{5}$$

where X_R represents images from the real CT dataset. To further enforce structural consistency, particularly in the air regions, we introduce an Air Segmentation Loss that minimizes the mean squared error (MSE) between the predicted air regions of the source CBCT and the generated CT:

$$L_{Air} = \mathbb{E}_{X \sim P(X)}[\|S(G(X)) - S(X)\|_2^2] \tag{6}$$

with $S(\cdot)$ denoting the air segmentation network. Finally, to align the anatomical structure similarity of pseudo-paired CT images and the HU value with real CT images, we employ a dual adversarial loss using two discriminators, D_P, and D_R. The dual adversarial loss is defined as:

$$\begin{aligned} L_{DualAdv} =& \mathbb{E}_{y \sim P_{real}(Y)}[\log D_R(y)] + \mathbb{E}_{x \sim P(X)}[\log(1 - D_R(G(x)))] + \\ & \mathbb{E}_{y \sim P_{pseudo}(Y)}[\log D_P(y)] + \mathbb{E}_{x \sim P(X)}[\log(1 - D_P(G(x)))] \end{aligned} \tag{7}$$

The total multi-task loss is the weighted sum of these terms:

$$\begin{aligned} L_{Total} =& \lambda_{DualAdv} L_{DualAdv} + \lambda_{LPIPS} L_{LPIPS} + \lambda_{Style} L_{Style} + \\ & \lambda_{Content} L_{Content} + \lambda_{PCL} L_{PCL} + \lambda_{Idt} L_{Idt} + \lambda_{Air} L_{Air} \end{aligned} \tag{8}$$

where $\lambda_{DualAdv} = 2$, $\lambda_{LPIPS} = 1$, $\lambda_{style} = 2$, $\lambda_{content} = 1$, $\lambda_{PCL} = 1$, $\lambda_{idt} = 1$, and $\lambda_{Air} = 1$ are weights determined empirically.

Table 1. Performance comparison of our method with other unpaired image-to-image translation networks."Structure" denotes the structure score, "Luminance" represents the luminance score, and "Contrast" indicates the contrast score.

Methods	Content & Resolution		Hounsfield Unit		
	Structure (↑)	GMSD (↓)	Luminance (↑)	Contrast (↑)	NCC (↑)
Cycle GAN	0.940 ± 0.049	0.196 ± 0.028	0.792 ± 0.011	0.919 ± 0.028	0.856 ± 0.051
UNIT	0.944 ± 0.028	0.234 ± 0.028	0.736 ± 0.133	0.919 ± 0.029	0.851 ± 0.069
CUT	0.965 ± 0.021	0.146 ± 0.027	0.761 ± 0.110	0.934 ± 0.016	0.863 ± 0.050
OURS	**0.988 ± 0.006**	**0.139 ± 0.015**	**0.809 ± 0.091**	**0.948 ± 0.019**	**0.888 ± 0.046**

Table 2. Performance comparison of HU values and mean absolute error (MAE) across hard tissue regions (enamel, dentin, cortical bone, and trabecular bone). The proposed method achieves HU values closest to CT and exhibits lower MAE compared to CBCT, CycleGAN, UNIT, and CUT in all regions except trabecular bone.

Methods	Enamel		Dentin		Cortical Bone		Trabecular Bone	
	HU	MAE	HU	MAE	HU	MAE	HU	MAE
CT	2577.67	–	1560.56	–	1355.32	–	557.84	–
CBCT	3453.14	875.46	1956.90	396.34	1953.16	597.83	1128.78	570.93
CycleGAN	2421.24	156.43	1669.25	108.94	1459.47	104.16	503.08	**54.77**
UNIT	2700.22	122.55	1973.66	413.52	1485.27	287.98	833.80	275.93
CUT	2758.98	181.25	2045.66	485.10	1732.15	339.86	398.39	163.96
Ours	2614.89	**37.21**	1620.85	**60.29**	1497.69	**67.35**	708.75	150.91

3 Experiment

3.1 Datasets and Implementation Details

Dataset. We collected CBCT and corresponding CT images from 40 subjects (20 males and 20 females; age range 21–80 years), CT scans were acquired using a Somatom Definition Edge CT scanner (Siemens AG, Erlangen, Germany) at 120 kVp and 120 mA, with voxel sizes of $0.49 \times 0.49 \times 0.5\,\mathrm{mm}^3$, and dimensions of 512×512 pixels. CBCT scans were acquired using a Green X 12 CBCT scanner (Vatech, Hwaseong-si, South Korea) at 94 kVp and 11.7 mA, with voxel sizes of $0.2 \times 0.2 \times 0.2\,\mathrm{mm}^3$, and dimensions of 600×600 pixels. For training, 10,625 axial slices from 25 subjects were used, and 6,375 axial slices from 15 subjects were reserved for evaluation. All images were resized to 576×576 pixels, and data augmentation including horizontal flipping, shifts of $\pm 0.1\%$, rotations of $\pm 20°$, and scaling variations of $\pm 0.1\%$ was applied to enhance generalization.
Training Setup. All models were trained by Adam optimizer for 50 epochs with an initial learning rate of 2×10^{-4}. A batch size of 1 was used, and training was performed on an NVIDIA GeForce RTX 4090 GPU (24 GB RAM). All models were implemented in Python3 using the PyTorch framework and executed in identical computing environments to ensure a fair comparison.
Evaluation Metrics. To assess the performance of the generated images, we computed the structure score from the structural similarity index measure (SSIM) [19] and the gradient magnitude similarity deviation (GMSD) [20] to evaluate structural and resolution similarity between the generated output and the source CBCT. Additionally, we used luminance and contrast scores derived from SSIM, along with normalized cross-correlation (NCC), to evaluate the similarity of the HU distribution between the generated output and the CT images.

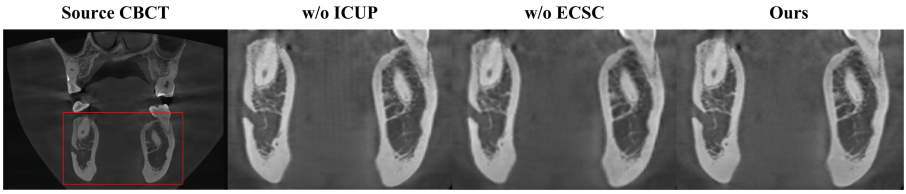


Fig. 4. Qualitative comparison of the impact of ICUP and ECSC. The red box indicates the region of interest. Without ICUP, checkerboard artifacts cause vertical line distortions, and without ECSC, resolution is degraded.

3.2 Experimental Results

Quantitative Comparison with Other Methods. We evaluated the performance of our proposed method against several unpaired image-to-image translation methods, including CycleGAN [9], UNIT [11], and CUT [10]. Our method achieves the best performance across all evaluation metrics. In particular, our method outperforms CUT with the second highest performance by obtaining 0.988 ± 0.006, 0.139 ± 0.015, 0.809 ± 0.091, 0.948 ± 0.018, and 0.888 ± 0.046 for Structure, GMSD, Luminance, Contrast, and NCC, as summarized in Table 1. Table 2 presents the analysis results for hard tissues. Among enamel, dentin, cortical bone, and trabecular bone, our method demonstrated most similar to CT images in all regions except trabecular bone by obtaining mean absolute error (MAE) 37.21, 60.29, 67.35, and 150.91.

Qualitative Comparison with Other Methods. Figure 3 shows that our method produces images that closely resemble the CT, preserving structural details without deformation while maintaining a resolution comparable to the source CBCT. Moreover, as demonstrated in Fig. 1, our approach generates realistic CT images that closely approximate the target CT, outperforming other methods. In particular, while other models exhibit slice inconsistency, our method achieves exact slice consistency.

Effectiveness of Proposed Components. We conducted ablation experiments to assess the effectiveness of our proposed ICUP and ECSC modules. Table 3 shows the quantitative performance variations with and without these components. The results indicate that ICUP effectively suppresses checkerboard artifacts, eliminating vertical line artifacts in the reconstructed volumes. Additionally, ECSC robustly prevents shape deformation and reduces resolution loss by explicitly integrating edge conditions into the skip connections (see Fig. 4).

Table 3. Quantitative performance comparison of the proposed ICUP and ECSC modules.

Methods	Content & Resolution		Hounsfield Unit		
	Structure (↑)	GMSD (↓)	Luminance (↑)	Contrast (↑)	NCC (↑)
Baseline	0.968 ± 0.003	0.150 ± 0.020	0.772 ± 0.120	0.945 ± 0.016	0.878 ± 0.046
+ICUP	0.970 ± 0.004	0.143 ± 0.028	0.782 ± 0.112	0.945 ± 0.015	0.877 ± 0.046
+ECSC	0.980 ± 0.003	0.141 ± 0.018	0.788 ± 0.106	0.946 ± 0.016	0.882 ± 0.045
+ICUP, ECSC (Ours)	**0.988 ± 0.006**	**0.139 ± 0.015**	**0.809 ± 0.091**	**0.948 ± 0.019**	**0.888 ± 0.046**

4 Conclusion

In this paper, we proposed a semi-supervised framework for CBCT-to-CT synthesis that ensures the exact preservation of anatomical shape, high resolution, and accurate HU value. Our two-stage approach employs an edge-conditioned unpaired translation model to generate pseudo-paired CT images, followed by a paired translation stage using a UNet++ generator enhanced by ICUP, ECSC, and dual discriminators. Extensive evaluations demonstrate that our method outperforms existing techniques, achieving HU values closest to the CT. In future work, we plan to further explore artifact-free CT synthesis to enhance clinical applicability.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2023R1A2C2 00532611). This work was also supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program-Advanced Biomaterials) (RS-2025-14322975), funded by the Ministry of Trade, Industry & Energy (MOTIE).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Choi, J.W., et al.: Relationship between physical factors and subjective image quality of cone-beam computed tomography images according to diagnostic task. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. **119**(3), 357–365 (2015)
2. Scarfe, W.C., Farman, A.G., Sukovic, P., et al.: Clinical applications of cone-beam computed tomography in dental practice. J.-Can. Dent. Assoc. **72**(1), 75 (2006)
3. Kiljunen, T., Kaasalainen, T., Suomalainen, A., Kortesniemi, M.: Dental cone beam Ct: a review. Physica Med. **31**(8), 844–860 (2015)
4. Ludlow, J.B., Ivanovic, M.: Comparative dosimetry of dental Cbct devices and 64-slice ct for oral and maxillofacial radiology. Oral Surgery, Oral Med. Oral Pathol. Oral Radiol. Endodontology **106**(1), 106–114 (2008)
5. Dahiya, N., et al.: Multitask 3d Cbct-to-ct translation and organs-at-risk segmentation using physics-based data augmentation. Med. Phys. **48**(9), 5130–5141 (2021)
6. Kang, S.R., et al.: Structure-preserving quality improvement of cone beam Ct images using contrastive learning. Comput. Biol. Med. **158**, 106803 (2023)

7. Yong, T.H., et al.: Qcbct-net for direct measurement of bone mineral density from quantitative cone-beam Ct: a human skull phantom study. Sci. Rep. **11**(1), 15083 (2021)
8. Suwanraksa, C., Bridhikitti, J., Liamsuwan, T., Chaichulee, S.: Cbct-to-ct translation using registration-based generative adversarial networks in patients with head and neck cancer. Cancers **15**(7), 2017 (2023)
9. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
10. Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y.: Contrastive learning for unpaired image-to-image translation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 319–345. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_19
11. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
12. Kumar, K., et al.: Melgan: generative adversarial networks for conditional waveform synthesis. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
13. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: a nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in Conjunction with MICCAI 2018, Granada, Spain, 20 September 2018, proceedings 4, pp. 3–11. Springer (2018)
14. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
15. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
16. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
20. Xue, W., Zhang, L., Mou, X., Bovik, A.C.: gradient magnitude similarity deviation: a highly efficient perceptual image quality index. IEEE Trans. Image Process. **23**(2), 684–695 (2013)

Sim-to-Real Transformer-Based Shape Reconstruction for Automated Orthopedic Fracture Reduction Planning

Sutuke Yibulayimu[1], Yanzhen Liu[1], Yudi Sang[3(✉)], Gang Zhu[3], Hui Li[1], Hao Lu[1], Chunpeng Zhao[2], Xinbao Wu[2], and Yu Wang[1(✉)]

[1] Key Laboratory of Biomechanics and Mechanobiology, Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, 100083 Beijing, China
{sutuk,wangyu}@buaa.edu.cn
[2] Department of Orthopaedics and Traumatology, Beijing Jishuitan Hospital, Capital Medical University, Beijing, China
[3] Beijing Rossum Robot Technology Co., Ltd., Beijing, China
sangyudi@rossumrobot.cn

Abstract. Accurate orthopedic fracture reduction planning is essential for ensuring successful postoperative recovery and improving patient outcomes. However, current methods are challenged by the complex and irregular fracture geometries and the scarcity of annotated training data. To address these challenges, we propose a novel approach that integrates learning-based shape restoration and fracture simulation. A transformer-based model is developed, which utilizes patch-to-patch restoration and recursive fragment registration to iteratively refine fracture reduction poses. To generate diverse and anatomically realistic fractured datasets for model training, we develop a fracture data simulation approach that combines statistical shape modeling with clinically representative fracture patterns, reducing reliance on annotated samples. Tested on extensive clinical data with hipbone and sacrum fractures, the proposed method achieved mean translational and rotational errors of 2.34 mm and 4.54 °, respectively, outperforming both template-based and existing learning-based methods. Our approach enhances learning and generalization for automated fracture reduction by connecting synthetic and real-world fracture data.

Keywords: Surgery planning · Point cloud deep learning · Transformer · Fracture simulation · Statistical shape model

1 Introduction

Fracture reduction, namely the process of realigning fractured bone fragments, is critical for optimizing biomechanical stability, promoting healing, and

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-032-04947-6_56.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 587–597, 2026.
https://doi.org/10.1007/978-3-032-04947-6_56

restoring patient mobility [14]. However, the planning process is highly subjective, time-consuming, and heavily dependent on surgeons' expertise [19]. Complex fractures, especially in the hipbones and sacrum, further complicate the task, where even minor misalignments can lead to chronic pain and impaired functionality. These challenges highlight the need for reliable and automated reduction planning methods to determine the target pose of each bone fragment [27].

Existing automated approaches fall primarily into two categories: fracture surface matching and template-based methods. Surface matching methods align fragments based on geometric features but often fail due to the inaccuracy in segmenting and defining fracture surfaces [8,13,15,20]. Template-based methods, such as the widely used mirror-template approach [5,8,26,28], rely on predefined reference shapes. Although effective for unilateral fractures, they cannot handle bilateral fractures or bones without contralateral references, such as the sacrum. Furthermore, the alignment errors caused by anatomical asymmetries between contralateral bones should not be ignored [6].

As an alternative to the mirror template, statistical models have been studied. Early methods construct mean shapes of bones but failed to account for individual variability [4]. Statistical shape models (SSMs) capture anatomical variability and enable adaptive deformations, providing flexibility in reconstructing diverse fracture configurations [7,12,21]. However, SSM-based methods require simultaneous optimization of fragment poses and deformation parameters, often leading to suboptimal results due to local minima.

Attempts have been made to apply deep leaning in reduction planning. A PointNet-based framework has been proposed to facilitate the planning process by establishing initial correspondence between fracture fragments and templates [3]. A convolutional pose estimation network has been used to address sacroiliac joint dislocation, but has not been applied to fractures of single bones where the fracture surfaces are more complex [11,22]. A key challenge in learning-based methods is the scarcity of annotated fracture datasets. The high variability of fracture patterns calls for models with high generalizability.

Synthetic data generation offers a promising solution to data scarcity. Bidirectional frameworks have been proposed to generate fracture data and learn non-rigid transformations between healthy and fractured morphologies [25]. This method relies on real fracture data with contralateral bones for fine-tuning, introducing challenges when differences between two sides exist. Furthermore, standard convolutional networks struggle to extract meaningful features from highly variable fracture morphologies. This underscores the need for tailored feature extraction methods designed to address such variability.

In this study, we present a learning-based approach for fracture reduction planning trained with synthetic data that is only based on small datasets. Our major contributions include: (1) A transformer-based point could processing network is developed to restore intact bone shape from fractured data through fragment-aware patch encoding and patch-to-patch translation. (2) A recursive refinement strategy is implemented to iteratively improve alignment accuracy,

effectively addressing diverse fracture configurations. (3) A simulation pipeline is developed, leveraging statistical shape modeling to capture anatomical variability while learning real fracture patterns.

2 Method

We aim to estimate the target pose of each fragment from surface point clouds. These point clouds are obtained by segmenting fragments in CT scans [10] and uniformly sampling points on their surfaces. As shown in Fig. 1, a transformer-based point cloud processing network is developed to restore bone shape, and a fracture data generation process is designed to support network training.

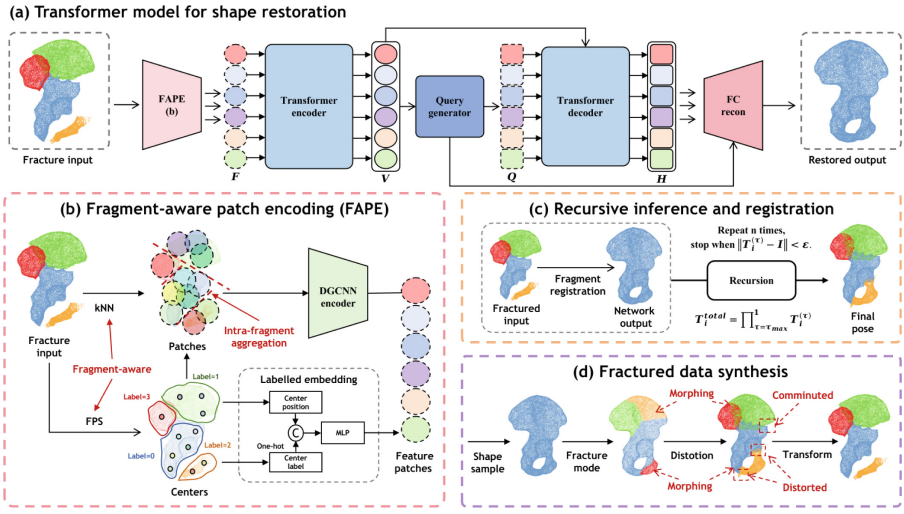


Fig. 1. Overview of the proposed method. (a) The transformer network performs patch-to-patch translation to restore bone shape. (b) FAPE encodes local features for each fragment. (c) Reduction plans are refined through recursive inference and registration. (d) Synthetic data is generated with rich variations to simulate realistic fracture.

2.1 Transformer Model for Shape Restoration

Considering the variability in the number, pose, and interactions of fracture fragments, we formulate the fracture reduction planning task as a *patch-to-patch translation* problem based on transformer network, where the input is the labeled point cloud of fractured fragments, and the output is the point cloud of the reconstructed shape. A naive unified encoding for all fragments can lead to undesired interference among fragments. To mitigate this problem, we design a transformer-based network that is aware of fragment labels throughout (Fig. 1a).

Fragment-Aware Patch Encoding. Fractured bones are represented by point clouds of multiple fragments. As shown in Fig. 1b, to prepare the fragments for the transformer-based model, we first sample N patch centers across all fragments. Direct application of farthest point sampling (FPS) to the entire fractured bone may result in incomplete coverage near fragment boundaries. To address this, we allocate sampling centers proportionally to the point distribution of each fragment, and then apply FPS within each fragment to obtain center points and labels $\{s_i, \eta_i\}_{i=1}^{N}$ For each center s_i, we form a local patch P_i by selecting its k nearest neighbors (kNN) within its entire fragment.

We applied a DGCNN to encode each patch P_i through dynamic graph construction and aggregation. Compared to the original model proposed in [16], both sampling and neighborhood aggregation are performed using the fragment-aware FPS and kNN methods described above. To embed fragment-specific identity, the patch center s_i is concatenated with its one-hot encoded fragment label η_i, and then processed by a multilayer perceptron (MLP):

$$F_i = \text{DGCNN}(P_i) + \text{MLP}([s_i, \text{one-hot}(\eta_i)]). \tag{1}$$

After processing all N patch centers, we obtain a unified set of patch features $\mathcal{F} = \{F_1, F_2, \ldots, F_N\}$, which serves as input to the transformer modules.

Patch-to-Patch Translation with Transformers. In this formulated patch-to-patch translation task, the input fractured patch features are transformed into reconstructed patch features. This process is defined as:

$$\mathcal{V} = \text{ME}(\mathcal{F}), \quad \mathcal{H} = \text{MD}(\mathcal{Q}, \mathcal{V}), \tag{2}$$

where ME and MD are the encoder and decoder models, $\mathcal{V} = \{V_1, V_2, \ldots, V_N\}$ are the encoded patch features, $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_N\}$ are the query embeddings, and $\mathcal{H} = \{H_1, H_2, \ldots, H_N\}$ are the reconstructed patch features. The architecture is based on the geometry-aware transformer proposed in [23].

Between the encoder and decoder, a query generator predicts query embeddings $\mathcal{Q}$ to guide reconstruction. Coarse query centers $\mathcal{C}$ are first predicted from the global feature of $\mathcal{V}$, which is extracted using max-pooling. These centers are then refined to produce the final query embeddings:

$$\mathcal{C} = \text{Proj}(\text{Pool}(\text{Linear}(\mathcal{V}))), \quad \mathcal{Q} = \text{MLP}([\mathcal{C}, \text{Pool}(\text{Linear}(\mathcal{V}))]), \tag{3}$$

where Proj is a linear layer that predicts 3D coordinates, and Pool is the max-pooling operation.

Coarse-to-Fine Reconstruction. We adopt a multi-scale reconstruction strategy to restore bone shape. The query generator predicts local centers $\mathcal{C}$, which serve as anchors for reconstructing the point cloud. A fully connected reconstruction head f refines these centers into detailed local geometries:

$$R_i = f(H_i) + c_i, \quad i = 1, 2, \ldots, N, \tag{4}$$

where R_i represents reconstructed points around each center c_i.

The reconstruction loss is computed both at sparse level on the patch centers $\mathcal{C}$ and at dense level on the restored shape $\mathcal{R}$, using the Chamfer distance (CD) to the ground-truth point cloud $\mathcal{G}$:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{sparse}} + \mathcal{L}_{\text{dense}} = \text{CD}(\mathcal{C}, \mathcal{G}) + \text{CD}(\mathcal{R}, \mathcal{G}), \tag{5}$$

where the same dense ground truth $\mathcal{G}$ is used for both dense and sparse supervision without additional downsampling, since Chamfer Distance does not require point-wise correspondence.

Recursive Inference and Registration. The restored shape serves as a template for fracture reduction. During testing, to correct errors from large initial displacements, we employ a recursive cascade approach (Fig. 1c). At each iteration τ, the network predicts a restored shape $R^{(\tau)}$, and fragments are aligned to $R^{(\tau)}$ using the iterative closest point (ICP) algorithm. This updates fragment poses and transformations $\{T_i^{(\tau+1)}\}$, which are iteratively refined until convergence ($\|T_i^{(\tau)} - I\| < \epsilon$) or a maximum iteration threshold is reached.

2.2 Training Data Simulation

To generate diverse and clinically realistic fractures for shape restoration, our method learns anatomical variations and fracture patterns from a limited set of intact and fractured samples by combining anatomical shape modeling and fracture mode labeling.

An SSM is constructed for each bone type using Point2SSM++ as the backbone network [1]. This approach establishes point correspondences across different bone geometries and applies principal component analysis (PCA) to efficiently model anatomical variability, enabling the generation of plausible bone shapes. Fracture patterns are then modeled by assigning fracture labels to the SSM based on clinical fractured data. Each SSM point is labeled according to the spatial distribution of fracture fragments observed in real samples. To ensure alignment between the SSM and fractured bones, shape deformation parameters and spatial transformations are jointly optimized [21]. This process, applied across multiple fractures, generates a diverse set of fracture modes within a unified coordinate system, forming the foundation for fracture synthesis.

Beyond the predefined fracture modes, additional variability is introduced through extensive post-processing. As illustrated in Fig. 1d, fractured samples are generated by sampling SSM shape parameters, applying a fracture mode, and incorporating morphing and distortion. The morphing step adjusts the label distribution of adjacent fracture surfaces, while distortion, which includes rotation, scaling, and shearing, alters the geometric structure to simulate irregularity in fracture surfaces commonly observed in clinical data. Further variability is introduced by randomly removing boundary regions to mimic missing shapes as in locally comminuted fracture cases. Since fragments originate from intact models without predefined interfaces, a mesh refinement algorithm [2] is employed

to close open surfaces. The final fractured samples undergo randomized pose adjustments, while the intact bone serves as the ground truth for training.

3 Experiments and Results

3.1 Experimental Setup

Datasets. We selected two of the most complex and challenging bone anatomies, hipbone and sacrum, to test our method. From the *CTPelvic1K* dataset [9], 430 healthy hipbone and sacrum samples were selected. They were divided into training and test sets in an 8:2 ratio for constructing the SSM. Additionally, 188 hipbone fracture cases and 52 sacrum fracture cases were selected from the *PENGWIN* dataset [10]. Due to the scarcity of fractured data, they were split in a 2:8 ratio, with 20% used for constructing the fracture modes in DFPM and 80% reserved for evaluation. In total, 191 real fracture samples were used for testing, representing one of the largest fracture test datasets to date. Ground-truth reduction poses were annotated by two clinical experts, each with more than five years of surgical planning experience.

Baseline Methods. We compared the proposed network model with several fracture reduction planning methods, including average templates (Mean) [4], statistical shape templates (SSM) [21], mirrored templates (Mirror) [26], and fracture surface matching based on mirrored templates (Match) [8]. Due to their reliance on the contralateral part, mirrored-template methods were restricted to unilateral hipbone fractures only. For a fair comparison, unilateral and bilateral hipbone fracture cases were tested separately in our experiments. Additionally, a transformer-based point cloud generation method, AdaPoinTr [24], was included as a learning-based baseline.

Evaluation. Reduction accuracy was assessed using translational error (in millimeter), defined as the Euclidean distance between the centers of reduced fragments and their ground truth, and rotational error (in degree), defined as the axis-angle difference between the planned and ground-truth pose matrices for each fragment. Furthermore, the reconstructed shapes were evaluated using L2 Chamfer distance (CD) (scaled by 10^{-4}) to the ground truth and part accuracy (PA) with the latter thresholded τ at 0.001 to enhance sensitivity, following established protocols [17].

Implementation. The proposed network was implemented in PyTorch, with an NVIDIA RTX 3070Ti GPU. The input point clouds of fractured bones were initially downsampled to 2,048 points using FPS. Fragment labels were encoded as one-hot vectors of seven dimension. The FAPE module generated 256 patch tokens, each with 128 dimensions. And the transformer used 384 dimensions. Reconstruction was performed at the coarse level with 256 points and the fine level with 4,096 points. AdamW optimizer was used with an initial learning rate

of 10^{-4} and weight decay of 5×10^{-4}, following a cosine annealing schedule. The training spanned 300 epochs with a batch size of 20. The recursive process allowed up to 10 iterations and used a convergence threshold ϵ of 1.

For data simulation, SSMs were constructed separately for the hipbone and sacrum using 344 samples, each with 38 and 11 fracture modes, respectively. Each bone had 20,000 synthesized samples, with random translation up to 15 mm and rotation up to 30° applied at each training iteration. The demonstration of data simulation is provided in the Supplementary Material. The training code and sample data are available at https://github.com/Sutuk/FracFormer.

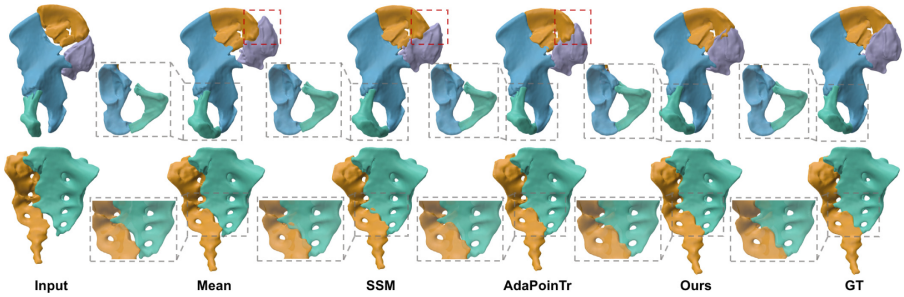


Fig. 2. Example reduction planning results for hipbone and sacrum fractures.

Table 1. Quantitative results on reduction planning for hipbone and sacrum fractures.

Method	Unilateral hipbone (76)				Bilateral hipbone (74)				Sacrum (41)			
	Rot.	Trans.	CD	PA	Rot.	Trans.	CD	PA	Rot.	Trans.	CD	PA
Mean [4]	6.07	5.25	6.51	67.43	7.48	5.89	8.20	61.26	6.57	3.65	9.90	56.91
SSM [21]	5.74	4.69	5.67	74.56	6.74	5.29	7.08	62.61	6.19	3.60	9.74	60.57
Mirror [26]	5.18	3.01	3.10	89.80	–	–	–	–	–	–	–	–
Matching [8]	6.40	3.08	4.74	87.17	–	–	–	–	–	–	–	–
AdaPoinTr [24]	5.08	3.04	2.85	91.89	6.24	3.16	3.27	90.77	5.74	2.44	6.38	78.05
Our method	**4.46**	**2.35**	**2.04**	**97.04**	**4.57**	**2.51**	**1.79**	**95.72**	**4.63**	**2.04**	**4.73**	**82.11**

3.2 Comparison of Reduction Planning Methods

The performance of our method was evaluated on unilateral hipbone fractures, bilateral hipbone fractures, and sacrum fractures. Results in Fig. 2 and Table 1 show that our method outperformed all baseline methods across all fracture types. Mirrored-template methods achieved competitive results due to their

reliance on a contralateral healthy template, however, inherent anatomical asymmetries often led to significant remaining errors, especially in complex fracture configurations. Fracture surface matching lacks robustness due to its inability to accurately identify and align fracture surfaces. Mean templates and statistical shape templates, though applicable to all cases, struggled with individual variability and parameter optimization, leading to suboptimal performance. AdaPoinTr, which was also trained using the synthesized fracture dataset, outperformed the non-learning methods. Paired t-test indicated that the improvements provided by our method is statistical significant, with $p < 0.05$ across all metrics. Our method achieved 2.34 mm translation and 4.54 ° rotation errors, well within clinically accepted thresholds for fracture reduction [18].

3.3 Ablative Experiment

The contributions of the network and data simulation components were analyzed in an ablative study on the hipbone dataset. As shown in Table 2, replacing the FAPE with standard unified encoding resulted in significant performance degradation, highlighting its importance for fragment shape and pose differentiation. The incorporation of fracture label embedding expedited convergence and ensured reconstruction consistency. And removing the recursive strategy significantly reduced alignment accuracy, underscoring its role in refining fragment poses. For fractured data generation, removing key techniques such as fracture surface generation, fragment morphing, and distortion led to noticeable drops in simulation quality. Fracture surface generation introduced critical variations, while distortion modeling captured clinically relevant misalignments. Morphing fracture areas increased data diversity, further boosting model robustness.

Table 2. Ablation study results. The left section evaluates different network components, while the right section examines variations in data simulation.

(a) Network Model Ablation					(b) Data Simulation Ablation				
Method	Rot.	Trans.	CD	PA	Method	Rot.	Trans.	CD	PA
w/o FAPE	5.22	2.52	2.14	95.50	w/o surface	5.34	3.20	2.93	90.33
w/o label	4.86	2.43	2.54	95.61	w/o morph.	5.17	2.56	2.06	94.89
w/o recursion	5.30	2.54	2.73	95.72	w/o distort.	5.30	2.79	2.31	95.00
Our method	**4.51**	**2.43**	**1.92**	**96.39**	Our method	**4.51**	**2.43**	**1.92**	**96.39**

4 Discussion and Conclusion

We present an integrated approach for realistic fracture simulation and robust reduction planning, addressing the challenges of learning fracture-specific features and the scarcity of training data. Our transformer-based network leverages

fracture-aware encoding, enabling accurate and generalizable fracture reduction, even in complex cases such as bilateral and sacrum fractures. To generate diverse and realistic synthetic fractures, our approach learns clinical fracture patterns from limited samples. This sim-to-real strategy enables the training of robust models without the need for extensive real fracture datasets. Additionally, by incorporating recursive inference and registration, our method achieves precise alignment and resilience against large displacements.

Extensive experiments demonstrate the superiority of our approach over traditional and learning-based methods, showing significant improvements in planning accuracy. Ablation studies further validate the contributions of individual components in both the reduction planning framework and the fracture data generation process. Additionally, our method achieved an average runtime of 1.45 s, a significant improvement over 42 min in the SSM-based method [7].

Since direct statistical modeling of fracture patterns in a latent space is challenging, we instead map clinical fractures onto simulated anatomies and introduce further variations. Preliminary results confirm the robustness and generalizability of this approach, with training on a limited number of fracture modes (38 for the hipbone and 11 for the sacrum) proving sufficient to infer unseen fracture patterns. However, failure cases still occur when fragments lack distinct geometric features, hindering accurate alignment. Future work will explore semantic-aware reconstruction to enhance fragment alignment, and validate the generalizability to irregular fractures and other anatomies.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant No. NSFC6247010104), Beijing Science and Technology Project (Grant No. Z221100003522007), and Beijing Jishuitan Research Funding (Grant No. ZR-202504).

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Adams, J., Elhabian, S.: Point2ssm++: self-supervised learning of anatomical shape models from point clouds. arXiv preprint arXiv:2405.09707 (2024)
2. Attene, M.: A lightweight approach to repairing digitized polygon meshes. Vis. Comput. **26**, 1393–1406 (2010)
3. Deng, Z., et al.: Tassembly: data-driven fractured object assembly using a linear template model. Comput. Graph. **113**, 102–112 (2023)
4. Ead, M.S., Palizi, M., Jaremko, J.L., Westover, L., Duke, K.K.: Development and application of the average pelvic shape in virtual pelvic fracture reconstruction. Int. J. Med. Robot. **17**(2), e2199 (2021). https://doi.org/10.1002/rcs.2199
5. Ead, M.S., Westover, L., Polege, S., McClelland, S., Jaremko, J.L., Duke, K.K.: Virtual reconstruction of unilateral pelvic fractures by using pelvic symmetry. Int. J. Comput. Assist. Radiol. Surg. **15**(8), 1267–1277 (2020). https://doi.org/10.1007/s11548-020-02140-z
6. Gnat, R., Saulicz, E., Biały, M., Kłaptocz, P.: Does pelvic asymmetry always mean pathology? Analysis of mechanical factors leading to the asymmetry. J. Hum. Kinet. **21**(2009), 23–32 (2009). https://doi.org/10.2478/v10078-09-0003-8

7. Han, R., et al.: Fracture reduction planning and guidance in orthopaedic trauma surgery via multi-body image registration. Med. Image Anal. **68**, 101917 (2021). https://doi.org/10.1016/j.media.2020.101917
8. Liu, J., et al.: An end-to-end geometry-based pipeline for automatic preoperative surgical planning of pelvic fracture reduction and fixation. IEEE Trans. Med. Imaging (2024). https://doi.org/10.1109/TMI.2024.3429403
9. Liu, P., et al.: Deep learning to segment pelvic bones: large-scale CT datasets and baseline models. Int. J. Comput. Assist. Radiol. Surg. **16**(5), 749–756 (2021). https://doi.org/10.1007/s11548-021-02363-8
10. Liu, Y., et al.: Pelvic fracture segmentation using a multi-scale distance-weighted neural network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 312–321. Springer (2023). https://doi.org/10.1007/978-3-031-43996-4_30
11. Liu, Y., et al.: Preoperative fracture reduction planning for image-guided pelvic trauma surgery: a comprehensive pipeline with learning. Med. Image Anal. 103506 (2025). https://doi.org/10.1016/j.media.2025.103506
12. Lu, S., et al.: Preoperative virtual reduction planning algorithm of fractured pelvis based on adaptive templates. IEEE Trans. Biomed. Eng. **70**(10), 2943–2954 (2023). https://doi.org/10.1109/TBME.2023.3272007
13. Luque-Luque, A., Pérez-Cano, F.D., Jiménez-Delgado, J.J.: Complex fracture reduction by exact identification of the fracture zone. Med. Image Anal. **72**, 102120 (2021). https://doi.org/10.1016/j.media.2021.102120
14. Olson, S.A., Furman, B., Guilak, F.: Joint injury and post-traumatic arthritis (2012). https://doi.org/10.1007/s11420-011-9247-7
15. Paulano-Godino, F., Jiménez-Delgado, J.J.: Identification of fracture zones and its application in automatic bone fracture reduction. Comput. Methods Programs Biomed. **141**, 93–104 (2017). https://doi.org/10.1016/j.cmpb.2016.12.014
16. Phan, A.V., Nguyen, M., Nguyen, Y.L.H., Bui, L.T.: Dgcnn: a convolutional neural network over large-scale labeled graphs. Neural Netw. **108**, 533–543 (2018). https://doi.org/10.1016/j.neunet.2018.09.001
17. Sellán, S., Chen, Y.C., Wu, Z., Garg, A., Jacobson, A.: Breaking bad: a dataset for geometric fracture and reassembly. Adv. Neural. Inf. Process. Syst. **35**, 38885–38898 (2022)
18. Smith, W., et al.: Clinical outcomes of unstable pelvic fractures in skeletally immature patients. JBJS **87** (2005). https://doi.org/10.2106/JBJS.C.01244v
19. Ulbrich, M., et al.: Advantages of a training course for surgical planning in virtual reality for oral and maxillofacial surgery: crossover study. JMIR Ser. Games **11**, e40541 (2023)
20. Willis, A.R., Anderson, D.D., Thomas, T.P., Brown, T.D., Marsh, J.L.: 3d reconstruction of highly fragmented bone fractures. In: SPIE Medical Imaging (2007). https://doi.org/10.1117/12.708683
21. Yibulayimu, S., et al.: Pelvic fracture reduction planning based on morphable models and structural constraints. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 322–332. Springer (2023). https://doi.org/10.1007/978-3-031-43996-4_31
22. Yibulayimu, S., et al.: Automatic pelvic structure restoration: a sim-to-real approach via recursive pose estimation network. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2024). https://doi.org/10.1109/ISBI56570.2024.10635211

23. Yu, X., Rao, Y., Wang, Z., Liu, Z., Lu, J., Zhou, J.: Pointr: diverse point cloud completion with geometry-aware transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12498–12507 (2021)
24. Yu, X., Rao, Y., Wang, Z., Lu, J., Zhou, J.: Adapointr: diverse point cloud completion with adaptive geometry-aware transformers. IEEE Trans. Pattern Anal. Mach. Intell. **45**(12), 14114–14130 (2023). https://doi.org/10.1109/TPAMI.2023.3309253
25. Zeng, B., Wang, H., Tao, X., Shi, H., Joskowicz, L., Chen, X.: A bidirectional framework for fracture simulation and deformation-based restoration prediction in pelvic fracture surgical planning. Med. Image Anal. **97**, 103267 (2024). https://doi.org/10.1016/j.media.2024.103267
26. Zhang, L.H., et al.: Computer-aided pelvic reduction frame for anatomical closed reduction of unstable pelvic fractures. J. Orthopaedic Res. **34**(1), 81–87 (2016). https://doi.org/10.1002/jor.22987
27. Zhao, C., Cao, Q., Sun, X., Wu, X., Zhu, G., Wang, Y.: Intelligent robot-assisted minimally invasive reduction system for reduction of unstable pelvic fractures. Injury **54**(2), 604–614 (2023). https://doi.org/10.1016/j.injury.2022.11.001
28. Zhao, C., et al.: Automatic reduction planning of pelvic fracture based on symmetry. Comput. Methods Biomech. Biomed. Eng. Imaging Vis. **10**(6), 577–584 (2022). https://doi.org/10.1080/21681163.2021.2012830



SMF-Net: Unlocking Multimodal Insights for Enhanced Stroke Lesion Segmentation

Meklit Atlaw[1], Geng Chen[1(✉)], Haotian Jiang[1], Xuyun Wen[2], Hengfei Cui[1], and Yong Xia[1]

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China
meklitmaki29@gmail.com, geng.chen@ieee.org, jianghaotian100@163.com, {hfcui,yxia}@nwpu.edu.cn

[2] College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China
wenxuyun@nuaa.edu.cn

Abstract. Stroke is a leading cause of death and disability worldwide, necessitating accurate lesion segmentation for effective diagnosis and treatment. Multimodal images provide complementary insights into stroke detection and progression. However, existing segmentation methods often struggle to fully leverage the distinct and dynamic sensitivities of these modalities. Current approaches, including encoder-decoder networks and SAM-based models, are either limited to single-modality data or rely on suboptimal fusion techniques, hindering their ability to adapt to the distinct nature of stroke lesions. To address these challenges, we propose SAM-driven Multimodal Fusion Network (SMF-Net) for enhanced stroke lesion segmentation. SMF-Net incorporates a multimodal Siamese image encoder based on the Swin Transformer to extract modality-specific features, alongside two novel fusion strategies: (1) Complementary dynamic fusion module, which uses pairwise co-attention and dynamic learnable weights to model interdependencies and adaptively combine multimodal features; and (2) Context-aware intermediate-layer fusion module, a lightweight, multi-layer fusion mechanism that captures multiscale features while preserving modality-specific information. Extensive experiments on an open benchmark dataset demonstrate that SMF-Net outperforms previous stroke lesion segmentation methods through effective multimodal integration.

Keywords: Stroke lesion segmentation · MRI · Multimodal fusion · SAM

1 Introduction

Stroke is a major neurological disorder, the second most common cause of death, and the third leading cause of disability worldwide. Its rising prevalence results in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 598–608, 2026.
https://doi.org/10.1007/978-3-032-04947-6_57

millions of deaths annually, leaving many survivors with lasting impairments [3]. Accurate stroke lesion segmentation in medical imaging is crucial for assessing lesion extent and severity, enabling timely clinical decisions to optimize treatment. Various imaging modalities capture different aspects of stroke evolution and location [11]. MRI techniques, such as Diffusion-Weighted Imaging (DWI), Apparent Diffusion Coefficient (ADC), and Fluid-Attenuated Inversion Recovery (FLAIR) provide critical insight into the progression and location of the lesion. Although the use of multiple imaging modalities improves the accuracy of diagnosis and segmentation, the dynamic nature of brain tissue damage and the varying imaging sensitivities require effective integration strategies for accurate segmentation of stroke lesions.

Recent studies have explored various deep learning-based approaches for stroke lesion segmentation, particularly using encoder-decoder networks [1,2,9,13,16,17]. D-UNet [25] combines 2D and 3D features to capture complex lesion patterns, while X-Net[14] focuses on capturing long-range dependencies to enhance segmentation accuracy. SAN-Net[21] is tailored for multi-site stroke lesion segmentation, while W-Net [19] addresses fuzzy stroke boundaries to improve precision. While these models have proven effective for single-modality data, efforts to address multi-modal imaging have led to methods such as [5,6,15], which attempt to integrate different modalities using image stacking or basic fusion techniques. More recently, the Segment Anything Model (SAM) [8] has inspired a wave of universal segmentation models, including MedSAM [12], Mobile-SAM [22], and EfficientSAM [20], which have shown promise in various medical imaging tasks.

Despite the advancements, these methods face significant challenges in stroke lesion segmentation. Most approaches are primarily single-modality focused, limiting their ability to capture the complementary information offered by different imaging techniques. When it comes to multimodal strategies, existing methods often rely on basic fusion techniques, such as stacking images from different modalities or using simple sub-networks for modality fusion. These methods often overlook the distinct sensitivities of each modality, leading to suboptimal utilization of modality-specific features and an inability to adapt to their changing relevance during stroke progression. Furthermore, SAM-based models and their adaptations, while effective in other tasks, are not yet optimized for the unique intensity patterns and lesion characteristics specific to stroke. In addition, these models are not tailored for multimodal integration, an essential aspect of accurate stroke diagnosis and clinical decision making. Therefore, addressing these gaps is crucial to improving stroke lesion segmentation.

To this end, we propose SAM-driven Multimodal Fusion Network (SMF-Net), a SAM-based framework leveraging Siamese networks and optimized fusion strategies to overcome challenges in multimodal stroke lesion segmentation. Particularly, we first use a multimodal Siamese image encoder based on the Swin Transformer to extract distinct features from each modality. We then introduce the Complementary Dynamic Fusion (CDF) with pairwise co-attention that models pairwise relationships between the modalities, capturing interde-

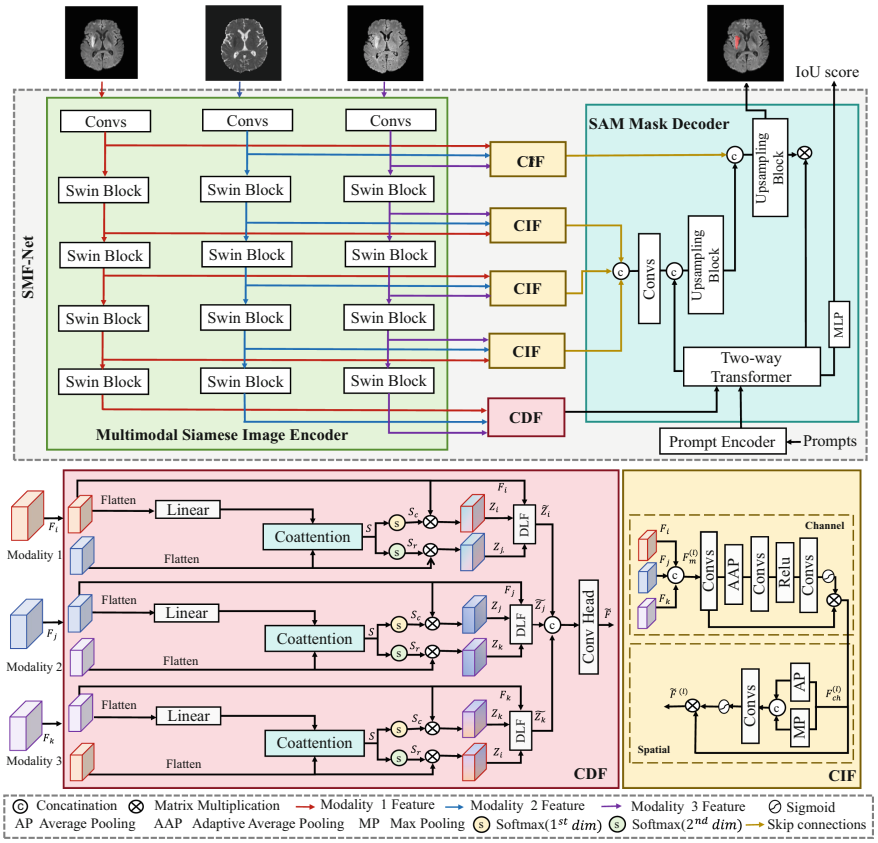


Fig. 1. An overview of the proposed SMF-Net and its components, CDF and CIF.

pendencies and enhancing feature representation, and dynamic learnable fusion that combines features from all modalities, adapting based on learned weights to optimize integration. Finally, we introduce Context-aware Intermediate-layer Fusion (CIF), a lightweight, multi-layer attention-based fusion strategy that captures features at multiple scales while preserving modality-specific information. CIF integrates features through skip connections between the encoder and decoder, enhancing the integration of multiscale features. Extensive experiments on open benchmark dataset demonstrate that SMF-Net improves multimodal stroke lesion segmentation by effectively integrating modality-specific information, outperforming previous methods.

2 Methodology

In this section, we first provide a comprehensive overview of SMF-Net in Subsect. 2.1. We then present a detailed explanation of its core components in Subsects. 2.2 and 2.3, outlining their structure and functional contributions.

2.1 Network Architecture

The overview of SMF-Net is shown in Fig. 1. Given input image modalities $I_i \in \mathbb{R}^{h\times w\times c}$, where h, w and c denote the height, width, and channel dimensions of the image and i denote different modalities, the multimodal Siamese encoder, based on a modified Swin Transformer [4], extracts modality-specific features. Using shared weights across modalities, the encoder ensures consistent feature extraction while maintaining computational efficiency. For each layer l, the encoder generates modality-specific feature maps $F_i^{(l)} \in \mathbb{R}^{h_l\times w_l\times c_l}$, where h_l, w_l, and c_l denote the spatial height, width, and channel dimensions of the feature in that layer. The CDF module simultaneously models the cross-modal relationships between modalities, refines, and fuse feature representations extracted from the final layer of the encoder. First, the CDF module captures the interactions between modality-specific features through pairwise co-attention, facilitating the integration of complementary information while preserving modality-specific characteristics. Then, adaptive weighting dynamically refines and fuses representations, ensuring that the most relevant modality contributions are emphasized while suppressing less informative features. Finally, the resulting features are concatenated and passed through a convolution operation to produce a unified encoded feature representation $F_{\text{CDF}}^{(l)} \in \mathbb{R}^{h_l\times w_l\times c_l}$. To preserve multiscale information, the CIF module dynamically fuses features from each intermediate layer of the encoder. This strategy captures modality-specific information in multiple layers, producing a fused representation $F_{\text{CIF}}^{(l)} \in \mathbb{R}^{h_l\times w_l\times c_l}$ in each layer. These fused multilayer features are then used as skip connections, enabling better multiscale feature propagation between the encoder and decoder. Skip connections are formed by concatenating the deeper $F_{\text{CIF}}^{(l)}$ outputs for early upsampling, while the shallowest output is integrated later to refine spatial details. At the final scale, the $F_{\text{CDF}}^{(l)}$ module fuses features and serves as a deep skip connection by providing a global image embedding to the decoder. The final segmentation output is generated by processing these fused features (i.e., $F_{\text{CDF}}^{(l)}$ and $F_{\text{CIF}}^{(l)}$) through a SAM mask decoder [8]. In addition, a prompt mechanism is kept the same as [4] to further refine the segmentation results.

2.2 Complementary Dynamic Fusion Module

To effectively model modality relationships while preserving modality-specific characteristics, we use a multimodal Siamese image encoder with shared weights, ensuring consistent feature extraction. The final-layer features F_i that contain deeper features are processed and fused by the CDF module, which integrates complementary information among modalities. CDF consists of pairwise

co-attention to model interactions and dynamic learnable fusion to adaptively combine co-attended features. Inspired by [10], we apply co-attention to model modality interactions, enabling the network to capture modality relationships and dynamically emphasize complementary features.

Pairwise co-attention. Given two modality-specific feature maps $F_i \in \mathbb{R}^{h\times w\times c}$ and $F_j \in \mathbb{R}^{h\times w\times c}$, extracted from the final layer of the encoder, the co-attention mechanism captures pairwise modality relationships. In particular, we first flatten the feature maps F_i and F_j. Then, F_i is projected using a learnable weight matrix W. The co-attention mechanism computes the attention matrix S by evaluating the pairwise interaction between F_i and F_j as:

$$S = F_i^\top W F_j, \tag{1}$$

where $W \in \mathbb{R}^{c\times c}$ is a learned weight matrix, and $F_i, F_j \in \mathbb{R}^{c\times hw}$ are the flattened feature maps of modalities i and j, respectively. We then compute the softmax along both the column and row dimensions of S, resulting in attention weights:

$$S_\mathrm{c} = \mathrm{softmax}(S, \mathrm{dim} = 1), \quad S_\mathrm{r} = \mathrm{softmax}(S, \mathrm{dim} = 2), \tag{2}$$

where S_c and S_r represent the column and row-wise attention weights, respectively. Using these attention maps, we generate co-attended feature representations by applying attention-weighted matrix multiplication:

$$Z_i = F_j S_\mathrm{c}, \quad Z_j = F_i S_\mathrm{r}^\top, \tag{3}$$

where Z_i and Z_j are the refined co-attended feature maps. Pairwise co-attention emphasizes informative regions across modalities, enhancing shared feature representation.

Dynamic Learnable Fusion. To integrate complementary information from multiple modalities, we introduce a dynamic learnable fusion mechanism that adaptively assigns spatial attention weights to the original and co-attended features via element-wise multiplication, producing the refined modality-specific feature $\tilde{Z}_i$ as:

$$S_\alpha, S_\lambda, S_\gamma = \mathrm{softmax}(\mathrm{Conv}(\mathrm{Concat}(F_i, Z_i, Z_j)), \tag{4}$$

$$\tilde{Z}_i = S_\alpha \circ F_i + S_\lambda \circ Z_i + S_\gamma \circ Z_j, \tag{5}$$

where $\mathrm{Concat}(\cdot)$ represents a concatenation operation followed by a convolution operation $\mathrm{Conv}(\cdot)$ with kernel size 1 and softmax function. The weights S_α , S_λ, and S_γ represent the learned importance of F_i, Z_i and Z_j, respectively. Once modality-specific features are refined, the final multimodal feature representation $\tilde{F}$ is calculated as:

$$\tilde{F} = \mathrm{Conv}(\mathrm{Concat}(\tilde{Z}_i, \tilde{Z}_j, \tilde{Z}_k)), \tag{6}$$

where $\tilde{Z}_i, \tilde{Z}_j$, and $\tilde{Z}_k$ denote the refined modality-specific feature corresponding to different imaging modalities.

2.3 Context-aware Intermediate-layer Fusion Module

We integrate multimodal features across layers using the CIF module, acknowledging the importance of multilayer fusion as highlighted in [23,24]. Modality-specific features from the intermediate layer undergo lightweight fusion followed by attention-based refinement. Given intermediate-layer modality features $F^{(l)}$, we concatenate them along the channel dimension and apply a convolution block to learn a weighted combination $F_{\mathrm{m}}^{(l)}$. We further apply sequential channel and spatial attention, following [18], which employs lightweight attention for adaptive refinement. Channel attention assigns modality-wise weights and refines $F_{\mathrm{m}}^{(l)}$, yielding channel-refined feature map $F_{\mathrm{ch}}^{(l)}$ via element-wise multiplication, as:

$$F_{\mathrm{ch}}^{(l)} = \mathcal{F}_{\mathrm{ch}}(F_{\mathrm{m}}^{(l)}) \circ F_{\mathrm{m}}^{(l)}, \tag{7}$$

where $\mathcal{F}_{\mathrm{ch}}(\cdot)$ denotes channel attention. $\mathcal{F}_{\mathrm{ch}}(\cdot)$ comprises of adaptive average pooling, two convolution operations with kernel size 1 and ReLU activation in between, and a final sigmoid function. To capture spatially significant features, we apply spatial attention. $F_{\mathrm{ch}}^{(l)}$ undergoes average and max pooling, followed by concatenation and convolution operation with kernel size 7 yielding the spatial attention map. This map refines $F_{\mathrm{ch}}^{(l)}$, producing layerwise multimodal feature representation $\tilde{F}^{(l)}$ via element-wise multiplication, as:

$$\tilde{F}^{(l)} = \mathcal{F}_{\mathrm{sp}}(F_{\mathrm{ch}}^{(l)}) \circ F_{\mathrm{ch}}^{(l)}, \tag{8}$$

where $\mathcal{F}_{\mathrm{sp}}(\cdot)$ represents spatial attention refinement. After modality-specific fusion at each layer, the fused features propagate through skip connections, preserving high-resolution details while deeper layers retain contextual information.

3 Experiments

3.1 Dataset

We evaluate SMF-Net on the Ischemic Stroke Lesion Segmentation (ISLES) 2022 dataset [7]. This dataset consists of 250 multimodal and multicenter MRI modalities for the segmentation of acute to subacute stroke lesions. Each scan includes DWI, ADC, and FLAIR sequences.

3.2 Implementation Details

Proper alignment between modalities is crucial in multimodal learning, and since FLAIR images were misaligned with DWI and ADC, we applied image registration to ensure alignment. Voxel dimensions were standardized to $2 \times 2 \times 2\,\mathrm{mm}^3$, and images were resized to 256×256 pixels using cropping and padding. We followed the preprocessing pipeline in [12]. Then, we included a margin of three axial slices before and after the lesion to provide additional context and ensure a balanced distribution of training samples. DWI consistently yielded the highest

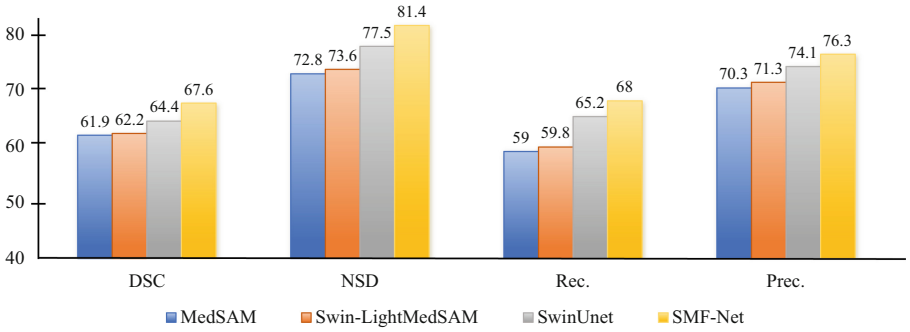


Fig. 2. Comparison of SMF-Net with previous medical image segmentation methods on the ISLES 2022 dataset.

Table 1. Ablation study results evaluating the effectiveness of different components in the SMF-Net and evaluating the contribution of different modalities.

Modality			Fusion			Evaluation Performance			
ADC	FLAIR	DWI	CDF	DLF	CIF	DSC	NSD	Rec.	Prec.
✓	✗	✗	✗	✗	✗	42.2	46.7	50.0	43.0
✗	✓	✗	✗	✗	✗	48.2	56.1	50.9	52.7
✗	✗	✓	✗	✗	✗	62.2	73.6	59.8	71.3
✓	✓	✓	✓	✗	✗	66.0	77.7	60.4	75.3
✓	✓	✓	✓	✗	✓	67.2	80.5	67.7	74.0
✓	✓	✓	✓	✓	✗	67.3	79.6	65.2	72.7
✓	✓	✓	✓	✓	✓	67.6	81.4	68.0	76.3

segmentation performance and was selected as our baseline modality due to its clinical relevance in acute stroke detection. The dataset was split 80%-20% for training and testing. Training was conducted on an NVIDIA GeForce RTX 3090 (24GB RAM) with a batch size of 4 for 70 epochs. We employed the AdamW optimizer with an initial learning rate of 1×10^{-4}, adjusted via the ReducedLROnPlateau scheduler. We use dice loss to optimize spatial overlap, focal loss to handle class imbalance by focusing on hard examples, and MSE loss to supervise the IoU prediction head. We followed the data augmentation strategy in [4], additionally we applied random affine transformations, including rotations ($\pm 90°$), scaling (0.8–1.2), and translations (± 10 pixels) with 50% probability. The network performance was evaluated using widely used metrics: Dice Score (DSC), Dice(NSD), Recall, and Precision.

3.3 Quantitative and Qualitative Results

To assess the performance of SMF-Net, we compared it with widely adopted segmentation approaches. These include Swin-Unet [1], MedSAM [12], and Swin-

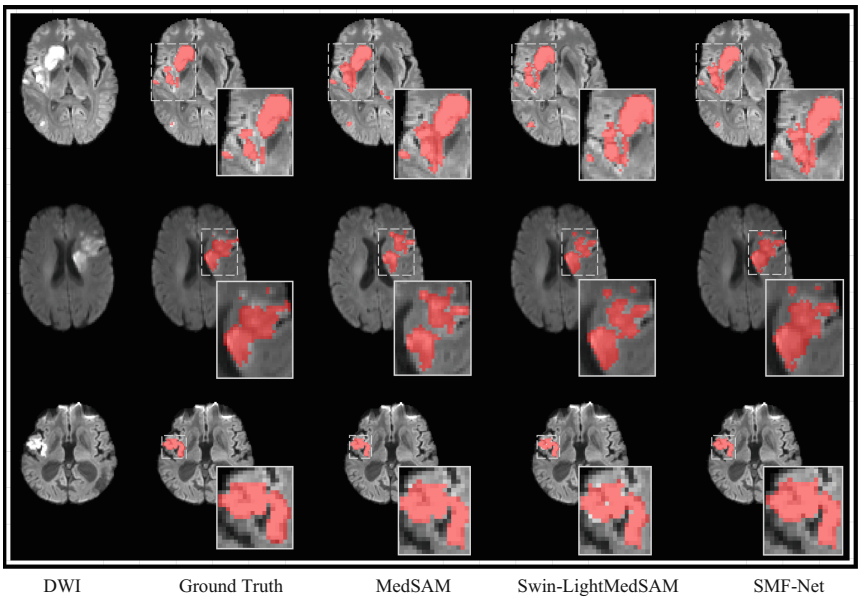


Fig. 3. Qualitative results of stroke segmentation on ISLES 2022 datasets, with the corresponding lesion area in red. (Color figure online)

LightMedSAM [4]. The results are shown in Fig. 2. SMF-Net demonstrates superior performance among SAM-based methods, achieving the highest DSC (67.6%), outperforming both MedSAM (DSC: 61.9%) and Swin-LightMedSAM (DSC: 62.2%). Furthermore, SMF-Net achieves a significantly greater performance gain over SwinUnet. A paired t-test between SMF-Net and the baseline on DSC scores confirms the improvement is statistically significant ($p < 0.005$). Furthermore, the qualitative assessment of the SMF-Net and SAM-based methods on ISLES 2022 is visually presented in Fig. 3. Swin-LightMedSAM shows improved performance over MedSAM but still leaves some gaps in the segmenting lesion regions. However, SMF-Net enhances lesion segmentation by reducing missing regions compared to other methods.

3.4 Ablation Study

We conducted an ablation study to evaluate the contribution of modalities and fusion strategies in SMF-Net. In Table 1, we summarize the ablation study results that assess the contribution of each imaging modality (ADC, FLAIR, DWI) and the fusion strategies (CDF, Dynamic Learnable Fusion (DLF), and CIF) in SMF-Net. Among the single-modality experiments, using DWI achieves the highest performance (DSC: 62.2%, NSD: 73.6%), likely because of its high sensitivity to ischemic regions. In contrast, using ADC produces the lowest scores (DSC: 42.2%, NSD: 46.7%) due to limited contrast, while using FLAIR falls in between

(DSC: 48.2%, NSD: 56.1%), highlighting its complementary role in stroke imaging. Building on the DWI baseline, introducing the CDF module yields a +3.8% increase in DSC and a +4.1% increase in NSD, emphasizing the benefits of multimodal fusion based on cross-modal interactions. Adding the CIF module further boosts performance by enabling multiscale feature fusion, whereas the DLF refines feature integration but does not replace CIF's benefits. The SMF-Net, integrating CDF, CIF, and DLF achieves the best overall performance in all evaluation metrics. These improvements underscore the importance of leveraging multimodal data and complementary fusion strategies for accurate stroke lesion segmentation.

4 Conclusion

In this work, we proposed SMF-Net, a multimodal stroke lesion segmentation framework that effectively integrates complementary information from multiple modalities. The model utilizes a multimodal Siamese image encoder for modality-specific feature extraction while maintaining computational efficiency. We introduced CDF with pairwise co-attention to enhance cross-modal interactions, dynamic learnable fusion for adaptive weighted fusion, and CIF to capture multiscale information. Each component of SMF-Net was evaluated, demonstrating its contribution to the overall performance. Experimental results show that SMF-Net outperforms existing methods, emphasizing the importance of structured multimodal feature integration in stroke lesion segmentation.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62201465 and 62271405.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cao, H., et al.: Swin-unet: unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision, pp. 205–218. Springer (2022)
2. Chen, J., et al.: Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
3. Feigin, V.L., Norrving, B., Mensah, G.A.: Global burden of stroke. Circ. Res. **120**(3), 439–448 (2017)
4. Gao, R., Lyu, D., Staring, M.: Swin-LiteMedSAM: a lightweight box-based segment anything model for large-scale medical image datasets. In: Medical Image Segmentation Challenge, pp. 70–82. Springer (2024)
5. Garcia-Salgado, B.P., et al.: Enhanced ischemic stroke lesion segmentation in MRI using attention u-net with generalized dice focal loss. Appl. Sci. **14**(18), 8183 (2024)
6. Gheibi, Y., Shirini, K., Razavi, S.N., Farhoudi, M., Samad-Soltani, T.: CNN-res: deep learning framework for segmentation of acute ischemic stroke lesions on multimodal MRI images. BMC Med. Inform. Decis. Mak. **23**(1), 192 (2023)

7. Hernandez Petzsche, M.R., et al.: ISLES 2022: a multi-center magnetic resonance imaging stroke lesion segmentation dataset. Sci. Data **9**(1), 762 (2022)
8. Kirillov, A., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
9. Liu, N., Li, H.: CMU-net: a cascaded mini u-network for retinal vessel segmentation. SSRN Electron. J. (2022)
10. Lu, X., Wang, W., Shen, J., Crandall, D., Luo, J.: Zero-shot video object segmentation with co-attention siamese networks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(4), 2228–2242 (2020)
11. Luo, J., Dai, P., He, Z., Huang, Z., Liao, S., Liu, K.: Deep learning models for ischemic stroke lesion segmentation in medical images: a survey. Comput. Biol. Med. 108509 (2024)
12. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nat. Commun. **15**(1), 654 (2024)
13. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
14. Qi, K., et al.: X-net: brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 247–255. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_28
15. Raju, C.S.P., Neelapu, B.C., Laskar, R.H., Muhammad, G.: Analysis of multimodal fusion strategies in deep learning for ischemic stroke lesion segmentation on computed tomography perfusion data. Multimedia Tools Appl. 1–26 (2024)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: gated axial-attention for medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4
18. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
19. Wu, Z., Zhang, X., Li, F., Wang, S., Huang, L., Li, J.: W-net: a boundary-enhanced segmentation network for stroke lesions. Expert Syst. Appl. **230**, 120637 (2023)
20. Xiong, Y., et al.: Efficientsam: leveraged masked image pretraining for efficient segment anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16111–16121 (2024)
21. Yu, W., Huang, Z., Zhang, J., Shan, H.: SAN-net: learning generalization to unseen sites for stroke lesion segmentation with self-adaptive normalization. Comput. Biol. Med. **156**, 106717 (2023)
22. Zhang, C., et al.: Faster segment anything: towards lightweight sam for mobile applications. arXiv preprint arXiv:2306.14289 (2023)
23. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-net: hybrid-fusion network for multi-modal mr image synthesis. IEEE Trans. Med. Imaging **39**(9), 2772–2781 (2020)

24. Zhou, T., Zhou, Y., Li, G., Chen, G., Shen, J.: Uncertainty-aware hierarchical aggregation network for medical image segmentation. IEEE Trans. Circ. Syst. Video Technol. (2024)
25. Zhou, Y., Huang, W., Dong, P., Xia, Y., Wang, S.: D-UNet: a dimension-fusion U shape network for chronic stroke lesion segmentation. IEEE/ACM Trans. Comput. Biol. Bioinf. **18**(3), 940–950 (2019)



Spatio-Temporal Pre-Trained Foundation Model for Neural Decoding with Fine-Grained Optimization

Ziyu Li[1], Zhiyuan Zhu[2], Yang Bai[3], Qing Li[1], and Xia Wu[1(✉)]

[1] Beijing Institute of Technology, Beijing 100081, China
wuxia@bit.edu.cn
[2] Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[3] Beijing Century TAL Education Technology Co., Ltd., Beijing 102200, China

Abstract. Traditional neural decoding methods are heavily based on fully annotated brain data, which are both expensive to produce and scarce in availability. This limitation hinders the development of accurate and generalizable decoding models. Drawing inspiration from the success of foundational AI models in reducing dependency on annotated data in fields such as natural language processing, we introduce a novel foundation model that leverages the inherent spatiotemporal covariation of functional brain networks, which enables effective neural decoding with minimal annotation requirements. Our framework incorporates three key innovations: 1) A spatiotemporal importance-guided augmentation strategy is designed to capture the synergistic relationships between brain regions and their dynamic changes; 2) A progressive spatiotemporal-aware encoder is proposed to learn local-to-global brain interaction information; 3) A fine-grained consistency optimization technique is developed to enhance the representations of overall brain function. Evaluations of publicly available fMRI datasets demonstrate that our proposed framework not only achieves superior decoding performance, but also exhibits strong generalizability and reveals patterns of nervous activity. Our research advances brain representation learning and provides an innovative solution for universal neural decoding models.

Keywords: Neural decoding · Spatiotemporal · Self-supervised learning · fMRI

1 Introduction

Neural decoding involves the systematic analysis and interpretation of neural activity patterns to infer brain states, cognitive processes, or behavioral intentions. This technique has significant applications in the elucidation of brain mechanisms, disease diagnosis, and the development of brain-inspired artificial intelligence [4]. Functional magnetic resonance imaging (fMRI) has garnered considerable attention in neural decoding due to its ability to capture cooperative

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 609–618, 2026.
https://doi.org/10.1007/978-3-032-04947-6_58

relationships between regions of interest (ROIs) within the brain [2]. Recently, Graph Neural Networks (GNNs) have gained widespread application in fMRI-based neural decoding due to their robust capability to learn complex contextual representations and provide interpretability [6] [16]. However, the superior performance of these models frequently depends on fully annotated fMRI datasets, which require expert knowledge for precise labeling. This dependency may undermine both the accuracy and generalizability of neural decoding models [11].

In contrast, graph foundation models leverage a substantial amount of unlabeled data through pre-training and fine-tuning techniques to adapt to complex graph tasks [9], thereby significantly reducing the cost of fMRI data annotation. During pre-training, general semantic information is extracted via GNNs-based self-supervised learning (GSSL), which facilitates downstream decoding tasks. Among them, contrastive-based GSSL has gained attention for its simplicity and efficiency [15], which predominantly focused on ROI connection analysis. However, this narrow-focused approach has a glaring limitation: it completely overlooks the crucial temporal dynamics within brain networks. Emerging research has provided compelling evidence that the temporal dependencies between ROIs play a pivotal role in neural activities. By neglecting this aspect, we are essentially missing out on a wealth of information that could potentially revolutionize our understanding of brain functions [12,14]. Numerous recent studies have shown that the incorporation of these temporal dependencies can lead to a substantial leap in self-supervised decoding performance [10]. Therefore, a novel graph foundation model for neural decoding is needed to fill this critical gap, by integrating the ignored temporal dynamics. The main challenges arise: i) How to design an efficient spatiotemporal representation extractor to capture more comprehensive and flexible brain representations, thereby improving the accuracy and reliability of neural decoding? and ii) What are the appropriate spatiotemporal augmented views for brain signals to provide a robust data foundation and enhance the model's capacity to interpret and analyze?

To address the aforementioned challenges, we propose an innovative SpatioTemporal Pre-Training Foundation model (STPTF) for neural decoding based on fMRI. STPTF employs spatiotemporal information of the brain, enabling pre-training with unlabeled fMRI data. With only a small amount of labeled fMRI data for fine-tuning, it can achieve substantial performance improvements in downstream decoding tasks. Specifically, STPTF introduces three key innovations: i) Strategy innovation through SpatioTemporal Importance Guided Augmentation (STIGA), STIGA comprises Spatial Centrality Augmentation (SCA) and Temporal Continuity Augmentation (TCA), which facilitate the model's deep analysis of brain representations from both spatial and temporal views; ii) Architectural innovation via Progressive SpatioTemporal-Aware encoder (PSTA), PSTA includes a simultaneously static and dynamic extractor, a continuous brain state calibration module, and a progressive temporal aggregation module, which help the model to thoroughly analyze global spatiotemporal features of brain activities; iii) Optimization innovation using Fine-Grained Consistency Optimization (FGCO), FGCO can effectively mitigate the suboptimal

representation problem caused by negative samples, which allows the model to fully and accurately interpret brain signals. We evaluate our proposed model, STPTF, on multiple medical imaging datasets, and the results demonstrate its state-of-the-art performance. To the best of our knowledge, this study represents one of the first attempts to integrate graph foundation models with spatio-temporal brain information, underscoring its significant potential in the field of neural decoding.

2 Method

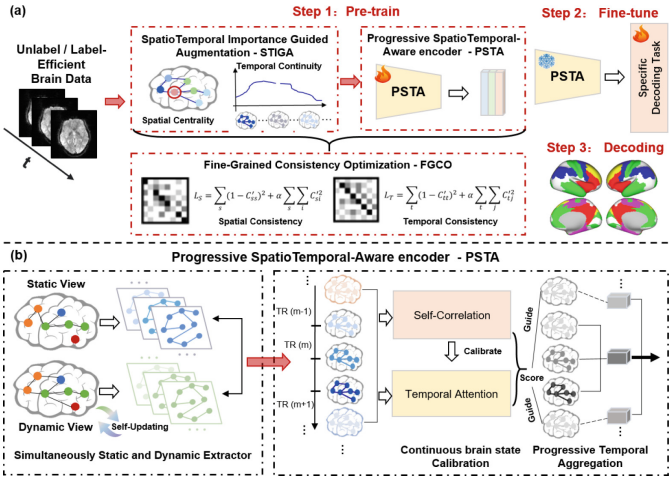


Fig. 1. (a) The overall framework of the proposed SpatioTemporal Pre-Training Foundation model (STPTF) for neural decoding. (b) Progressive SpatioTemporal-Aware encoder (PSTA).

2.1 Overall Framework

Implementing neural decoding based on our proposed STPTF is divided into three stages (Fig. 1): pre-training a general spatiotemporal encoder with unlabeled brain data, performing linear fine-tuning for various decoding tasks using the pre-trained encoder, and subsequently decoding different brain states based on the fine-tuned encoder. Specifically, STPTF integrates three core components to enhance fMRI-based neural decoding: i) Strategy innovation through STIGA; ii) Architectural innovation via PSTA (Fig. 1(b)); iii) Optimization innovation using FGCO.

2.2 SpatioTemporal Importance Guided Augmentation - STIGA

Spatial Centrality Augmentation SCA. Research shows that key ROIs, which serve as hubs in brain networks, frequently engage in interactions with other ROIs, thus facilitating neuronal communication and functional integration [5]. Based on these findings, we introduce SCA to generate an alternative view. In graph theory, nodes with high centrality are considered critical hubs [1]. We select degree centrality as a straightforward yet effective measure of node importance and adaptively assign masking probabilities according to the degree centrality of each ROI. Specifically, hubs are allocated higher masking probabilities, which aids in capturing the global semantic representation of brain networks.

Temporal Continuity Augmentation - TCA. Considering the typically consistent ROI connections between adjacent timestamps and the concentration of semantic information, there could be redundant temporal information, indicating that changes in ROI connections exhibit continuity [13]. Based on this observation, we design the TCA for the other view. Specifically, we define the temporal change rate as the displacements between consecutive timestamps across the entire fMRI time series and adaptively assign mask probabilities according to the change rate at each timestamp. Timestamps with higher change rates are assigned greater mask probabilities, thereby compelling the model to learn the global temporal representation of the brain network.

2.3 Progressive SpatioTemporal-Aware Encoder - PSTA

First, we extract the spatial representation of the original fMRI time series through a simultaneously static and dynamic extractor and then use GNN to capture the latent collaborative relationships among different ROIs. This approach allows us to simultaneously derive both static and dynamic brain representations, thereby reflecting both the integrity and the time-varying characteristics of the brain's functional network. The static and dynamic representations are complementary [8], and their joint learning provides a more comprehensive understanding of brain function. We incorporate automatic weight learning to capture the latent relationship between these two types of features, thus improving the adaptability and robustness of the model.

$$
\begin{aligned}
Z_{Spatial} &= \beta_1 Z_s + \beta_2 Z_d, \\
\beta_c &= \frac{exp(-\alpha_c)}{\sum_{k=1}^{2} exp(-\alpha_k)}, c \in 1, 2
\end{aligned}
\tag{1}
$$

From a static perspective, we construct a static brain network by calculating partial correlation coefficients between pairs of ROIs after excluding potential confounding factors. We then employ GNNs to generate static representations of the brain denoted as Z_s. From a dynamic perspective, we adapt the adaptive

brain topology learning module [7], previously proposed, to learn brain networks that dynamically update with brain activities. By integrating GNN, the dynamic representations Z_d of the brain over time are automatically refined.

Assume that the original fMRI sequence is represented as X. For the learned spatial representations ($Z_{Spatial}$) of the brain, we further explore their progressive temporal representations. We introduce continuous brain state calibration to address the long-range error estimation challenge faced by typical attention mechanisms when establishing timestamp associations. Continuous brain state calibration adaptively determines the extent of the brain state transition information to be aggregated, without being constrained by the length of the fMRI sequence. Consequently, it facilitates the extraction of more refined long-term dependency information within the fMRI sequence and reveals contextual relationships of brain states within an appropriate temporal range.

$$\begin{aligned} Z_{temporal} &= Conv_2(Conv_V(Z_{Spatial}) \cdot CAM) + X, \\ CAM &= Softmax(Conv_1([CCM, TAW])), \\ CCM &= Softmax(Z_{Spatial}^T \cdot Z_{Spatial}), \\ TAW &= Softmax(Conv_Q(Z_{Spatial}) \cdot Conv_K(Z_{Spatial})^T) \end{aligned} \tag{2}$$

In order to improve the ability to model complex dynamic brain activities, we further design a progressive temporal aggregation module, which divides the temporal dimension of $Z_{temporal}$ into multiple subsegments. By integrating attention mechanism, progressive temporal aggregation computes the importance score for each timestamp within each sub-segment and derives the global representation of each sub-segment through weighted aggregation of these scores. Subsequently, progressive temporal aggregation calculates the importance score of each subsegment and progressively refines the global representation (Z_{st}) of the entire sequence, therefore effectively captures temporal dependencies from local to global levels.

2.4 Fine-Grained Consistency Optimization - FGCO

The spatially augmented view obtained through SCA is denoted as X_S, while the temporally augmented view obtained through TCA is denoted as X_T. By inputting X, X_S, and X_T into PSTA, the spatiotemporal features of each view are extracted. FGCO is designed to maximize the information correlation between the original view and its spatially/temporally augmented views by utilizing the cross-correlation matrix across different views. Specifically, this process is described below.

$$\begin{aligned} L_{total} &= L_S + L_T, \\ L_S &= \sum_s \left(1 - C'_{ss}\right)^2 + \alpha \sum_s \sum_i C'^2_{si}, \\ L_T &= \sum_t \left(1 - C'_{tt}\right)^2 + \alpha \sum_t \sum_j C'^2_{tj} \end{aligned} \tag{3}$$

This loss function encourages the augmented views to be closer to the original views while simultaneously minimizing redundancy within each view, thereby obtaining more informative and discriminative representations. The parameter α is utilized to balance the trade-off between consistency and redundancy. Through this optimization, a highly generalizable brain representation encoder named PSTA is obtained. Additionally, a linear layer is incorporated to fine-tune the entire network when applying it to downstream decoding tasks.

Table 1. Comparative Results (mean (std)). SM and SSM indicates graph supervised methods and self-supervised methods, respectively.

Dataset	Type	Method	Accuracy	AUC	Recall
Rest	SM	BrainGNN	66.60(3.63)	65.73(3.65)	**80.17(6.60)**
		STGCN	79.18(3.31)	78.55(2.68)	80.12(4.65)
	SSM	BrainGSL	68.10(1.19)	69.02(2.34)	67.56(4.83)
		GATE	65.93(2.26)	65.67(2.32)	68.63(2.25)
		STPTF	**83.37(1.77)**	**79.81(1.62)**	77.81(3.75)
Task	SM	BrainGNN	69.23(2.10)	82.22(0.99)	69.57(1.65)
		STGCN	68.59(3.63)	73.82(3.08)	55.14(5.28)
	SSM	BrainGSL	68.10(1.19)	69.02(2.34)	67.56(4.83)
		GATE	66.09(2.18)	65.82(2.20)	68.97(2.35)
		STPTF	**89.62(0.29)**	**90.62(0.35)**	**83.02(0.75)**

3 Experiments

3.1 Data Description and Implementation Details

The experimental data was from the publicly available fMRI dataset, Human Connectome Project (HCP). The resting state data from 1091 subjects and the task state data from 1007 subjects who participated fully in seven cognitive tasks are utilized to evaluate the effectiveness of STPTF. Despite originating from the same source, they represent distinct paradigms and cognitive states, providing a diverse evaluation. Specifically, the resting-state data examine gender differences in brain states, while the task-state data investigate variations in brain states across different cognitive tasks. To address computational complexity and information fitting challenges, we segment the entire brain into 22 major regions. These regions are derived by merging finer subdivisions from the existing brain atlas as [3].

We utilize a five-fold cross-validation to ensure the reliability of the decoding performance. Classification accuracy, area under the curve (AUC), and recall were used to evaluate decoding performance. In addition, a warm-up stage is incorporated into the training process to enhance stability. The learning rate

begins at a small initial value, gradually increases to its maximum as training progresses, and subsequently follows a cosine decay schedule, leading to a gradual decrease along a cosine curve. Compared with step decay, cosine decay offers a smoother adjustment, which can accelerate convergence and improve decoding performance.

Table 2. Ablation Results (mean (std)).

Dataset	Type	Method	Accuracy	AUC	Recall
Rest	STIGA	RSA + RTA	82.51(1.31)	79.50(1.29)	76.91(4.42)
		w/o/TCA	81.18(1.41)	78.29(1.14)	68.41(4.44)
		w/o/SCA	81.61(1.75)	79.17(2.02)	76.31(2.75)
	PSTA	GCN	58.62(2.02)	56.37(2.38)	39.46(4.90)
		LSTM	77.73(1.69)	75.61(0.15)	75.03(4.78)
		STGCN	77.14(3.81)	75.06(4.99)	73.91(5.54)
		STPTF	**83.37(1.77)**	**79.81(1.62)**	**77.81(3.75)**
Task	STIGA	RSA + RTA	89.15(0.44)	90.18(0.24)	82.39(0.45)
		w/o / TAM	89.17(3.57)	90.15(0.22)	82.29(0.32)
		w/o / SCM	89.25(0.50)	90.15(0.79)	82.85(0.60)
	PSTA	GCN	61.63(0.60)	72.73(0.34)	52.38(0.65)
		LSTM	88.93(0.28)	89.38(0.45)	81.34(0.28)
		STGCN	80.09(0.83)	80.22(0.69)	64.93(1.92)
		STPTF	**89.62(0.29)**	**90.62(0.35)**	**83.02(0.75)**

3.2 Comparative Results

To assess the advancement of the proposed model, we conduct a comparative analysis against several state-of-the-art graph supervised methods: including BrainGNN [6] and STGCN [3], which share the same goal as our work. They specialize in spatial and spatiotemporal brain representation mining, respectively; graph self-supervised methods based on brain spatial pattern, namely BrainGSL [14] and GATE [12]. BrainGSL is a representative example of the generative-based approach to the diagnosis of brain diseases. GATE is one of the pioneering contrastive-based methods for disease diagnosis. Table 1 illustrates the decoding performance of STPTF in two datasets. It is evident that most graph self-supervised methods exhibit a substantial performance gap compared to spatiotemporal supervised methods. Nevertheless, the proposed STPTF demonstrates comparable, and in some cases superior, decoding capabilities relative to state-of-the-art supervised spatio-temporal decoding methods. Moreover, STPTF significantly outperforms existing graph self-supervised decoding

approaches, it may stem from the neglect of temporal dynamics and the limitations inherent in the backbone (GNNs), which restrict the effectiveness of existing graph self-supervised methods.

Table 3. The robustness results under different sample sizes (mean (std), %).

	Rest			Task		
%	Accuracy	AUC	Recall	Accuracy	AUC	Recall
20	78.45(0.84)	74.42(1.79)	71.71(6.24)	86.81(0.16)	88.18(0.32)	78.92(0.55)
40	81.41(1.04)	77.81(0.57)	69.95(2.08)	88.90(0.10)	90.48(0.09)	82.92(0.15)
60	81.27(0.37)	78.51(1.03)	72.70(2.07)	88.43(0.11)	90.12(0.14)	82.28(0.24)
80	83.37(1.77)	79.81(1.62)	77.81(3.75)	89.62(0.29)	90.62(0.35)	83.02(0.75)

3.3 Ablation Results

To further validate the effectiveness of our STPTF, we perform comprehensive ablation experiments on the proposed modules, including STIGA and PSTA. For STIGA, we replace the proposed spatial and temporal augmented strategies (SCA and TCA) with random spatial and temporal augmention (RSA and RTA). For PSTA, we replace it with only spatial encoders and traditional spatiotemporal encoders. Quantitative results (Table 2) show that STIGA and PSTA have the ability to effectively guide the model to focus on key information, map features to a more general representation space, and thus enhance the performance of downstream decoding tasks. It should be noted that, although the improvements brought about by STIGA may seem modest, they demonstrate consistency across various metrics and datasets, highlighting the robustness of STIGA. Furthermore, even minor improvements can bring it closer to practical application scenarios.

3.4 Robustness and Transferability

Robustness. Considering that the ultimate goal of STPTF is to facilitate its application to different downstream decoding tasks and achieve efficient decoding even with a small amount of labeled fMRI data, we perform multiple experiments to fine-tune the sample size. The results show that even with a small sample size (20%), STPTF achieves better results than previous GSSL-based brain decoding methods and is comparable to supervised methods (Table 3).

Transferability. Current GSSL decoding algorithms are intra-domain, meaning that pre-training and fine-tuning are based on the same dataset. To validate

that the proposed STPTF has strong transferability and adapts well to different downstream decoding tasks, we further explore its performance in inter-domain scenarios. For example, we pre-train on task-related data and then fine-tune for resting-state data decoding, and vice versa. The results (Table 4) show that STPTF still achieves good performance when transferring between different datasets, similar to intra-domain results. This indicates that STPTF can precisely extract generalizable knowledge from pre-trained datasets.

Table 4. The transferability results on different fMRI sequence length, Task → Rest represents training on task-related data while fine-tuning on resting data, and vice versa (mean (std)).

Task→Rest			Rest→Task		
Accuracy	AUC	Recall	Accuracy	AUC	Recall
83.37(1.86)	80.01(1.45)	77.72(4.88)	88.18(0.49)	88.42(0.40)	79.40(0.63)

4 Conclusion

In this study, we introduce an innovative fine-grained spatio-temporal pre-training and decoding framework (STPTF) based on GSSL. This framework offers a novel technical approach to gain deeper insights into the mechanisms of brain state transitions. Specifically, STPTF can learn universal representations of brain states from unlabeled data by leveraging STIGA strategy, PSTA miner, and multi-view FGCO. Experimental results in multiple public datasets demonstrate that STPTF achieves significant improvements in unsupervised brain decoding, likely due to its incorporation of self-supervised temporal information. Furthermore, STPTF's effectiveness in downstream decoding tasks with limited data and across different datasets highlights its robustness and transferability. With broader data-sharing initiatives, STPTF has the potential to be applied to larger and more diverse decoding datasets, further validating its generalization capabilities.

Acknowledgments. This work was supported by the China Postdoctoral Science Foundation (Grant No. 2024M764143), the National Science Fund for Distinguished Young Scholars of China (Grant No. 62325601), the MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China (Grant No. 2421001), and the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62406049).

Disclosure of Interests. The authors have no competing interests in declaring that they are relevant to the content of this paper.

References

1. Borgatti, S.P., Everett, M.G.: A graph-theoretic perspective on centrality. Soc. Netw. **28**(4), 466–484 (2006)
2. Cocchi, L., et al.: Neural decoding of visual stimuli varies with fluctuations in global network efficiency. Hum. Brain Mapp. **38**(6), 3069–3080 (2017)
3. Gadgil, S., et al.: Spatio-temporal graph convolution for resting-state fMRI analysis. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12267, pp. 528–538. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59728-3_52
4. Greene, A.S., Horien, C., Barson, D., Scheinost, D., Constable, R.T.: Why is everyone talking about brain state? Trends Neurosci. **46**(7), 508–524 (2023)
5. Van den Heuvel, M.P., Sporns, O.: Network hubs in the human brain. Trends Cogn. Sci. **17**(12), 683–696 (2013)
6. Li, X., et al.: BrainGNN: interpretable brain graph neural network for FMRI analysis. Med. Image Anal. **74**, 102233 (2021)
7. Li, Z., Li, Q., Zhu, Z., Hu, Z., Wu, X.: Multi-scale spatio-temporal fusion with adaptive brain topology learning for FMRI based neural decoding. IEEE J. Biomed. Health Inform. **28**(1), 262–272 (2024)
8. Lin, X., et al.: Aberrant static and dynamic functional brain network in depression based on EEG source localization. IEEE/ACM Trans. Comput. Biol. Bioinf. **20**(3), 1876–1889 (2022)
9. Liu, J., et al.: Towards graph foundation models: a survey and beyond. ArXiv **abs/2310.11829** (2023), https://api.semanticscholar.org/CorpusID:264288909
10. Luppi, A.I., et al.: Consciousness-specific dynamic interactions of brain integration and functional diversity. Nat. Commun. **10**(1), 4616 (2019)
11. Lutnick, B., et al.: An integrated iterative annotation technique for easing neural network training in medical image analysis. Nat. Mach. Intell. **1**(2), 112–119 (2019)
12. Peng, L., Wang, N., Xu, J., Zhu, X., Li, X.: GATE: graph CCA for temporal self-supervised learning for label-efficient FMRI analysis. IEEE Trans. Med. Imaging **42**(2), 391–402 (2022)
13. Wang, M., Huang, J., Liu, M., Zhang, D.: Modeling dynamic characteristics of brain functional connectivity networks using resting-state functional MRI. Med. Image Anal. **71**, 102063 (2021)
14. Wen, G., et al.: Graph self-supervised learning with application to brain networks analysis. IEEE J. Biomed. Health Inform. **27**(8), 4154–4165 (2023)
15. Zhang, K., et al.: Self-supervised learning for time series analysis: taxonomy, progress, and prospects. arXiv preprint arXiv:2306.10125 (2023)
16. Zhang, Y., Tetrel, L., Thirion, B., Bellec, P.: Functional annotation of human cognitive states using deep graph convolution. Neuroimage **231**, 117847 (2021)

Speech Audio Generation from Dynamic MRI via a Knowledge Enhanced Conditional Variational Autoencoder

Yaxuan Li[1], Han Jiang[2], Yifei Ma[1], Shihua Qin[3], Jonghye Woo[4], and Fangxu Xing[4(✉)]

[1] Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong, China
[2] School of Software Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China
[3] Medical Center Boulevard, Wake Forest University School of Medicine, Winston-Salem, NC, USA
[4] Department of Radiology, Harvard Medical School, Boston 08544, USA
fxing1@mgh.harvard.edu

Abstract. Dynamic Magnetic Resonance Imaging (MRI) of the vocal tract has become an increasingly adopted imaging modality for speech motor studies. Beyond image signals, systematic data loss, noise pollution, and audio file corruption can occur due to the unpredictability of the MRI acquisition environment. In such cases, generating audio from images is critical for data recovery in both clinical and research applications. However, this remains challenging due to hardware constraints, acoustic interference, and data corruption. Existing solutions, such as denoising and multi-stage synthesis methods, face limitations in audio fidelity and generalizability. To address these challenges, we propose a Knowledge Enhanced Conditional Variational Autoencoder (KE-CVAE), a novel two-step "knowledge enhancement + variational inference" framework for generating speech audio signals from cine dynamic MRI sequences. This approach introduces two key innovations: (1) integration of unlabeled MRI data for knowledge enhancement, and (2) a variational inference architecture to improve generative modeling capacity. To the best of our knowledge, this is one of the first attempts at synthesizing speech audio directly from dynamic MRI video sequences. The proposed method was trained and evaluated on an open-source dynamic vocal tract MRI dataset recorded during speech. Experimental results demonstrate its effectiveness in generating natural speech waveforms while addressing MRI-specific acoustic challenges, outperforming conventional deep learning-based synthesis approaches (https://github.com/YaxuanLi-cn/KE-CVAE).

Y. Li and H. Jiang—Equal contribution.
Y. Li—This work was done by Yaxuan Li as a research assistant at Harvard Medical School.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-032-04947-6_59.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 619–629, 2026.
https://doi.org/10.1007/978-3-032-04947-6_59

Keywords: Speech audio generation · dynamic MRI · vocal tract · knowledge enhancement · variational inference

1 Introduction

Capturing real-time deformations of the vocal tract during human speech through medical imaging is essential for various speech research applications [17,29], as accurately characterizing the functional behavior of vocal articulators has multiple clinical implications. Several imaging modalities are currently used for this task, including electromagnetic articulography (EMA), ultrasound, and magnetic resonance imaging (MRI). However, EMA requires attaching multiple sensors to the articulators, which can potentially disrupt natural speech patterns, and it can only track a limited number of points on the tongue's surface [25]. Although ultrasound is non-invasive, it has a restricted field of view and struggles to clearly visualize deeper structures such as the palate and pharyngeal walls [21]. In contrast, MRI offers a more effective solution by providing high-contrast anatomical imaging with high resolution while remaining non-invasive. Advances in MRI technology have led to the development of high-speed dynamic cine imaging, enabling rapid and real-time imaging for in vivo speech [9,16].

However, simultaneously recording the subjects' speech audio signals during image acquisition remains challenging due to several factors. (1) Hardware constraints require specialized MRI-compatible microphones that function within strong magnetic fields without interference from the MRI scanner or to the image quality [8]. (2) Acoustic interference from MRI scanners (the loud pulse sequence sounds) generates substantial and unpredictable noise, varying with scanning protocols/parameters and making it difficult to capture clean recordings of human voice [2]. (3) Data integrity issues arise as the process of dynamic MRI reconstruction can affect sound recordings, with synchronization difficulties, MRI-induced noise, missing data or artifacts leading to corrupted or incomplete audio segments [23]. Despite these challenges, capturing speech waveforms remains a critical requirement in many speech imaging studies [18,30].

Multiple solutions have been proposed to obtain high-quality speech sound files in MRI environments. Denoising techniques enhance audio clarity but can only reduce noise rather than fully eliminate it and cannot reconstruct missing speech segments [7,10]. More recently, an innovative approach has emerged that synthesizes speech directly from MRI data. Liu et al. [18] developed a system that converts 4D tagged-MRI data into audio using Non-negative Matrix Factorization (NMF). While effective for processing short phrases, this method relies on a precomputed sequence of deformation fields to synthesize audio, limiting its ability to directly process dynamic cine images. Additionally, its multi-step process is computationally intensive and may not generalize well to longer or more varied speech patterns. Although these methods represent significant progress, they still fall short of producing fully accurate and complete speech sound files for comprehensive research applications.

In non-medical imaging settings, several methods have been proposed for audio generation [12,15] that utilize an end-to-end variational inference frame-

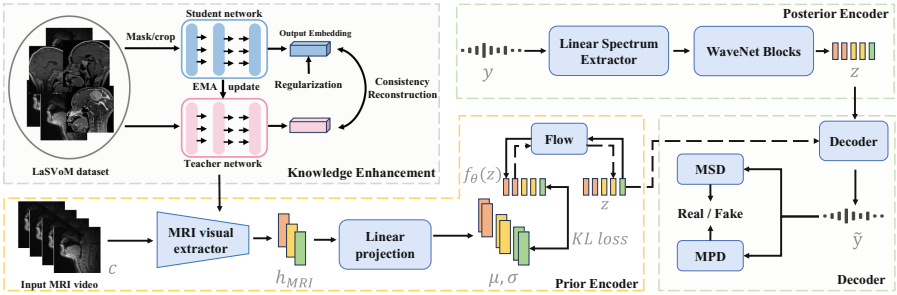


Fig. 1. Illustration of our KE-CVAE model. The dashed lines in the figure represent the inference process.

work, producing more natural-sounding signals than earlier two-stage models. Inspired by these approaches, we propose a Knowledge Enhanced Conditional Variational Autoencoder (KE-CVAE) with two key innovations: (1) A novel two-step "knowledge enhancement + variational inference" framework for speech audio generation from dynamic MRI sequences. Specifically, we have compiled a collection of thirteen public head and neck MRI image/video datasets for the training process using a self-supervised knowledge enhancement strategy. (2) Adoption of a variational inference framework to enhance the expressive power of generative modeling (Fig. 1). We conducted comprehensive experiments to generate speech sound using the open-access speech dynamic MRI dataset described in [16], demonstrating the effectiveness of KE-CVAE over conventional CNN and transformers in multiple metrics such as correlation, PESQ, and a subjective MOS score (detailed in the results section).

2 Methods

2.1 Knowledge Enhancement with Domain-Specific Data

Recent advances in self-supervised pre-training on images [4,24] in computer vision have demonstrated that it is possible to learn robust and meaningful low-dimensional features without explicitly labeled supervision. Inspired by this concept, we propose a self-supervised knowledge enhancement strategy to learn latent variables from large-scale MRI data. Specifically, our model comprises a teacher network and a student network (Fig. 1), both based on vision transformers (ViTs) [19], denoted as $F_{teacher}$ and $F_{student}$, respectively. By leveraging a large-scale curated set of vocal tract MRI, our strategy enables the visual extractor to focus on articulatory-relevant regions and capture subtle distinctions in vocal tract configurations during speech production. To effectively learn domain-specific knowledge without annotations, we introduce three complementary loss functions: consistency loss for representation alignment, reconstruction loss for masked image modeling, and KoLeo regularization for feature distribution optimization.

Consistency Loss. The consistency loss is designed to maximize the alignment of the output classification ($[CLS]$) token embeddings between $F_{teacher}$ and $F_{student}$, facilitating better representation learning. Specifically, we pass the $[CLS]$ token through a standard multilayer perceptron (MLP) model to generate a vector of scores, followed by a softmax function to obtain the pseudo-class probability p. We compute consistency loss at both the global and local levels, corresponding to image-level and patch-level augmentations, respectively. The global consistency loss is defined as the cross-entropy loss between the teacher output p_t and the student output p_s. The local consistency loss is computed as the cross-entropy loss between the $[CLS]$ embeddings of both networks, p_{sl} and p_{tl}. We apply Sinkhorn-Knopp centering [3] to the teacher network output to enhance distribution alignment. The final consistency loss integrates both global and local components, formulated as follows:

$$\mathcal{L}_{con} = -\sum_{i=1}^{N} p_t^{(i)} \log p_s^{(i)} - \sum_{i=1}^{N}\sum_{j=1}^{M} p_{tl}^{(i,j)} \log p_{sl}^{(i,j)}, \tag{1}$$

where N is the batch size and M is the number of patches per sample.

Reconstruction Loss. We randomly replace the patches in the input image with masks using the $[MASK]$ token or random patch features with a certain probability. The masked input MRI image $\mathbf{V}_i$ can be represented as $\mathbf{V}_{i_mask}$ where the embeddings of some patches $\{v_i\}_{i\in B}$ in $\mathbf{V}_{i_mask}$ are replaced by trainable $[MASK]$ token embeddings. The reconstruction loss is computed by comparing the output embeddings of the $[MASK]$ tokens from both the teacher and student models:

$$\mathcal{L}_{rec} = -\sum_{i=1}^{N}\sum_{j=1}^{M} h_j \cdot F_{teacher}(\mathbf{V}_i) \log F_{student}(\mathbf{V}_{i_mask}), \tag{2}$$

where $h_j = 1$ indicates that the token at position j has been masked, and $h_j = 0$ indicates otherwise. Incorporating this reconstruction loss complements the consistency loss, facilitating the learning of robust MRI image embeddings at multiple levels.

KoLeo Regularization Loss. The KoLeo regularization loss [6] has been designed to encourage a uniform span of the features within a batch. Given a batch of N output $[CLS]$ token embeddings from $\{F_{student}(\mathbf{V}_i)[CLS]\}_{i=1}^{N}$ of the student network, KoLeo regularization loss is defined as:

$$\mathcal{L}_{KoLeo} = -\frac{1}{N}\sum_{i=1}^{N} \log(d_{N,i}), \tag{3}$$

where $d_{N,i} = \min_{j\neq i} \|F_{student}(\mathbf{V}_i)[CLS] - F_{student}(\mathbf{V}_j)[CLS]\|$ is the minimal distance between $F_{student}(\mathbf{V}_i)[CLS]$ and any other student network $[CLS]$ embedding within the batch. $F_{student}(\mathbf{V}_i)[CLS]$ is also ℓ_2-normalized before computing the KoLeo regularization loss.

2.2 Variational Inference Framework on Dynamic Speech MRI

We formulate the proposed model KE-CVAE as a conditional variational autoencoder. In this framework, $p(y|z)$ represents the likelihood function for generating an audio waveform data point y given the latent variable z, while $q(z|y)$ denotes the approximate posterior distribution that infers z from y. Additionally, $p(z|c)$ describes the prior distribution of the latent variables z, conditioned on the MRI sequence c. The objective of the CVAE is to maximize the variational lower bound, also known as the evidence lower bound (ELBO), of the intractable marginal log-likelihood of the data $\log p(y|c)$. This objective can be decomposed into two terms: the reconstruction loss and the KL divergence, expressed as follows:

$$\mathcal{L}_{ELBO} = \mathcal{L}_{recon} + KL(q(z|y) \,||\, p(z|c)). \tag{4}$$

The reconstruction loss, $\mathcal{L}_{recon}$, is defined as $\mathcal{L}_{recon} = \|y_{mel} - \tilde{y}_{mel}\|_1$, where y_{mel} denotes the mel-spectrogram of the input audio, and $\tilde{y}_{mel}$ represents the mel-spectrogram reconstructed by the decoder. The $\|\cdot\|_1$ notation refers to the ℓ_1 norm.

The KE-CVAE model primarily comprises three components: a posterior encoder, a prior encoder, and a decoder. We detail the three modules as follows:

Posterior Encoder. The posterior encoder extracts the latent representation $z \sim q(z|y)$ from the input waveform y. We first transform the raw waveform into its linear spectrum. Several non-causal WaveNet residual blocks [11] are applied to extract an embedding sequence. Then we employ a linear layer to project the mean and variance from the normal posterior distribution $p(z|y)$.

Prior Encoder. In our proposed setting, dynamic MRI sequences serve as the condition for speech generation. The prior encoder models the conditional prior distribution $p(z|c)$ given the MRI sequence c. As described in Sect. 2.1, the pretrained MRI visual extractor $F_{teacher}$ is applied in this module to obtain the hidden representation h_{MRI}. The linear projection layer following the blocks produces the mean μ and variance σ of the normal posterior distribution. To further improve the scalability of the approximated posterior distributions, we use a normalizing flow f_θ to apply a sequence of invertible transformations [28]. Therefore, the KL divergence is calculated by:

$$\begin{aligned} KL(q(z|y) \,||\, p(z|c)) &= \log q(z|y) - \log p(z|c), \\ p(z|c) &= \mathcal{N}(f_\theta(z); \mu(c), \sigma(c)) \left| \det \frac{\partial f_\theta(z)}{\partial z} \right|. \end{aligned} \tag{5}$$

where det denotes the determinants.

Decoder. The decoder uses the latent variable z to reconstruct the waveform $\tilde{y} \sim p(y|z)$. In addition, we adopt the adversarial training strategy. The discriminator D follows HiFi-GAN's multi-period discriminator (MPD) and multi-scale

discriminator (MSD) architecture [13]. Specifically, the adversarial losses [22] for the generator G and the discriminator D are defined as:

$$\mathcal{L}_{adv}(D) = \mathbb{E}_{(y,z)}\left[(D(y)-1)^2 + (D(G(z)))^2\right], \tag{6}$$

$$\mathcal{L}_{adv}(G) = \mathbb{E}_{z}\left[(D(G(z))-1)^2\right]. \tag{7}$$

where $\mathbb{E}$ denotes the expectation. Specifically, we utilize the feature matching loss [14] as an element-wise reconstruction loss for more stable training procedure.

3 Experiments and Results

3.1 Datasets and Evaluation Metrics

LaSVoM Dataset. To address the scarcity of large-scale head and neck MRI resources, we constructed a "large-scale vocal track MRI" data collection (LaSVoM) by aggregating and curating data from thirteen public MRI image and video datasets in the head and neck region. Through systematic frame extraction and resizing, we collected over 30,000 high-quality mid-sagittal head and neck MRI images, each resized to focus on the vocal tract region with a size of 84 × 84 pixels. These images encompass a wide range of phonetic contexts, speakers, and articulation patterns, providing a comprehensive resource for the proposed knowledge enhancement process.

Variational Inference Dataset. We utilized the open dynamic MRI in speech dataset in [16], which consists of cine MRI capturing dynamic vocal tract movements in the sagittal plane, accompanied by synchronized audio recordings. The dataset includes cine sequences from 75 participants performing various speech tasks. For training and testing, we split the dataset into an 8:2 ratio.

Evaluation Metrics. We perform all experiments in a speaker-independent manner so that the metrics we used are not affected by the change of speakers. We calculated conventional objective metrics for audio waveform evaluation, including 2D Pearson's correlation coefficient (Corr2D) [5] and perceptual evaluation of speech quality (PESQ) [27]. We also performed a subjective MOS (mean opinion score) test to evaluate all compared methods. Specifically, we randomly selected 30 from 1662 test audios for subjective listening and asked 10 human raters to assess their quality. The raters evaluated the audio quality by comparing each reconstructed audio sample with its corresponding ground truth, rating them on a scale from 1 to 5 based on their similarity and overall quality.

3.2 Implementation and Training Protocols

Due to significant imaging differences across scanner platforms, we normalized each image in the LaSVoM dataset to a mean of 0.5 and a standard deviation of 0.5. For both the training and test datasets, we applied a noise reduction

Table 1. Quantitative evaluation results of the audio quality using different synthesis methods on various metrics

Method	#Param	Corr2D ↑	PESQ ↑	MOS ↑
Vanilla (CNN) *w/* KE	124.0 M	0.672	1.135	3.48 (±0.14)
Vanilla (CNN)	124.0 M	0.641	1.116	3.16 (±0.09)
Vanilla (Transformer) *w/* KE	146.5 M	0.688	1.140	3.57 (±0.15)
Vanilla (Transformer)	146.5 M	0.649	1.131	3.22 (±0.12)
Ours	121.1 M	**0.818**	**1.251**	**4.13 (±0.08)**
Ours *w/o* adversarial training	60.0 M	0.759	1.190	3.80 (±0.13)
Ours *w/o* Flow	120.6 M	0.798	1.237	4.02 (±0.10)

Table 2. Ablation study in our knowledge enhancement strategy

Method	Corr2D ↑	PESQ ↑	MOS ↑
Baseline	0.716	1.152	3.74 (± 0.12)
$+\mathcal{L}_{con}$	0.758	1.170	3.88 (± 0.14)
$+\mathcal{L}_{con} + \mathcal{L}_{rec}$	0.802	1.226	4.06 (± 0.10)
$+\mathcal{L}_{con} + \mathcal{L}_{rec} + \mathcal{L}_{Koleo}$	0.818	1.251	4.13 (± 0.08)

algorithm[1] to remove background noise from the audio, followed by audio magnitude normalization. Since the dataset contained a substantial number of silent segments (audio volume below –60 dB for more than 0.2 s), we used FFmpeg[2] to detect and segment audio sequences, retaining only the voiced segments. Finally, we resampled the audio to 16,000 Hz and the video to 80 fps.

In the proposed model, The MRI visual extractor was composed of 12 ViTs blocks. The dimension of the WaveNet was 512. The decoder and discriminator version of the HiFi-GAN we used was V1. The networks were trained using the AdamW optimizer [20] with $\beta_1 = 0.8$, $\beta_2 = 0.99$ and weight decay $\lambda = 0.01$. We implemented our approach using the Pytorch toolbox and trained for a total of 48 h on an NVIDIA Tesla A100 GPU, where 40 epochs were used for the knowledge enhancement stage and 100 epochs were used for the variational inference stage.

3.3 Benchmarking

Quantitative Analysis. To validate the effectiveness of the proposed KE-CVAE framework, we compared it against two vanilla network variants following [1,26]. Specifically, we replaced the variational inference step with conventional CNN and transformer architectures with a similar number of parameters for the

[1] https://github.com/timsainb/noisereduce.
[2] https://github.com/FFmpeg/FFmpeg.

same speech audio generation task. In these vanilla variants, we directly used the MRI sequences as input and passed them through a series of encoders to obtain the audio output. Both vanilla models were optimized using the reconstruction loss $\mathcal{L}_{recon}$ as defined in our method. From Table 1, it is prominent that the performances of both architectures are evidently lower than that of KE-CVAE. The p-value obtained from the one-tailed Student's t-test is less than 0.05, indicating statistical significance. Notably, our knowledge enhancement step (KE in Table 1) also proves to be beneficial for the two vanilla methods, demonstrating its general applicability and effectiveness.

We further conducted an ablation study to assess the contributions of different components in the KE-CVAE framework. Results in Table 1 indicate that removing adversarial training and Flow mechanisms leads to a noticeable performance drop, highlighting the critical role of either component in achieving optimal performance. Meanwhile, every component in Sect. 2.1 plays an important role in the whole framework. To evaluate the impact of each objective function, we performed additional ablation studies and presented the results in Table 2, where we used the CVAE model without knowledge enhancement as baseline. Results show that all the proposed loss functions in Sect. 2.1 are essential in improving the final performance.

Qualitative Analysis. Moreover, we visualized the spectrograms of the generated audio clips from the test stage, with one example shown in Fig. 2. As can be seen, the generated spectrogram and its corresponding audio waveform from KE-CVAE exhibit superior alignment with the ground truth. In contrast, the CVAE framework without the knowledge enhancement step (*w/o* KE in Fig. 2) and the vanilla transformer model's quality are visually lower than that of the proposed full KE-CVAE.

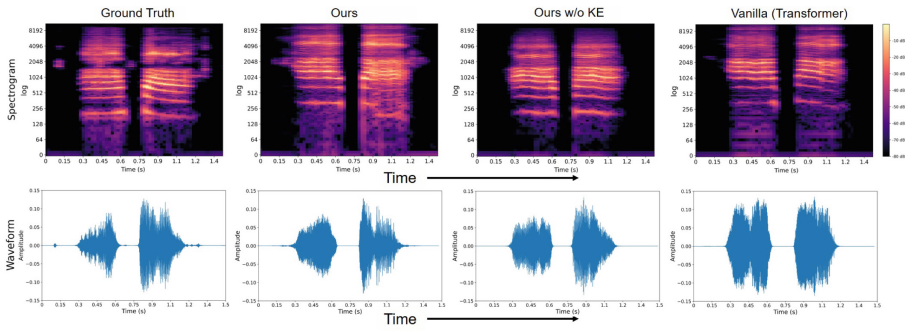


Fig. 2. An example of generated audio spectrograms and waveforms using different methods compared to the ground truth.

4 Conclusion

In this work, we developed a novel two-step "knowledge enhancement + variational inference" framework to synthesize high-quality speech waveforms from dynamic MRI sequences, addressing critical challenges in synchronizing speech audio with real-time MRI recordings. In the knowledge enhancement phase, we proposed a robust self-supervised training pipeline that leverages large-scale MRI data to learn domain-specific features without explicit annotations. Following this, we integrated a variational inference framework to enhance the model's expressiveness. Specifically, we employed a posterior encoder, a prior encoder, and a decoder to effectively map latent variables to speech waveforms, further augmenting them with normalizing flow and adversarial training to improve the overall training process. Experimental results have demonstrated the efficacy of KE-CVAE in generating high-quality, temporally accurate speech from dynamic MRI data. This method has the potential to significantly enhance the accuracy of speech analysis in both clinical and research settings.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akbari, H., Arora, H., Cao, L., Mesgarani, N.: Lip2audspec: speech reconstruction from silent lip movements video. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2516–2520. IEEE (2018)
2. Bresch, E., Nielsen, J., Nayak, K., Narayanan, S.: Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. J. Acoust. Soc. Am. **120**(4), 1791–1794 (2006)
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural Inf. Process. Syst. **33**, 9912–9924 (2020)
4. Chen, X., He, K.: Exploring simple Siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15750–15758 (2021)
5. Chi, T., Ru, P., Shamma, S.A.: Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. **118**(2), 887–906 (2005)
6. Delattre, S., Fournier, N.: On the kozachenko-leonenko entropy estimator. J. Stat. Plan. Inference **185**, 69–93 (2017)
7. Ertürk, M.A., Bottomley, P.A., El-Sharkawy, A.M.M.: Denoising MRI using spectral subtraction. IEEE Trans. Biomed. Eng. **60**(6), 1556–1562 (2013)
8. Jacobs, S.M., et al.: Image quality and subject experience of quiet t1-weighted 7-t brain imaging using a silent gradient coil. Eur. Radiol. Exp. **6**(1), 36 (2022)
9. Jin, R., et al.: Optimization of 3D dynamic speech mri: poisson-disc undersampling and locally higher-rank reconstruction through partial separability model with regional optimized temporal basis. Magn. Reson. Med. **91**(1), 61–74 (2024)
10. Jin, R., et al.: Enhancing linguistic research through 2-mm isotropic 3d dynamic speech mri optimized by sparse temporal sampling and low-rank reconstruction. Magn. Reson. Med. **89**(2), 652–664 (2023)

11. Kim, J., Kim, S., Kong, J., Yoon, S.: Glow-tts: a generative flow for text-to-speech via monotonic alignment search. Adv. Neural Inf. Process. Syst. **33**, 8067–8077 (2020)
12. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning, pp. 5530–5540. PMLR (2021)
13. Kong, J., Kim, J., Bae, J.: Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis. Adv. Neural Inf. Process. Syst. **33**, 17022–17033 (2020)
14. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: International Conference on Machine Learning, pp. 1558–1566. PMLR (2016)
15. Lee, S.H., Kim, S.B., Lee, J.H., Song, E., Hwang, M.J., Lee, S.W.: Hierspeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis. Adv. Neural Inf. Process. Syst. **35**, 16624–16636 (2022)
16. Lim, Y., et al.: A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images. Scientific Data **8**(1), 187 (2021)
17. Lim, Y., Zhu, Y., Lingala, S.G., Byrd, D., Narayanan, S., Nayak, K.S.: 3D dynamic mri of the vocal tract during natural speech. Magn. Reson. Med. **81**(3), 1511–1520 (2019)
18. Liu, X., et al.: Speech Audio Synthesis from Tagged MRI and Non-negative Matrix Factorization via Plastic Transformer. In: Greenspan, H., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. MICCAI 2023. LNCS, vol. 14226, pp. 435–445. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43990-2_41
19. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Lulich, S.M., Berkson, K.H., de Jong, K.: Acquiring and visualizing 3d/4d ultrasound recordings of tongue motion. J. Phon. **71**, 410–424 (2018)
22. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802 (2017)
23. Menchón-Lara, R.M., Simmross-Wattenberg, F., Casaseca-de-la Higuera, P., Martín-Fernández, M., Alberola-López, C.: Reconstruction techniques for cardiac cine mri. Insights Imaging **10**, 1–16 (2019)
24. Oquab, M., et al.: Dinov2: learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
25. Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T.: Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. J. Acoust. Soc. Am. **92**(6), 3078–3096 (1992)
26. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: Learning individual speaking styles for accurate lip to speech synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13796–13805 (2020)
27. Recommendation, I.T.: Perceptual evaluation of speech quality (pesq): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Rec. ITU-T P. 862 (2001)

28. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International Conference on Machine Learning, pp. 1530–1538. PMLR (2015)
29. Toutios, A., et al.: Illustrating the production of the international phonetic alphabet sounds using fast real-time magnetic resonance imaging. In: Interspeech, pp. 2428–2432 (2016)
30. Zheng, R.C., Ai, Y., Ling, Z.H.: Incorporating ultrasound tongue images for audiovisual speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. (2024)



Steerable Anatomical Shape Synthesis with Implicit Neural Representations

Bram de Wilde[1,3(✉)], Max T. Rietberg[2], Guillaume Lajoinie[1], and Jelmer M. Wolterink[3]

[1] Physics of Fluids Group, Technical Medical (TechMed) Centre, University of Twente, Enschede, The Netherlands
[2] Multi-Modality Medical Imaging, Technical Medical (TechMed) Centre, University of Twente, Enschede, The Netherlands
[3] Department of Applied Mathematics, Technical Medical (TechMed) Centre, University of Twente, Enschede, The Netherlands
contact@bramdewilde.com

Abstract. Generative modeling of anatomical structures plays a crucial role in virtual imaging trials, which allow researchers to perform studies without the costs and constraints inherent to in vivo and phantom studies. For clinical relevance, generative models should allow targeted control to simulate specific patient populations rather than relying on purely random sampling. In this work, we propose a steerable generative model based on implicit neural representations. Implicit neural representations naturally support topology changes, making them well-suited for anatomical structures with varying topology, such as the thyroid. Our model learns a disentangled latent representation, enabling fine-grained control over shape variations. Evaluation includes reconstruction accuracy and anatomical plausibility. Our results demonstrate that the proposed model achieves high-quality shape generation while enabling targeted anatomical modifications.

Keywords: Shape Synthesis · Implicit Neural Representations · Latent Space Disentanglement

1 Introduction

Evaluating the performance of novel medical imaging modalities through patient trials is expensive in terms of patient risk, time, and cost [1]. Simulating *in vivo* trials with tissue-mimicking phantoms is common practice, but since recreating human tissue is complex, experiments with high mimicking accuracy remain costly [2]. Virtual imaging trials, where both the patient and the imaging modality are simulated, provide a compelling alternative: they allow for inexpensive, rapid and flexible testing while closely matching patient tissue [3]. However, these

B. de Wilde and M. T. Rietberg—These authors contributed equally to this work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 630–639, 2026.
https://doi.org/10.1007/978-3-032-04947-6_60

trials are only meaningful if the underlying anatomical model of the patients is accurate, but also covers the variations seen in the population.

Generative models have the potential to synthesize anatomical structures [4], enhancing existing (open source) anatomical datasets, such as the Gold Atlas project [5] and MedShapeNet [6]. However, in medical imaging, factors not directly related to the clinical question can strongly influence the image and its interpretation. For example, imaging might be affected by anatomical variation and tissue composition, which purely random generative approaches cannot disentangle. There is therefore a need for models that provide control over such factors, allowing virtual cohorts to be generated in line with study objectives.

In recent years, implicit neural representations (INRs) have emerged as a powerful and flexible platform for shape modeling [7,8]. To describe shapes, INRs can represent surfaces as the zero-level set of their signed distance function (SDF), modeled in a neural network. Unlike more traditional template-based approaches, INRs are amenable to various conditioning mechanisms and naturally allow for modeling topological changes in a population [9,10]. Conditioning INRs on relevant shape features introduces an additional level of control for shape synthesis, which has previously been demonstrated outside [11,12], and inside the medical domain [13–15]. Furthermore, the ability of INRs to represent organic shapes with varying or changing topology has been previously demonstrated in biomedical applications [16]. Topological changes in anatomy can, for instance, occur due to certain diseases like tissue adhesion, renal fusion, or osteosarcoma, or due to surgical procedures like a laryngectomy.

In this paper, we investigate steerable anatomical shape synthesis in a population with topological variations. As a concrete use case we opt to model the thyroid gland, which is diverse in its shape and bilateral symmetry, and not topologically consistent across the patient population. The thyroid consists of two lobes, connected by a bridge called the *isthmus*. However, up to 33% of all patients show agenesis of the isthmus, having two separate lobes instead of a single connected thyroid [17]. We show that INRs are capable of synthesizing anatomically feasible thyroids. Additionally, we condition INRs on three key anatomical features (volume, isthmus cross-sectional area and symmetry) to generate and edit thyroids in a steerable way. Finally, we experiment with a simple correlation loss term to promote feature disentanglement.

2 Methods

2.1 Model

To model three-dimensional shapes, we use coordinate-based multilayer perceptrons (MLPs) to encode the SDF. Similar to [7], we use a single MLP to represent multiple shape instances by conditioning the MLP on a latent code $\boldsymbol{z}$, which is concatenated to the input coordinates. We use MLPs with 3 hidden layers of 256 nodes and ReLU activations. For every target shape i, we sample the SDF value $s \in \mathbb{R}$ for a set of coordinates $\boldsymbol{x} \in \mathbb{R}^3$:

$$X_i = \{(\boldsymbol{x}, s) : SDF_i(\boldsymbol{x}) = s\} \tag{1}$$

The parameters θ of the MLP f_θ are optimized such that the model approximates the SDF for each shape i when conditioned on its latent code $\boldsymbol{z}_i$:

$$f_\theta(\boldsymbol{x}, \boldsymbol{z}_i) \approx SDF_i(\boldsymbol{x}) \tag{2}$$

2.2 Training

Both the parameters θ and the latent codes $\boldsymbol{z}_i$ are optimized using the mean squared error loss on the predicted SDF values. Additionally, we apply L_2 regularization to the latent codes as we assume Gaussian noise on the SDF values, leading to the following total loss $\mathcal{L}$:

$$\mathcal{L}(f_\theta, X, \boldsymbol{z}) = \sum_{\boldsymbol{x} \in X} ||f_\theta(\boldsymbol{x}, \boldsymbol{z}) - SDF(\boldsymbol{x})||_2^2 + \lambda ||\boldsymbol{z}||_2^2 \tag{3}$$

2.3 Disentanglement

Training a model as described in Sect. 2.1 does not impose any structure on the latent space. Hence, randomly sampled codes provide novel samples, but there is no easy way to control anatomical aspects, such as the volume of the generated shape. To promote steerability of output shape characteristics, we split the latent code $\boldsymbol{z}_i$ for each shape into a fixed part $\boldsymbol{z}_{i,fixed}$, which is not updated during training, and a trainable part $\boldsymbol{z}_{i,trainable}$. By letting $\boldsymbol{z}_{i,fixed}$ represent anatomical features, the model is directly conditioned on them, such that we can investigate disentanglement of $\boldsymbol{z}_{i,fixed}$ with respect to the trainable features. If $\boldsymbol{z}_{i,fixed}$ is, for instance, set to the volume of each shape i, then $\boldsymbol{z}_{i,trainable}$ will ideally model all shape changes *but* the volume. Concretely, we investigate the two following disentanglement strategies:

Fixed Conditioning. The anatomical feature(s) of interest are added as fixed latent code features and training is done as described in Sect. 2.1. This is a baseline approach to investigate how much the model disentangles the features by itself without any special strategies.

Correlation Loss. In addition to the **Fixed conditioning** strategy, we calculate a correlation loss term at the end of each epoch to update the latent codes. The loss term promotes disentanglement of the fixed features from the trainable features. Specifically, the loss term calculates the mean Pearson correlation coefficient between the fixed and trainable latent features:

$$\mathcal{L}_{corr} = \frac{1}{N} \sum_{j}^{N} \left| \frac{\mathrm{Cov}(\boldsymbol{z}_{fixed}, \boldsymbol{z}_{trainable,j})}{\sigma_{fixed} \sigma_{trainable}} \right| \tag{4}$$

Here N is the number of trainable latent features, j is the latent feature index. If a model is perfectly disentangled, the fixed feature should have no correlation to any of the trainable dimensions, and hence have a low loss value.

2.4 Inference

Shapes are synthesized by conditioning the model on a latent code $\boldsymbol{z}$ and sampling the predicted SDF values $f_\theta(\boldsymbol{x}, \boldsymbol{z})$ for $\boldsymbol{x} \in [0, 1]^3$, where $\boldsymbol{x}$ is sampled on a 64^3 grid. Meshes are obtained with marching cubes. Novel samples are synthesized by sampling randomly from the latent space. For each trainable latent dimension, we fit and draw random values from a normal distribution. For fixed latent dimensions in conditioned models, we sample directly from the distribution seen in the dataset, such that the synthesized population follows the training population.



Fig. 1. Examples from the dataset illustrating the variety in thyroid anatomy. From left to right: (1) a typical connected thyroid, (2) a typical split thyroid, (3) a very large thyroid, (4) a very small thyroid, (5) a highly asymmetric thyroid.

2.5 Data

We collect thyroid shapes from the TotalSegmentator training dataset [18]. The dataset consists of 1228 CT scans on which 117 structures, including the thyroid gland, have been annotated. Of these 1228 scans, there are 415 scans where the thyroid gland is fully in the field of view and is completely annotated. We additionally collect the trachea shape for each thyroid to center the meshes. Due to the diverse nature of the dataset, not all cases are annotated consistently. After a check for annotation quality, which includes checking for a number of connected components and watertightness, we discard 62 cases. Finally, we train all models on a set of 353 thyroids. Representative examples from the dataset are shown in Fig. 1.

We choose to center all meshes on the trachea, instead of on thyroid center of mass, because asymmetric thyroids would give off-center results. To allow for

centering, we first convert all binary voxel grid thyroid masks to meshes with marching cubes, and then center them on the center of mass of the trachea. The meshes are finally normalized to the largest extent in the dataset in each dimension, such that all shapes fall within a common unit cube, but size variations are preserved.

For each thyroid mesh, we sample the SDF values for 50,000 coordinates for training. 40,000 coordinates are randomly sampled on the mesh surface and hence have an SDF of 0. The remaining 10,000 points are a random sample of the 40,000 surface points plus a randomly sampled displacement in each direction from a Gaussian distribution with a standard deviation of 0.1. We publish the processed meshes, pre-sampled SDF values and our source code online.[1]

2.6 Validation

To validate that the baseline model is able to capture the anatomical variations in the dataset, we evaluate reconstruction accuracy with the Chamfer distance between reconstructed and reference meshes. To validate that the baseline model can also *generate* meaningful novel samples, we validate that they are anatomically realistic with respect to three key anatomical features, of which we compare training to generated population. Specifically, the anatomical features we consider are thyroid **volume**, **isthmus area**, and **symmetry**.

Volume is a natural measure for size variations. Furthermore, isthmus area allows for a continuous description from a split thyroid (area = 0) to a connected one. It is calculated as the thyroid area in the midsagittal plane. Finally, to characterize the large variation in symmetry in the dataset, we flip one half of the thyroid onto the other by mirroring at the midsagittal plane, and calculating the intersection over union between both halves. In this measure, a perfectly symmetric thyroid has a symmetry score of 1, whereas a completely asymmetric thyroid has a score of 0.

To validate steerability of the conditioned models (*fixed* and *correlation*), we randomly generate 1000 thyroids and evaluate the correlation between the conditioned anatomical features and the actual generated features with the Pearson correlation coefficient (PCC).

3 Results

All models are trained with the Adam optimizer with a learning rate of $3 \cdot 10^{-4}$. Each model is trained for 10,000 epochs, where each epoch consists of 1000 coordinate-SDF pairs for each shape, sampled randomly from X_i. Models are trained with a trainable latent code of size $N = 64$, with their values initialized randomly from $\mathcal{N}(0, 0.01^2)$ and where applicable, features added as 3-vectors. Throughout the results we compare three different models:

– *Baseline* A model conditioned only on a trainable latent code.

[1] https://github.com/MIAGroupUT/steerable-shape-synthesis.

- *Fixed* The baseline model, but additionally conditioned on volume, isthmus area, and symmetry. This model is trained with the default loss (3).
- *Correlation* The fixed model, with the correlation loss (4) added.

3.1 Reconstruction Quality

The mean Chamfer distance between training meshes and their corresponding reconstructed meshes were 1.56 ± 0.38 mm (std) for the *baseline* model, 1.63 ± 0.12 mm for the *fixed* model, and 1.60 ± 0.06 mm for the *correlation model*. This shows that the baseline model is able to fit all shapes in the dataset well, and that adding extra conditioning to the latent code does not impact reconstruction quality, as the dimensions of the shapes in the training set are $44.75 \pm 9.79 \times 29.32 \pm 5.91 \times 52.18 \pm 7.03$ mm.

3.2 Generation Quality

To demonstrate the generative performance of the *baseline* model, 1000 meshes were randomly generated. The generated distributions of volume, isthmus area and symmetry compared to the training distributions are shown in Fig. 2. These results indicate that our baseline approach is able to generate thyroids which cover the entire training distribution. Moreover, the learned latent space follows the anatomical distribution of the training data. Furthermore, to verify that we are generating new shapes and are not simply replicating shapes in the training set, we computed the pairwise Chamfer distance between all shapes in the training set. This showed that the average Chamfer distance to the closest shape was 4.13 ± 1.33 (std). For the 1000 randomly generated shapes, the average distance to the closest shape in the training set was 3.67 ± 0.51, 3.72 ± 0.95, and 3.69 ± 0.76 for the *baseline*, *fixed* and *correlation* versions, respectively. Hence, while shapes are samples from the same distribution, synthetic shapes have no match in the training set.

3.3 Steerability

To evaluate to which extent latent conditioning on volume, isthmus area and symmetry works, we generate 1000 random meshes for both the *fixed* and the *correlation* model. Figure 3 shows the correlation between features the model is conditioned on and the actual generated features for the both models. Generated volume correlates well with the conditioned volume with a PCC of 0.98 for both models, indicating that volume is a relatively easy feature to disentangle. Isthmus area and especially symmetry appear more complex features to disentangle, and here the use of the correlation loss improves the PCC in both cases: from 0.88 to 0.93 for the isthmus area, and from 0.32 to 0.43 for symmetry.

To investigate the ability of the *correlation* model to independently vary specific features, we show the editing of a particular training mesh in Fig. 4. While varying a specific feature, the other features show almost no difference,

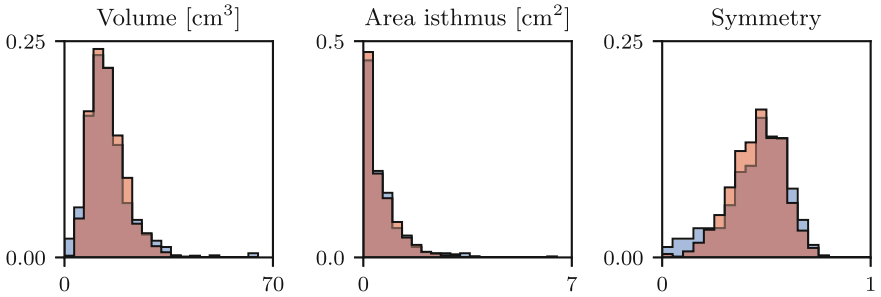


Fig. 2. Comparison between volume, isthmus area and symmetry for the training data (blue) and 1000 randomly generated meshes using the *baseline* model (red). The dark red denotes the overlap of both distribution. (Color figure online)

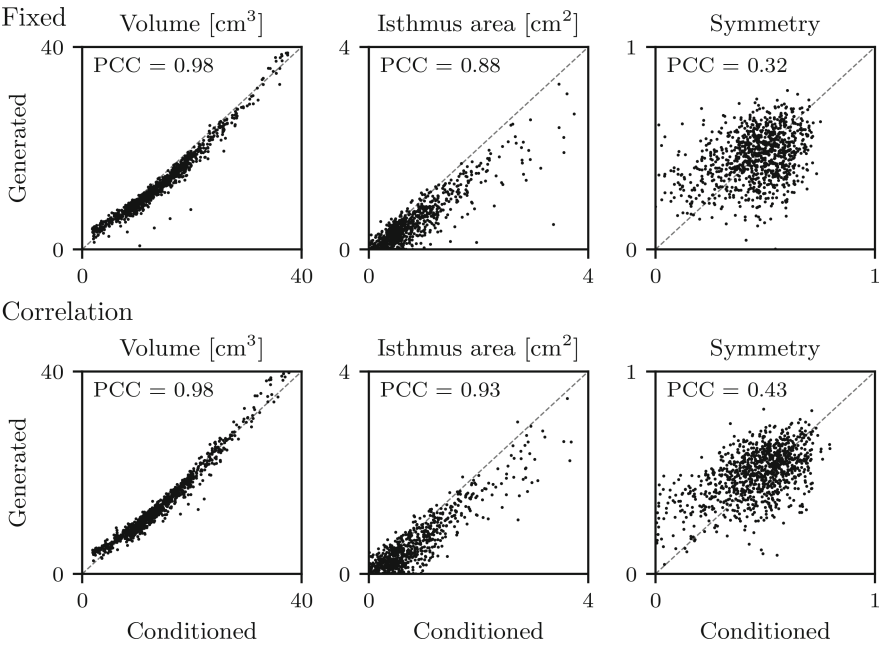


Fig. 3. Correlation between conditioned and generated anatomical features for the *fixed* model (top row) and the *correlation* model (bottom row). The inset shows the Pearson correlation coefficient.

with exception of the isthmus area when increasing the symmetry value. We hypothesize this might be because the cross-sectional area increases when you tilt a thyroid, i.e., make it less symmetrical. This could be a reason why it is harder for the model to disentangle symmetry from the isthmus area than from the volume.

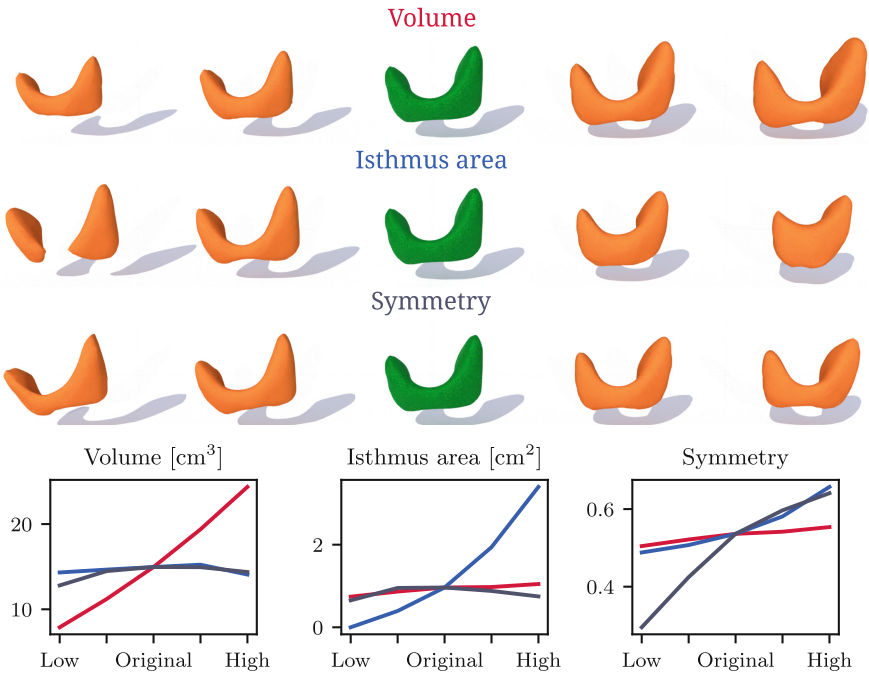


Fig. 4. Editing a training mesh (middle column, green) by independently varying volume (red), isthmus area (blue) and symmetry (grey). The plots show each anatomical feature for each of the rows, demonstrating that features can be independently varied. (Color figure online)

4 Discussion

We show that INRs are capable of synthesizing anatomically accurate thyroid glands, including the ability to model topological changes across a patient population. By conditioning the models on volume, isthmus area, and symmetry, it is possible to synthesize thyroids in a controlled manner. While prior work by Sørensen et al. [14] has shown that anatomy synthesis with INRs can be conditioned on explicit patient characteristics, we here show that explicitly minimizing the correlation between fixed and trainable features enables disentanglement of these features in the latent space. The resulting model can be used to generate patient cohorts with specific anatomical characteristics and allows editing anatomical features of existing thyroids with a high degree of independence.

Both Fig. 3 and 4 show that the three anatomical features we include are not equally complex for the model. Volume is fitted well by both models, but isthmus area is already more difficult, which might be due to fact that it is a more local feature. Symmetry is clearly the most complex, although Fig. 2 and 4 show that the models can model the feature to some extent. In order for the

model to disentangle symmetry, it arguably needs to learn some representation of that feature, which might be difficult for our relatively small MLP models.

While the thyroid is a complex use-case due to variations in topology and symmetry, the shapes themselves are still relatively smooth. Whether our approach generalizes to organs with sharper features that require a representation containing high frequencies is an interesting direction for future research. In this case, investigating alternative model architectures such as Siren, which has been shown to have a stronger capacity for representing sharp features, is probably a fruitful direction [19].

Another interesting avenue for future investigation is extending our approach to a multi-organ setting. In this setting, a model could be conditioned on a unique latent code for each output organ. The generative process can then be steered on an organ-level basis, while still providing a coherent set of generated organs, respecting the anatomical hierarchy. Using the proposed correlation loss could help models to disentangle organ representations in latent space, keeping their representation independent, and allowing for imposing variations in single organs.

Acknowledgments. All authors acknowledge the funding from KWF, TKI-Life Sciences and Health and Seno Medical Instruments in project THYNAS+. Additionally, Guillaume Lajoinie acknowledges funding from the HORIZON.1.1 program of the European Research Council for the Super-FALCON project, grant agreement ID: 101076844.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fogel, D.B.: Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp. Clin. Trials Commun. **11**, 156–164 (2018). https://doi.org/10.1016/j.conctc.2018.08.001
2. McGarry, C.K., et al.: Tissue mimicking materials for imaging and therapy phantoms: a review. Phys. Med. Biol. **65**(23), 23TR01 (2020). https://doi.org/10.1088/1361-6560/abbd17
3. Abadi, E., et al.: Virtual clinical trials in medical imaging: a review. J. Med. Imaging **7**(4), 042805 (2020). https://doi.org/10.1117/1.JMI.7.4.042805
4. Liu, Y., Dwivedi, G., Boussaid, F., Bennamoun, M.: 3D brain and heart volume generative models: a survey. ACM Comput. Surv. **56**(6), 137 (2024). https://doi.org/10.1145/3638044
5. Nyholm, T., et al.: MR and CT data with multiobserver delineations of organs in the pelvic area part of the gold atlas project. Med. Phys. **45**(3), 1295–1300 (2018). https://doi.org/10.1002/mp.12748
6. Li, J., et al.: Medshapenet a large-scale dataset of 3d medical shapes for computer vision. Biomed. Eng./ Biomedizinische Technik **70**(1), 71–90 (2025). https://doi.org/10.1515/bmt-2024-0396

7. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 165–174 (2019)
8. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12348, pp. 523–540. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58580-8_31
9. Dupont, E., Kim, H., Eslami, S.M.A., Rezende, D., Rosenbaum, D.: From data to functa: your data point is a function and you can treat it like one. https://doi.org/10.48550/arXiv.2201.12204
10. Liu, H.T.D., Williams, F., Jacobson, A., Fidler, S., Litany, O.: Learning smooth neural functions via lipschitz regularization. https://doi.org/10.48550/arXiv.2202.08345
11. Mu, J., Qiu, W., Kortylewski, A., Yuille, A., Vasconcelos, N., Wang, X.: A-SDF: learning disentangled signed distance functions for articulated shape representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13001–13011 (2021)
12. Miao, J., et al.: Manipulating vehicle 3D shapes through latent space editin. https://doi.org/10.48550/arXiv.2410.23931
13. Kong, F., Stocker, S., Choi, P.S., Ma, M., Ennis, D.B., Marsden, A.L.: SDF4CHD: generative modeling of cardiac anatomies with congenital heart defects. Med. Image Anal. **97**, 103293 (2024). https://doi.org/10.1016/j.media.2024.103293
14. Sørensen, K., et al.: Spatio-temporal neural distance fields for conditional generative modeling of the heart. In: Linguraru, M.G., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024. LNCS, vol. 15003, pp. 422–432. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72384-1_40
15. Dannecker, M., Kyriakopoulou, V., Cordero-Grande, L., Price, A.N., Hajnal, J.V., Rueckert, D.: CINA: conditional implicit neural atlas for spatio-temporal representation of fetal brains. In: Linguraru, M.G., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024. LNCS, vol. 15009, pp. 181–191. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72114-4_18
16. Wiesner, D., et al.: Generative modeling of living cells with so (3)-equivariant implicit neural representations. Med. Image Anal. **91**, 102991 (2024)
17. Ranade, A.V., et al.: Anatomical variations of the thyroid gland: possible surgical implications. Singapore Med. J. **49**(10), 831 (2008)
18. Wasserthal, J., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiol.: Artif. Intell. **5**(5), e230024 (2023). https://doi.org/10.1148/ryai.230024
19. Sitzmann, V., Martel, J.N.P., Bergman, A.W., Lindell, D.B., Wetzstein, G.: Implicit neural representations with periodic activation functions. https://doi.org/10.48550/arXiv.2006.09661



T2I-Diff: fMRI Signal Generation via Time-Frequency Image Transform and Classifier-Free Denoising Diffusion Models

Hwa Hui Tew[1], Junn Yong Loo[1(✉)], Yee-Fan Tan[1], Xinyu Tang[1], Hernando Ombao[2], Fuad Noman[1], Raphaël C.-W. Phan[1], and Chee-Ming Ting[1(✉)]

[1] School of Information Technology, Monash University Malaysia, Subang Jaya, Malaysia
{loo.junnyong,ting.cheeming}@monash.edu
[2] Statistics Program, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Abstract. Functional Magnetic Resonance Imaging (fMRI) is an advanced neuroimaging method that enables in-depth analysis of brain activity by measuring dynamic changes in the blood oxygenation level-dependent (BOLD) signals. However, the resource-intensive nature of fMRI data acquisition limits the availability of high-fidelity samples required for data-driven brain analysis models. While modern generative models can synthesize fMRI data, they often underperform because they overlook the complex non-stationarity and nonlinear BOLD dynamics. To address these challenges, we introduce T2I-Diff, an fMRI generation framework that leverages time-frequency representation of BOLD signals and classifier-free denoising diffusion. Specifically, our framework first converts BOLD signals into windowed spectrograms via a time-dependent Fourier transform, capturing both the underlying temporal dynamics and spectral evolution. Subsequently, a classifier-free diffusion model is trained to generate class-conditioned frequency spectrograms, which are then reverted to BOLD signals via inverse Fourier transforms. Finally, we validate the efficacy of our approach by demonstrating improved accuracy and generalization in downstream fMRI-based brain network classification. The code is available at repository.

Keywords: fMRI · Time-Frequency Image · Diffusion Models

1 Introduction

With the rapid advancement of artificial intelligence, deep generative modeling has shown promising capability in generating realistic variations of neuroimaging data [26]. Although the generation of brain connectivity has been extensively studied, research on the direct generation of fMRI signals remains limited [1,29].

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 640–650, 2026.
https://doi.org/10.1007/978-3-032-04947-6_61

fMRI signals are particularly valuable as they provide critical insights into neural activity, enabling more precise assessment and identification of neuropsychiatric and neurodevelopmental disorders [13,14]. Despite their significance, acquiring fMRI signals is challenging, often resulting in small sample sizes that limit the performance of data-driven brain analysis models [17]. These limitations can lead to data imbalances and a lack of temporal dynamics in acquired brain signals, ultimately affecting the accuracy of predictive and diagnostic models for neurological and psychiatric conditions [2,22]. To address these challenges, generative techniques have been explored for synthesizing fMRI signals, enabling data augmentation to enhance various downstream applications [15,19].

Generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have been widely used to synthesize realistic time series data [3]. While GANs are known for generating high-quality samples, VAEs are favored for their faster generation speed. However, both models face challenges in achieving stable training, often struggling with issues like mode collapse and optimization difficulties. For instance, Yoon et al. proposes TimeGAN, which jointly trains adversarial and supervised loss components within a learned embedding space, successfully preserving both the static and dynamic characteristics of synthetic time-series data [27]. Apart from that, TimeVAE incorporates time-series components into its encoder-decoder network, enhancing interpretability in time-series generation, where it has demonstrated success in reducing overall training time compared to adversarial methods [5].

Recently, denoising diffusion probabilistic models (DDPMs) have gained traction in time-series applications due to their robustness against model collapse [10]. However, their potential for capturing the complex intrinsic properties of brain signals remains underexplored [1,29]. For example, Coletta et al. propose different types of constraints to improve time-series generation: hard constraints enforce fixed points and global minima; soft constraints introduce penalties to guide the model towards desired temporal trends [4]. Furthermore, ImagenTime leveraged the time-frequency images to generate synthetic data for general time-series benchmarks, demonstrating the feasibility of capturing intricate spectral and temporal patterns [12]. The recently proposed Diffusion-TS model introduces a novel non-autoregressive diffusion model for time-series data, including fMRI signals. This framework explicitly captures temporal dynamics through transformer-based model architecture and disentangled seasonal-trend representations, achieving success across various generation tasks [28].

To address the challenge of insufficient inductive bias in dynamic brain characterization, this paper introduces T2I-Diff, an fMRI generation framework that leverages time-frequency image representations of BOLD signals and classifier-free denoising diffusion. This approach effectively captures both spectral and dynamic features underlying fMRI BOLD signals, which are essential for downstream brain disorder classification. Additionally, our framework incorporates classifier-free denoising diffusion and EDM sampling for high-fideliy generation of time-windowed fMRI spectrograms conditioned on subject classes. The generated time-frequency images are then reverted to BOLD signals via image-to-time

series transforms. Our method eliminates the need for additional classifiers to guide brain profiles, thereby reducing training redundancy and simplifying model optimization. We summarized our main contributions as follows:

1. Our proposed T2I-Diff framework is the first to integrate a time-frequency image transform to capture complex spatiotemporal and spectral features for fMRI BOLD signal generation and brain disorder classification.
2. We introduce classifier-free denoising diffusion and EDM sampling to generate conditioned time-windowed fMRI spectrograms without requiring an additional classifier. This enables T2I-Diff to dynamically modulates the diffusion process to guide the model in learning different brain signal profiles.
3. Our results shows that T2I-Diff demonstrates competitive performance on time-frequency image generation and brain disorder (MDD) classification complemented by class-conditioned synthetic fMRI BOLD signals.

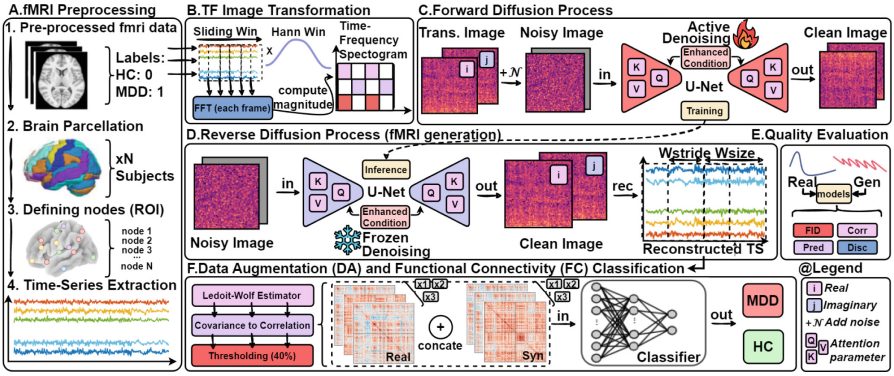


Fig. 1. Our framework T2I-Diff begins by (A.) preprocessing raw fMRI data in time domain and transforming into (B.) time-frequency spectrogram (image) using windowed Fourier transform. Next, the images are fed into the diffusion model executes both (C.) forward and (D.) reverse processes. Finally, we perform (E.) quality evaluation of the generation and apply (F.) DA on top of the real fMRI signals for MDD classification.

2 Methods

Time-Frequency Image Transformation. Figure 1 and 2 provides an overview of our proposed framework. Given high-dimensional fMRI signals from S subjects, denoted as $\mathcal{X}_S = \{\mathbf{x}_s\}_{s=1}^{S}$, where each subject $\mathbf{x}_s \in \mathbb{R}^{D \times T}$ consists of D regions of interest (ROIs) recorded over T time points, our objective is to learn the underlying real data distribution $p_{\text{data}}(\mathcal{X}_S)$ and generate a synthetic distribution $p_\theta(\mathcal{X}_S)$ that is statistically indistinguishable from the real data.

Unlike conventional time-series generative tasks that operate exclusively in the time domain, our approach transforms fMRI time series into a time-frequency spectrogram image using the windowed Fourier transform (WFT), defined as:

$$\mathbf{X}_s(m,k) = \sum_{n=0}^{N-1} x_s[n+m\,h]\,w[n]\,e^{-j2\pi\frac{k}{N}n}, \tag{1}$$

which serves as the input to the generative model. Here, $n \in \{1,2,\ldots,N\}$ is the local time index within an WFT window of length N; $m \in \{1,2,\ldots,M\}$ is the window index, where the total number of windows are computed as $M = \lfloor (T-N)/h \rfloor + 1$ with signal length T and hop size h; and $k \in \{1,2,\ldots,K\}$ is the frequency index, with the number of frequency bins defined as $K = N/2+1$ since only the first half of the spectrum is retained for real-valued signals.

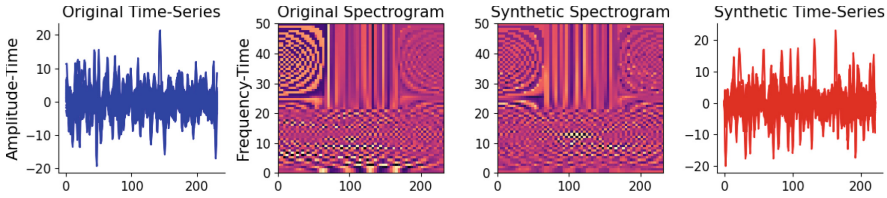


Fig. 2. Original vs synthetic BOLD signals and generated normalized spectrograms. Our framework directly generates BOLD signals as opposed to correlation matrices.

Given an univariate input signal $x_s \in \mathbb{R}^{1\times T}$, the WFT produces a spectrogram image representation $\mathbf{X}_s \in \mathbb{R}^{2\times K\times M}$, where the channels are doubled (2 channels) to store both the real $\mathbf{X}_s^{\mathrm{Re}}$ and imaginary $\mathbf{X}_s^{\mathrm{Im}}$ components of the WFT output, and the spatial dimensions (height and width) of the image are determined by the frequency bins K and time frames M, respectively. We further perform component-wise normalization to accentuate between high and low frequencies. This structured representation allows generative models to process time-series data as images while preserving both spectral and temporal characteristics, enabling robust generative modeling of the fMRI BOLD signals.

To reconstruct the original fMRI time series spectrogram representation, we first unnormalize the real and imaginary components, we merge them to form a complete complex spectrogram $\hat{\mathbf{X}}_s(m,k) = \mathbf{X}_s^{\mathrm{Re}}(m,k) + j\,\mathbf{X}_s^{\mathrm{Im}}(m,k)$. For each window indexed by m and for local time indices n, we compute the inverse fourier transform to reconstruct the fMRI BOLD signals as follows:

$$x_s[n+m\,h] = \frac{1}{N}\sum_{k=0}^{K-1} \mathbf{X}_s(m,k)\,e^{j\,2\pi\frac{k}{N}n}. \tag{2}$$

Finally, these segments are recombined using the overlapping method with Hann window to obtain the reconstructed time-domain signal $\tilde{x_s}$. This process ensures that both the spectral and temporal characteristics of the original fMRI BOLD signals are preserved.

Classifier-Free Denoising Diffusion. Rather than incrementally adding noise through a Markov chain, we employ EDM framework [7], which perturb the ground-truth fMRI spectrogram images using a single noise level σ_t drawn from a lognormal distribution. At each diffusion step $t \in \{0, 1, \ldots, T\}$, the images are corrupted according to the perturbation kernel $p_\theta(x_t \mid x_0)$ which corresponds to $x_t = x_0 + n_t$, where the total noise $n_t = \sigma_t \cdot \epsilon_t$ with $\sigma_t = \exp(\eta_t \cdot P_{\text{std}} + P_{\text{mean}})$, $\eta_t \sim \mathcal{N}(0, 1)$ and $\epsilon_t \sim \mathcal{N}(0, I)$. Here, n_t represents a sample from a Gaussian distribution whose variance is modulated by the lognormal factor σ_t. This process gradually transforms the fMRI spectrogram images into non-informative Gaussian noise. The reverse process aims to learn the conditional distribution $p_\theta(x_{t+1} \mid x_t)$ to reconstruct the original fMRI spectrogram images. In particular, a deep neural network denoiser $\epsilon_\theta(x_t, \sigma_t)$ is trained to predict the clean image x_0 from its noisy version x_t at noise level σ_t across time steps. Specifically, the noise scheduling is defined as $\sigma_t = \left(\sigma_{\max}^{1/\rho} + \frac{t}{T-1}\left(\sigma_{\min}^{1/\rho} - \sigma_{\max}^{1/\rho}\right)\right)^{\rho}$. Using Heun's second-order method, the reverse-diffusion update is given by:

$$x_{t+1} = x'_t + (\sigma_{t+1} - \sigma'_t) \cdot \frac{1}{2}\left[\frac{x'_t - \epsilon_\theta(x'_t, \sigma'_t)}{\sigma'_t} + \frac{x'_{t+1} - \epsilon_\theta(x_{t+1}, \sigma_{t+1})}{\sigma_{t+1}}\right], \quad (3)$$

where $\sigma'_t = \sigma_t + \gamma_t \sigma_t$ and $x'_t = x_t + \sqrt{\sigma'^2_t - \sigma^2_t}\,\epsilon'_t$ is perturbed by noise $\epsilon_t \sim \mathcal{N}(0, S^2_{\text{noise}} I)$, where both γ_t and S_{noise} are EDM hyperparameters. By iteratively applying this update from $t = T$ down to $t = 0$, a synthetic fMRI spectogram image $x_0 \sim p_\theta(x)$ is sampled from the marginal data distribution.

To achieve class-conditioned generation, we apply classifier-free guidance and parameterize the conditional distribution $p_\theta(x_t|c)$ via a conditional noise-prediction model $\epsilon_\theta(x_t, t, c)$, where c is the subject class. For unconstrained generation, we then input a null token $\varnothing$ for the unconditional model $\epsilon_\theta(x_t, t) = \epsilon_\theta(x_t, t, \varnothing)$. Concurrently, these noise-prediction models are trained to minimize the discrepancy between the true noise ϵ and the predicted noise ϵ_θ across timesteps, via the noise-matching objective $\mathcal{L}(\theta) = \mathbb{E}_{x_t, \epsilon, t, c}\left[\|\epsilon - \epsilon_\theta\|^2\right]$. The unconditional and conditional models are jointly trained by randomly setting c to the unconditional class identifier $\varnothing$ with some discrete probability $p_\varnothing$.

3 Experiments and Results

Data Acquisition and Pre-processing. We preprocessed the dataset from REST-meta-MDD Consortium database [24] using Data Processing Assistant for Resting-State fMRI (DPARSF) [25]. This resting-state fMRI dataset consist of 250 healthy controls (HC) and 227 Major Depressive Disorder (MDD) patients. The data were acquired using a Siemens (Tim Trio 3T) scanner (TR/TE = 2000/30 ms, 3 mm slice thickness). The brain was parcellated into 116 region-of-interests (ROIs), including cortical and subcortical areas, and the mean time series of 232 time points for each ROI was extracted using the Automated Anatomical Labeling (AAL) atlas.

Table 1. Comparison of fMRI signal generation quality with SOTA time-series generation and our proposed T2I-Diff.

	CoT-GAN	DiffTime	DiffWave	TimeVAE	TimeGAN	Diffusion-TS	**T2I-Diff**
Context-FID	7.813±.550	0.340±.015	0.244±.018	14.449±.969	0.126±.002	0.105±.006	1.384±.107
Correlational	26.824±.449	1.501±.048	3.927±.049	17.296±.526	23.502±.039	1.411±.042	4.121±.094
Discriminative	0.492±.018	0.245±.051	0.402±.029	0.476±.044	0.484±.042	0.167±.023	0.400±.059
Predictive	0.185±.003	0.100±.000	0.101±.000	0.113±.003	0.126±.002	0.099±.000	0.102±.001

Table 2. Ablation study of short-term and long-term fMRI signal generations across sequence lengths 24, 64, 128 and 256.

24	64	128	256	24	64	128	256
Context-FID				**Correlation**			
1.384±0.107	7.141±0.789	10.655±1.001	1.661±0.146	4.121±0.094	4.737±0.136	3.525±0.101	4.124±0.243
Discriminative				**Predictive**			
0.400±0.059	0.280±0.268	0.176±0.090	0.360±0.164	0.102±0.001	0.100±0.001	0.098±0.001	0.095±0.002

State-of-the-Art Baselines. i) Quality Evaluation: Our proposed T2I-Diff model is compared against six time-series generative model baselines. These baselines include generative adversarial networks (GANs) and diffusion models such as CoT-GAN [23], DiffTime [4], DiffWave [9], TimeVAE [5] and Diffusion-TS [28]. ii) Classification Score: We further evaluate the performance of our proposed model with a special designed classifier for brain connectvity [8]. The baselines include 1D-Deep Convolutional GAN (1D-DCGAN) [16] and Wasserstein GAN Gradient Policy(WGAN-GP) [6].

Implementation Details. i) Connectivity Network Construction: The functional connectivity of brain networks is constructed using Ledoit-Wolf (LDW) regularized shrinkage covariance estimator to keep the strongest $\tau = 40\%$ connections and set other connections to zeros. This results in FCs of size 116×116 for each subject. ii) T2I-Diff Training: The proposed T2I-Diff framework generates the fMRI signals corresponding to the subjects' condition (HC and MDD). The BrainNetCNN classifier then discriminates between the HC and MDD subjects. We train the T2I-Diff via an Adam optimizer using a learning rate of $3e^{-4}$ for 1000 epochs. For all experiments, we use 18 diffusion sampling steps and experiment in image sizes of 8×8, 16×16, 32×32, and 64×64. iii) Data Augmentation and Classifier Training: The trained T2I-Diff is used to augment fMRI signals of $1\times$, $2\times$ and $3\times$ on top of the real fMRI signals. For our classifier, the L2 regularization weight decay from 10^{-8} to 10^{-2}, scheduler learning rate reduce factor from 0.1 to 0.9, batch size from 5 to 16 same as in [18]. These model hyperparameters are selected based on a 5-fold stratified cross-validation.

3.1 Overall Performance

We first trained our diffusion models unconstrained to produce similar outputs. Then, we followed the standard setting for the evaluation of time series generation from Diffusion-TS [28]. The purpose of our experiments is to demonstrate the effectiveness of the proposed time-frequency representation and classifier-free diffusion models in achieving a trade-off between FID and predictive scores, comparable to other time-series generative models. Rather than solely aiming for state-of-the-art sample quality metrics on these benchmarks, our primary goal is to validate the ability of our approach to model complex spatiotemporal patterns and excel in capturing conditional distribution over time.

Table 3. Classification performance of different classifiers trained on the ground-truth data and increasing amount of augmented time series data using our proposed model.

Method	Train Set	Accuracy	Recall	Precision	F1-Score	ROC
W/O Augmentation	Real	58.90 ± 2.98	58.90 ± 2.98	59.56 ± 2.74	58.39 ± 3.09	59.00 ± 2.56
1D-DCGAN	Real + Synth 1×	62.94 ± 2.01	62.94 ± 2.01	63.43 ± 2.20	62.23 ± 2.68	62.71 ± 2.26
	Real + Synth 2×	65.04 ± 2.02	65.04 ± 2.02	66.35 ± 2.13	64.12 ± 2.10	64.74 ± 2.08
	Real + Synth 3×	58.21 ± 2.98	58.21 ± 2.98	55.70 ± 6.58	52.86 ± 4.14	57.38 ± 3.11
WGAN-GP	Real + Synth 1×	66.02 ± 4.25	66.02 ± 4.25	66.22 ± 4.24	65.93 ± 4.20	65.95 ± 4.13
	Real + Synth 2×	64.76 ± 4.25	64.76 ± 4.25	65.67 ± 4.08	64.23 ± 4.52	64.73 ± 4.14
	Real + Synth 3×	64.56 ± 3.18	64.56 ± 3.18	64.78 ± 3.17	64.38 ± 3.15	64.41 ± 3.08
T2I-Diff (Ours)	Real + Synth 1×	66.87 ± 3.22	66.87 ± 3.22	67.06 ± 3.34	66.83 ± 3.21	67.26 ± 6.00
	Real + Synth 2×	65.41 ± 2.37	65.41 ± 2.37	66.30 ± 1.67	64.73 ± 2.80	65.75 ± 3.22
	Real + Synth 3×	66.03 ± 1.75	66.03 ± 1.75	66.50 ± 1.32	65.85 ± 1.82	66.58 ± 5.33

fMRI Signal Generation Quality. Table 1 compares our T2I-Diff framework with state-of-the-art time-series generative models across four evaluation metrics for a time-series generation length of 24. T2I-Diff demonstrates competitive overall performance, surpassing strong time-series generation models such as CoT-GAN, DiffWave, TimeGAN, and TimeVAE. Notably, the low discriminative and predictive scores suggest that T2I-Diff effectively generates synthetic fMRI BOLD signals with in-distribution temporal patterns and spectral features that may not be fully captured in the time domain. By leveraging the WFT image transform, the diffusion model not only preserves global contextual information but also models the complex spatiotemporal dynamics inherent in fMRI signals. While T2I-Diff does not achieve the highest overall scores, our results demonstrate that the time-series-to-image generative framework is well-suited for modeling high-dimensional, complex fMRI time series.

Ablation Study. Table 2 assess the impact of sequence length on model performance, we conducted ablation studies on time-series generation lengths of 64, 128, and 256 to contrast the sequence length of 24. Shorter sequences exhibit a

low context-fid score and high predictive score, but their lower predictive score suggests a limited ability to capture temporal dynamics due to insufficient contextual information. Mid-length sequences (64 and 128) show improved context-fid scores but struggle with increased complexity. The longest sequence (256) achieves the best context-fid and predictive scores, effectively capturing both global and fine-grained features.

Classification Score. To validate the fidelity of the generated samples, we evaluate the classification performance of BrainNetCNN and compare it against ID-DCGAN and WGAN-GP on our proposed fMRI dataset. Here, we maintain the original sequence length rather than truncating, as supported by the performance of long sequence generation in Table 1. Table 3 presents the classification results on the 5-fold cross-validation test set. Notably, T2I-Diff achieves the highest accuracy in the 1× data augmentation setting. Furthermore, the results indicate that as the augmentation range increases, our model exhibits lower variance, suggesting that the synthetic data generalizes well across varying augmentation levels while maintaining consistent classification performance. The results highlight that T2I-Diff not only enhances the diversity of synthetic fMRI samples but also effectively captures and preserves critical structural and functional patterns for downstream tasks such as brain disorder classification.

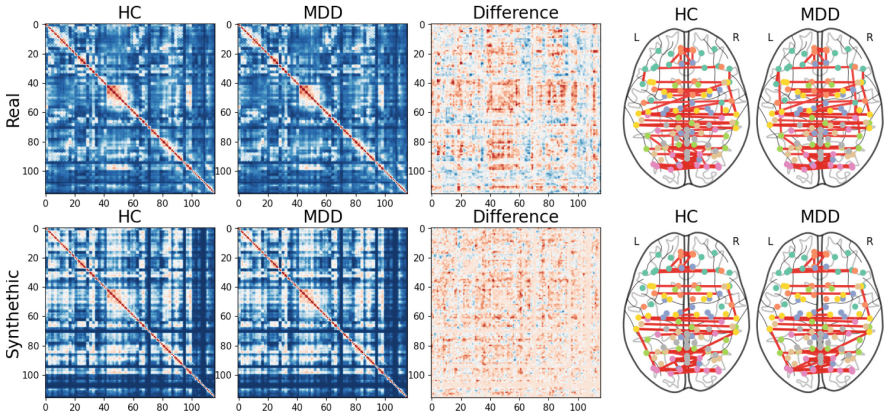


Fig. 3. Real and synthethic (left) connectivity pattern and (right) brain networks.

3.2 Functional Connectivity (FC) Visualization

To further evaluate the quality of generated data, we derived functional connectivity (FC) measures from the synthetic fMRI BOLD signals to assess differences between healthy controls (HC) and major depressive disorder (MDD) patients. We computed the average connectivity using a threshold of 0.6 to highlight significant connections. Our analysis reveals that the synthetic FC closely aligns

with the functional changes observed in the real FC distribution. Furthermore, connectogram comparisons between HC and MDD in both real and synthetic FC data indicate a reduction in connectivity within the left superior frontal gyrus (FrontalSupL) and decreased connectivity between the left middle frontal gyrus (FrontalMidL) and the anterior cingulate cortex (CingulumAntL). The results suggest impaired cognitive functions, such as difficulties in decision-making and emotion regulation, reinforcing the biological plausibility of the generated data.

4 Conclusions and Future Work

In this paper, we propose T2I-Diff, which effectively captures both temporal dynamics and spectral evolution underlying the ground-truth data distribution for accurate brain signal generation. For future work, we aim to further validate MDD classification using graph-based deep learning models [20,21]. Furthermore, we plan to incorporate energy-based models to identify out-of-distribution (OOD) patterns in brain spectogram associated with neurological disorders [11].

Acknowledgments. This work was supported by the Advanced Computing Platform (ACP), Monash University Malaysia. The work of Junn Yong Loo is supported by Monash University under the SIT Collaborative Research Seed Grants 2024 I-M010-SED-000242.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Avberšek, L.K., Repovš, G.: Deep learning in neuroimaging data analysis: applications, challenges, and solutions. Front. Neuroimaging **1**, 981642 (2022)
2. Bollmann, S., Barth, M.: New acquisition techniques and their prospects for the achievable resolution of fmri. Prog. Neurobiol. **207**, 101936 (2021)
3. Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G.: Deep generative modelling: a comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. IEEE Trans. Pattern Anal. Mach. Intell. **44**(11), 7327–7347 (2021)
4. Coletta, A., Gopalakrishnan, S., Borrajo, D., Vyetrenko, S.: On the constrained time-series generation problem. Adv. Neural Inf. Process. Syst. **36**, 61048–61059 (2023)
5. Desai, A., Freeman, C., Wang, Z., Beaver, I.: Timevae: a variational auto-encoder for multivariate time series generation. arXiv preprint arXiv:2111.08095 (2021)
6. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Adv. Neural Inf. Process. Syst. **30** (2017)
7. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Adv. Neural Inf. Process. Syst. **35**, 26565–26577 (2022)
8. Kawahara, J., et al.: Brainnetcnn: convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage **146**, 1038–1049 (2017)

9. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: a versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020)
10. Lin, L., Li, Z., Li, R., Li, X., Gao, J.: Diffusion models for time-series applications: a survey. Front. Inf. Technol. Electron. Eng. **25**(1), 19–41 (2024)
11. Loo, J.Y., et al.: Learning energy-based generative models via potential flow: a variational principle approach to probability density homotopy matching. Trans. Mach. Learn. Res. (2025)
12. Naiman, I., Berman, N., Pemper, I., Arbiv, I., Fadlon, G., Azencot, O.: Utilizing image transforms and diffusion models for generative modeling of short and long time series. Adv. Neural Inf. Process. Syst. **37**, 121699–121730 (2024)
13. Noman, F., Ting, C.M., Kang, H., Phan, R.C.W., Ombao, H.: Graph autoencoders for embedding learning in brain networks and major depressive disorder identification. IEEE J. Biomed. Health Inform. **28**(3), 1644–1655 (2024)
14. Noman, F., Yap, S.Y., Phan, R.C.W., Ombao, H., Ting, C.M.: Graph autoencoder-based embedded learning in dynamic brain networks for autism spectrum disorder identification. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 2891–2895 (2022)
15. Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E.: Methods to detect, characterize, and remove motion artifact in resting state fmri. Neuroimage **84**, 320–341 (2014)
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
17. Tan, Y.F., Loo, J.Y., Ting, C.M., Noman, F., Phan, R.C.W., Ombao, H.: BrainFC-CGAN: a conditional generative adversarial network for brain functional connectivity augmentation and aging synthesis. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1511–1515. IEEE (2024)
18. Tan, Y.F., Ting, C.M., Noman, F., Phan, R.C.W., Ombao, H.: Graph-regularized manifold-aware conditional wasserstein gan for brain functional connectivity generation. arXiv preprint arXiv:2212.05316 (2022)
19. Tan, Y.F., Ting, C.M., Noman, F., Phan, R.C.W., Ombao, H.: A unified framework for static and dynamic functional connectivity augmentation for multi-domain brain disorder classification. In: Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP), pp. 635–639 (2023)
20. Tew, H.H., et al.: ST-HCSS: deep spatio-temporal hypergraph convolutional neural network for soft sensing. arXiv preprint arXiv:2501.02016 (2025)
21. Tew, H.H., et al.: Kans: knowledge discovery graph attention network for soft sensing in multivariate industrial processes. In: 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4377–4383. IEEE (2024)
22. Ting, C.M., Skipper, J.I., Noman, F., Small, S.L., Ombao, H.: Separating stimulus-induced and background components of dynamic functional connectivity in naturalistic fmri. IEEE Trans. Med. Imaging **41**(6), 1431–1442 (2022)
23. Xu, T., Wenliang, L.K., Munn, M., Acciaio, B.: COT-GAN: generating sequential data via causal optimal transport. Adv. Neural Inf. Process. Syst. **33**, 8798–8809 (2020)
24. Yan, C.G., et al.: Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. Proc. Natl. Acad. Sci. **116**(18), 9078–9083 (2019)
25. Yan, C., Zang, Y.: DPARSF: a matlab toolbox for "pipeline" data analysis of resting-state fmri. Front. Syst. Neurosci. **4**, 1377 (2010)

26. Yap, S.Y., et al.: A deep probabilistic spatiotemporal framework for dynamic graph representation learning with application to brain disorder identification. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pp. 5353–5361 (2024)
27. Yoon, J., Jarrett, D., Van der Schaar, M.: Time-series generative adversarial networks. Adv. Neural Inf. Process. Syst. **32** (2019)
28. Yuan, X., Qiao, Y.: Diffusion-ts: interpretable diffusion for general time series generation. arXiv preprint arXiv:2403.01742 (2024)
29. Zhao, H., et al.: Diffusion models for computational neuroimaging: a survey. arXiv preprint arXiv:2502.06552 (2025)

Tied Prototype Model for Few-Shot Medical Image Segmentation

Hyeongji Kim[1(✉)], Stine Hansen[2], and Michael Kampffmeyer[1,3]

[1] Department of Physics and Technology, UiT - The Arctic University of Norway, Tromsø, Norway
hyeongji.kim@uit.no
[2] Norwegian Centre for Clinical Artificial Intelligence (SPKI), University Hospital of North-Norway, Tromsø, Norway
[3] Norwegian Computing Center, Oslo, Norway

Abstract. Common prototype-based medical image few-shot segmentation (FSS) methods model foreground and background classes using class-specific prototypes. However, given the high variability of the background, a more promising direction is to focus solely on foreground modeling, treating the background as an anomaly—an approach introduced by ADNet. Yet, ADNet faces three key limitations: dependence on a single prototype per class, a focus on binary classification, and fixed thresholds that fail to adapt to patient and organ variability. To address these shortcomings, we propose the Tied Prototype Model (TPM), a principled reformulation of ADNet with tied prototype locations for foreground and background distributions. Building on its probabilistic foundation, TPM naturally extends to multiple prototypes and multi-class segmentation while effectively separating non-typical background features. Notably, both extensions lead to improved segmentation accuracy. Finally, we leverage naturally occurring class priors to define an ideal target for adaptive thresholds, boosting segmentation performance. Taken together, TPM provides a fresh perspective on prototype-based FSS for medical image segmentation. The code can be found at https://github.com/hjk92g/TPM-FSS.

Keywords: Few-Shot Segmentation · Medical Image Segmentation · Multi-Class Segmentation · Prototype Model

1 Introduction

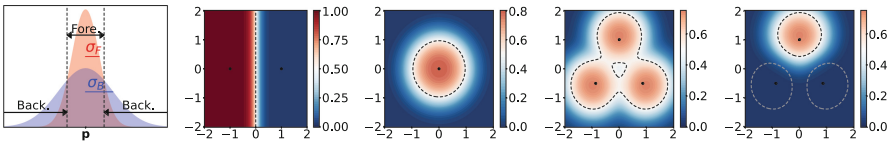
Medical image segmentation is a critical component of numerous clinical applications such as diagnostics [19] and treatment planning [3]. While supervised deep learning approaches can achieve good performance, their applications are constrained by the limited availability of annotated medical images. To address this, few-shot segmentation (FSS) approaches have been proposed to effectively adapt models trained with labeled datasets to new, previously unseen, classes.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 651–661, 2026.
https://doi.org/10.1007/978-3-032-04947-6_62

The dominant FSS approach employs prototype networks [16], first introduced by PANet [21], where each class is represented by a single prototype obtained through mean average pooling (MAP). Recognizing the limitation of a single prototype in modeling rich feature variants, PPNet [11] proposed using multiple prototypes for each class to improve expressiveness. More recently, there has been growing interest in leveraging self-supervised learning (SSL) to circumvent the need for labeled medical data altogether. ALPNet [13] pioneered the use of SSL by using pseudo-labels generated from superpixels, with additional local prototypes aiding intra-class local information. However, the background class typically exhibits significant heterogeneity, making it difficult to model using a fixed number of prototypes. Recognizing this, ADNet [6] employs an anomaly-detection-inspired approach, focusing on modeling the foreground prototype while treating the background as anomalous and separating regions via inside-outside (IO) classification. It assigns anomaly scores to each spatial location, using a fixed threshold at inference for segmentation. This simple method has demonstrated greater robustness to background heterogeneity relative to ALPNet and has been successfully used for the detection of brain tumors [1], ischemic stroke lesions [18], and lung lesions [17].

However, ADNet model [6] has three main drawbacks. First, its reliance on a single foreground prototype limits the expressiveness of the foreground class in medical images with significant intra-class variation. While non-anomaly-detection multi-prototype methods exist [5,13], they struggle to capture the full diversity of the background class. Meanwhile, existing anomaly-detection-based multi-prototype variants [14,23,25,26] of ADNet involve cumbersome modifications such as the introduction of new layers or hyperparameters. Hence, it is desirable for a more grounded yet simple approach that efficiently leverages multiple foreground prototypes at the feature level. Second, ADNet is limited to binary classification settings. While ADNet++ [7] extends it to multi-class classification, it only adapts the inference phase and thus does not effectively model the class relationships in training. Finally, ADNet's fixed threshold is insufficient to accommodate the inherent variability across patients and organs in diverse applications. While adaptive thresholds have been proposed to address this by utilizing support features [4], query information [15], or both [24], their threshold values are learned through cross-entropy-based losses, which may not necessarily optimize segmentation accuracy as further discussed in Sect. 3.4.

To address these issues, we take a step back and propose the Tied Prototype Model (TPM), a principled reformulation of ADNet [6] that leverages tied (shared) prototype locations for foreground and background distributions. As visualized in Fig. 1, the distributional foundation of TPM allows for its natural extension to both multiple prototypes and multi-class classification settings while retaining its simplicity and ability to separate non-typical background features. In particular, we propose a Gaussian mixture model (GMM)-based approach for multi-prototype segmentation, demonstrating its effectiveness in improving segmentation accuracy. Moreover, we introduce multi-foreground training for multi-class segmentation and show that it consistently offers superior represen-



(a) 1D-distrib. (b) Std. model (c) TPM-SP (d) TPM-MP (e) TPM-MC

Fig. 1. (a) TPM class distributions. The (first) foreground class probability $p(F|x)$, based on (b) the standard prototype model, (c) TPM with a single prototype (SP), similar to the ADNet [6]—which is limited to spherical space, (d) TPM with multiple prototypes (MP), and (e) TPM for multi-class (MC) classification. For (b)–(e), the xy-axis shows the 2D Euclidean feature space positions, and the points denote prototypes. Dashed lines indicate decision boundaries for the shown (black) and the other (gray) foreground classes. Compared to standard prototype models, all TPM models achieve IO separation.

tation learning capability. Lastly, by employing naturally occurring class priors and feature distance distributions, we provide an ideal threshold estimation approach, aiming to enhance segmentation accuracy.

2 Preliminaries

2.1 Self-supervised Few-Shot Learning Problem Setting

We adopt the self-supervised few-shot learning from [6], where pseudo-labels for unlabeled volumes are generated by performing supervoxel segmentation and randomly sampling a supervoxel as the foreground. Two 2D slices with the supervoxel are then selected to serve as support and query images, with one being augmented. The support image $\mathbf{X}^s$ and its pseudo-label $\mathbf{Y}^s$, where $\mathbf{Y}^s_c$ denotes the mask of class c, guide the segmentation of the query image $\mathbf{X}^q$, all of size (H, W). A feature extractor network f_θ encodes images into feature maps $\mathbf{F}^s = f_\theta(\mathbf{X}^s)$ and $\mathbf{F}^q = f_\theta(\mathbf{X}^q)$, with spatial dimensions (H', W') and feature dimension d. The upsampled predicted query label $\hat{\mathbf{Y}}^q$ is then compared to label $\mathbf{Y}^q$, training the network f_θ with a cross-entropy-based segmentation loss.

2.2 Anomaly Detection-Inspired Few-Shot Segmentation

In ADNet [6], the foreground prototype $\mathbf{p}$ is computed by upsampling the feature maps to the original size and applying masked average pooling (MAP) as:

$$\mathbf{p} = \frac{\sum_{\mathbf{r}} \mathbf{F}^s(\mathbf{r}) \odot \mathbf{Y}^s_F(\mathbf{r})}{\sum_{\mathbf{r}} \mathbf{Y}^s_F(\mathbf{r})}, \tag{1}$$

with $\odot$ as the Hadamard product, F the foreground class, and $\mathbf{r}$ the pixel location. It defines the anomaly score $S(\mathbf{r}) = -\alpha \cos(\mathbf{F}^q(\mathbf{r}), \mathbf{p})$ for each query feature $\mathbf{F}^q(\mathbf{r})$, where $\alpha = 20$ is a scaling factor following [12], and $\cos(\cdot, \cdot)$ is the cosine similarity. Then, ADNet estimates foreground class probability as:

$$p(F|\mathbf{F}^q(\mathbf{r})) = 1 - \text{sig}(S(\mathbf{r}) - T_S), \tag{2}$$

where sig$(\cdot)$ is the sigmoid function with a steepness parameter $\kappa = 0.5$, and T_S is a learnable anomaly score threshold. This estimate is then upsampled and, given the binary task focus, is combined with its complement to generate the predicted label $\hat{\mathbf{Y}}^q$.

3 Method

3.1 Tied Prototype Model

The tied prototype model (TPM), with our established equivalence to ADNet [6] under a certain condition, adopts a probabilistic view of discriminative classification. This provides valuable insights for advancing few-shot medical image segmentation while addressing ADNet's limitations without modifying its architecture.

As illustrated in Fig. 1a, the key idea of our model is to use class distributions with a tied (shared) center position $\mathbf{p}$ for foreground and background classes while differing in their dispersion parameters. This contrasts with standard classification, which uses distinct centers for different classes. Though counterintuitive, enforcing the same center for both foreground and background classes is essential in separating the background class, as shown in Fig. 1c and Theorem 1.

TPM assumes that the foreground and background class distributions follow multivariate normal distributions, $\mathcal{N}(\mathbf{p}, \sigma_F^2 I)$ and $\mathcal{N}(\mathbf{p}, \sigma_B^2 I)$, respectively. Here, the tied prototype $\mathbf{p}$ serves as the location parameter, while σ_F and σ_B are class standard deviations, with $\sigma_F < \sigma_B$. The symbol I denotes the identity matrix. By applying Bayes' theorem, the foreground class probability is given by:

$$p(F|\mathbf{F}^q(\mathbf{r})) = \frac{p_F \phi(\mathbf{p}, \sigma_F; \mathbf{r})}{p_F \phi(\mathbf{p}, \sigma_F; \mathbf{r}) + p_B \phi(\mathbf{p}, \sigma_B; \mathbf{r})} = \frac{1}{1 + \frac{p_B}{p_F} \frac{\phi(\mathbf{p}, \sigma_B; \mathbf{r})}{\phi(\mathbf{p}, \sigma_F; \mathbf{r})}} \tag{3}$$

where p_F and p_B denote class priors and $\phi(\mathbf{p}, \sigma; \mathbf{r})$ represents the normal distribution density function with a mean $\mathbf{p}$ and covariance matrix $\sigma^2 I$. The choices of priors are discussed in Sect. 3.4.

Figure 1c visualizes the estimated class probability of TPM with a single prototype, assigning higher probabilities to features closer to the prototype $\mathbf{p}$ while giving lower values to those further from the prototype. This inside-outside (IO) classification behavior highlights the model's effectiveness in distinguishing typical features from diverse non-typical features, which is particularly useful for foreground and background classification. As illustrated in Fig. 1b, this beneficial IO separation property is absent in the standard prototype model, which relies on distinct (non-tied) prototype positions and assigns class probabilities along the direction of the prototypes' difference.

Theorem 1. *In one foreground classification with a single prototype with a unit spherical embedding $\mathbb{S}^{d-1}$, the tied prototype model is equivalent to ADNet [6].*

Proof. To establish equivalence, we must show that Eqs. (2 and 3) are identical under a proper parameter correspondence. We first define $\mathrm{TP}(\mathbf{r})$ to simplify Eq. (3) and replace the multiplication with the exponential function, yielding:

$$\mathrm{TP}(\mathbf{r}) = \frac{1}{1 + \frac{p_B}{p_F} \frac{\phi(\mathbf{p}, \sigma_B; \mathbf{r})}{\phi(\mathbf{p}, \sigma_F; \mathbf{r})}} = \frac{1}{1 + \exp\left(\ln\left(\frac{\phi(\mathbf{p}, \sigma_B; \mathbf{r})}{\phi(\mathbf{p}, \sigma_F; \mathbf{r})}\right) + \ln\left(\frac{p_B}{p_F}\right)\right)}. \tag{4}$$

By substituting the density function $\phi(\cdot, \cdot)$ with its definition, we obtain:

$$\mathrm{TP}(\mathbf{r}) = \frac{1}{1 + \exp\left(\frac{1}{2} \left\|\mathbf{F}^q(\mathbf{r}) - \mathbf{p}\right\|^2 \left(\frac{1}{\sigma_F^2} - \frac{1}{\sigma_B^2}\right) + d \ln\left(\frac{\sigma_F}{\sigma_B}\right) + \ln\left(\frac{p_B}{p_F}\right)\right)}. \tag{5}$$

We then set $\alpha = 2\left(\frac{1}{\sigma_F^2} - \frac{1}{\sigma_B^2}\right)$ and $T_S = 2 \ln\left(\frac{p_F}{p_B}\right) - 2d \ln\left(\frac{\sigma_F}{\sigma_B}\right) - \alpha$ to obtain:

$$\mathrm{TP}(\mathbf{r}) = \frac{1}{1 + \exp\left(\frac{1}{2}\left(\left(\frac{1}{2}\left\|\mathbf{F}^q(\mathbf{r}) - \mathbf{p}\right\|^2 - 1\right)\alpha - T_S\right)\right)}. \tag{6}$$

Since we use the unit spherical embedding $\mathbb{S}^{d-1}$, embedding vectors $\mathbf{F}^q(\mathbf{r})$ and $\mathbf{p}$ satisfy the equation $\frac{2 - \left\|\mathbf{F}^q(\mathbf{r}) - \mathbf{p}\right\|^2}{2} = \cos(\mathbf{F}^q(\mathbf{r}), \mathbf{p})$, leading to:

$$\mathrm{TP}(\mathbf{r}) = \frac{1}{1 + \exp\left(\frac{1}{2}\left(-\alpha \cos(\mathbf{F}^q(\mathbf{r}), \mathbf{p}) - T_S\right)\right)}. \tag{7}$$

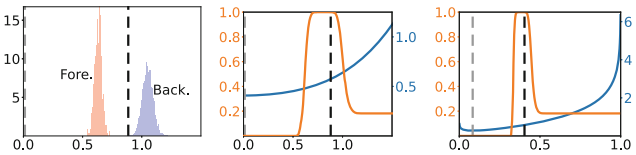
Using the definitions of the anomaly score $S(\mathbf{r})$ and the sigmoid function $\mathrm{sig}(\cdot)$, we can observe that Eq. (7) is equivalent to Eq. (2). □

Theorem 1 reveals that our TPM, under specific conditions, is a reparameterized ADNet [6]. Unlike ADNet which inherently uses spherical embedding due to its reliance on cosine similarity, our model can utilize other geometries. Furthermore, it enables extensions beyond a single prototype as discussed below.

3.2 Binary Classification with Multiple Prototypes

Rooted in its distributional formulation, TPM naturally extends to multiple prototypes modeled by Gaussian mixture models (GMMs). As illustrated in Fig. 1d, this extension allows it to capture diversity in foreground features while retaining the ability to distinguish non-typical background features.

To learn the GMM parameters from the foreground features, we apply the EM algorithm from [22], but without enforcing equal mixture weights. This algorithm alternates between the E-step and M-step, to extract multiple prototypes $\mathbf{p_m}$ along with their weights $w_\mathbf{m}$, satisfying $\sum w_\mathbf{m} = 1$. When multiple prototypes are obtained for the foreground class, we can simply use the same prototypes for the background class as well. Specifically, GMM $\sum_\mathbf{m} w_\mathbf{m} \phi(\mathbf{p_m}, \sigma_F; \mathbf{r})$ is used for the foreground class distribution, while the corresponding GMM $\sum_\mathbf{m} w_\mathbf{m} \phi(\mathbf{p_m}, \sigma_B; \mathbf{r})$ represents the background class distribution. Using Bayes'



(a) Dist. distrib. (b) Varying T_D (c) Varying p_F

Fig. 2. (a) Distributions of distances to a prototype for Gaussian samples projected onto a sphere. Dice score (orange) vs. CE (blue) for (b) varying distance threshold T_D and (c) varying foreground class prior p_F. Dashed lines indicate values for the ideal threshold (black) and the CE-minimizing threshold (gray). (Color figure online)

theorem, the foreground class probability $p_{\mathrm{MP}}(F|\mathbf{F}^q(\mathbf{r}))$, estimated from multiple prototypes and associated class priors, is given by:

$$p_{\mathrm{MP}}(F|\mathbf{F}^q(\mathbf{r})) = \frac{p_{F;\mathrm{MP}} \sum_{\mathbf{m}} w_{\mathbf{m}} \phi(\mathbf{p}_{\mathbf{m}}, \sigma_F; \mathbf{r})}{p_{F;\mathrm{MP}} \sum_{\mathbf{m}} w_{\mathbf{m}} \phi(\mathbf{p}_{\mathbf{m}}, \sigma_F; \mathbf{r}) + p_{B;\mathrm{MP}} \sum_{\mathbf{m}} w_{\mathbf{m}} \phi(\mathbf{p}_{\mathbf{m}}, \sigma_B; \mathbf{r})}. \quad (8)$$

3.3 Multi-class Classification

The distributional foundation of TPM allows for another expansion: multi-class classification. While multiple prototypes can be used for each foreground class, this work focuses on the simplified case of using a single prototype $\mathbf{p}_{F_i}$, obtained by MAP in Eq. (1), for foreground class F_i. Figure 1e visualizes the class probability $p(F_1|x)$ of the first foreground class with prototype $\mathbf{p}_{F_1} = (0, 1)$. The model effectively distinguishes features from other foreground classes while ensuring the separation of non-typical background features.

With a similar assumption on class distributions as in Sects. 3.1 and 3.2, the class probability $p_{\mathrm{MC}}(F_i|\mathbf{F}^q(\mathbf{r}))$ of foreground class F_i can be obtained as:

$$p_{\mathrm{MC}}(F_i|\mathbf{F}^q(\mathbf{r})) = \frac{p_{F_i} \phi(\mathbf{p}_{F_i}, \sigma_F; \mathbf{r})}{\sum_{i'} p_{F_{i'}} \phi(\mathbf{p}_{F_{i'}}, \sigma_F; \mathbf{r}) + \sum_{i'} p_B \phi(\mathbf{p}_{F_{i'}}, \sigma_B; \mathbf{r})}, \quad (9)$$

where p_{F_i} is the class prior of F_i. The background class probability can then be computed as $p_{\mathrm{MC}}(B|\mathbf{F}^q(\mathbf{r})) = 1 - \sum_{i'} p_{\mathrm{MC}}(F'_i|\mathbf{F}^q(\mathbf{r}))$.

Ignoring the background class, i.e., when $p_B = 0$, Eq. (9) corresponds to the normalized softmax [20]. This suggests that the model enables training to separate different foreground features into their respective prototype positions while pushing background features away from the prototypes. This is not possible in multi-class classification with ADNet++ [7], as its use of the max operation over foreground classes inherently makes it a piecewise binary classification model.

3.4 Targeting the Ideal Threshold

The foreground-background imbalance is a critical challenge in few-shot medical segmentation. While adaptive thresholds [4,15,24] attempt to tackle this with cross-entropy (CE), as illustrated in Fig. 2, minimizing CE does not necessarily yield the highest Dice score, showing that CE minimization does not directly optimize segmentation accuracy. Furthermore, using the *ideal* proportion—where the predicted foreground pixel counts match the true counts— achieves a near-optimal Dice score. Inspired by this and the significant performance gains observed in [2], we propose ideal thresholds of training query images as targets for learning adaptive thresholds. For brevity, we consider the single-prototype binary classification setting, which readily extends to the other settings.

First, we compute the upsampled Euclidean distance between features and prototypes. We then calculate the ideal distance threshold (IDT) T_D^* for each training query image. Mathematically, $T_D^* = \frac{D;\text{sort}_{|F|} + D;\text{sort}_{|F|+1}}{2}$, where D;sort is a vector of distances sorted in ascending order and $|F|$ is the foreground pixel count. By leveraging sorted distances, IDT ensures the number of predicted foreground pixels matches the true count $|F|$. To employ Eq. (3), we convert IDT into the corresponding ideal class prior (ICP) as $p_F^* = 1 - \text{sig}(-T_D^{*2}\left(\frac{1}{\sigma_F^2} - \frac{1}{\sigma_B^2}\right) - 2d\ln(\frac{\sigma_F}{\sigma_B}))$. Replacing p_F and the background prior with ICP p_F^*, in Eq. (3), shifts its decision boundary distance to IDT T_D^*.

4 Experiments

4.1 Setup

Dataset. Two public abdominal datasets are used for evaluation: ABD-MRI and ABD-CT. ABD-MRI [9] consists of 20 abdominal MRI scans with annotations for the liver, left kidney (L.K.), right kidney (R.K.), and spleen (Spl.). ABD-CT [10] contains 30 abdominal CT scans with the same organ annotations as ABD-MRI. For supervoxel generation, based on [6,7], we set the size parameter to 5000 for ABD-MRI and 2000 for ABD-CT.

Implementation Details. Following [6] for the binary-class and [7] for the multi-class setting, we conduct experiments using 5-fold cross-validation, performing three training runs per split, and leverage a single support image slice to segment the entire query volume and measure the mean Dice score. We initialize the feature extractor f_θ with a pre-trained ResNet-101 [8], using randomly transformed and reshaped 256×256 images, and optimize for 50k iterations using SGD with a learning rate of 10^{-3}. Features are 256-dimensional and normalized onto a spherical embedding, with $d = 1$ for simplicity. We run experiments in PyTorch (v1.9.0) on an NVIDIA RTX 3090 GPU.

Training and Threshold Options. To showcase the effectiveness of our approaches, we compare results across various options. For binary segmentation, we assess single-prototype-trained models using both single- and five-prototype evaluations. The baseline is standard ADNet [6], trained with the additional T loss, which regularizes T_S. In contrast, our TPM training excludes the T loss

Table 1. Binary segmentation. Best result in **bold**, second-best underlined. * denotes our contributions and OCP is the oracle result using test labels.

Proto.	Inference	ABD-MRI [9]					ABD-CT [10]				
		Liver	L.K.	R.K.	Spl.	Mean	Liver	L.K.	R.K.	Spl.	Mean
Single	ADNet[6]	74.86	59.41	84.27	59.54	69.52	56.18	34.82	37.56	49.14	44.42
	CE-T*	**75.31**	56.91	78.87	56.93	67.00	71.76	29.27	28.99	46.90	44.23
	AvgEst*	69.94	70.38	**85.99**	69.40	73.93	33.36	39.42	**41.49**	56.47	42.68
	LinEst*	73.98	**71.23**	85.64	69.11	74.99	65.34	38.92	40.88	58.49	50.91
	OCP	85.71	87.19	89.47	81.05	85.85	81.01	65.65	67.34	76.15	72.54
Multi (5)	CE-T*	74.85	54.92	77.60	56.75	66.03	**75.50**	28.48	28.79	46.06	44.71
	AvgEst*	73.60	68.08	85.82	69.65	74.29	46.26	39.97	40.49	58.71	46.36
	LinEst*	74.83	70.18	85.61	**70.18**	**75.20**	70.40	**40.08**	41.40	**58.96**	**52.71**
	OCP	85.39	86.00	89.00	81.04	85.36	82.96	64.14	66.00	76.14	72.31

and only optimizes the segmentation loss, as our proposed estimation schemes provide more principled thresholds. To assess the impact of aligning predicted and true pixel counts, we report TPM results using the CE-trained threshold (CE-T), estimated ideal class prior (ICP) values from training data—specifically, the average estimation (AvgEst) and a linear estimation (LinEst), with the latter based on the support foreground size $|F|$ and query slice location—as well as the oracle ideal class prior (OCP) for test queries. For multi-class segmentation, we compare results from single- and five-foreground training, using ADNet++ [7] as the baseline. While it is possible to consider each class count separately, we report LinEst estimated by matching the sum of all foreground class counts.

4.2 Results

Table 1 presents the binary segmentation results, confirming the merits of the ICP and multi-prototype (MP) approaches. In the single-prototype (SP) setting, while ADNet [6] tends to outperform CE-T (ADNet equivalent without T loss), AvgEst generally yields higher Dice scores than ADNet, except for the liver class, which has a relatively large foreground size. Moreover, LinEst further improves performance by incorporating size information, revealing ICP's benefit in this setting. The notably higher Dice scores of the OCP, which leverages oracle information, suggest significant room for improvement with advanced ICP estimation. Compared to SP predictions, the proposed MP extension further enhances the performance of the estimated ICPs. Table 2 presents the results for TPM's multi-class segmentation extension, demonstrating the efficacy of ICP and multi-foreground (MF) training. Focusing on single foreground (SF) training, TPM with estimated ICPs significantly outperforms the ADNet++ baseline [7]. Furthermore, MF consistently enhances TPM's performance over SF. This highlights the superior representation learning capability of the proposed extension of our TPM.

Table 2. Multi-class segmentation. * denotes our contributions.

Train	Inference	ABD-MRI [9]					ABD-CT [10]				
		Liver	L.K.	R.K.	Spl.	Mean	Liver	L.K.	R.K.	Spl.	Mean
SF	ADNet++ [7]	64.13	54.62	59.86	41.36	54.99	65.58	20.67	16.66	19.59	30.63
	LinEst*	65.50	69.43	70.72	47.15	63.20	61.97	31.77	24.46	**29.15**	36.84
MF (5)	ADNet++ [7]	57.22	45.85	54.57	29.93	46.89	61.14	15.43	12.00	12.88	25.36
	LinEst*	**71.91**	**71.00**	**73.85**	**49.94**	**66.67**	**66.84**	**37.90**	**27.55**	28.95	**40.31**

5 Conclusion

In this work, we introduce the tied prototype model (TPM), a probabilistic reformulation of ADNet [6] that advances its capabilities. Notably, TPM enables seamless extensions to multi-prototype and multi-class segmentation. Furthermore, we highlight the significance of ideal thresholds as target thresholds. Our experimental results demonstrate the performance improvements of each component, paving new research directions in prototype-based few-shot segmentation.

Acknowledgments. This work was supported by The Research Council of Norway (Visual Intelligence, grant no. 309439 as well as FRIPRO grant no. 315029 and IKT-PLUSS grant no. 303514).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Balasundaram, A., Kavitha, M.S., Pratheepan, Y., Akshat, D., Kaushik, M.V.: A foreground prototype-based one-shot segmentation of brain tumors. Diagnostics **13**(7), 1282 (2023)
2. Boudiaf, M., Kervadec, H., Masud, Z.I., Piantanida, P., Ben Ayed, I., Dolz, J.: Few-shot segmentation without meta-learning: a good transductive inference is all you need? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13979–13988 (2021)
3. Chen, X., et al.: A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiother. Oncol. **160**, 175–184 (2021)
4. Cheng, Z., Wang, S., Xin, T., Zhang, H.: Frequency-aware adaptive filtering network for few-shot medical image segmentation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1890–1897. IEEE (2024)
5. Cheng, Z., Wang, S., Xin, T., Zhou, T., Zhang, H., Shao, L.: Few-shot medical image segmentation via generating multiple representative descriptors. IEEE Trans. Med. Imaging (2024)

6. Hansen, S., Gautam, S., Jenssen, R., Kampffmeyer, M.: Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. Med. Image Anal. **78**, 102385 (2022)
7. Hansen, S., Gautam, S., Salahuddin, S.A., Kampffmeyer, M., Jenssen, R.: Adnet++: a few-shot learning framework for multi-class medical image volume segmentation with uncertainty-guided feature refinement. Med. Image Anal. **89**, 102870 (2023)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Kavur, A.E., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Med. Image Anal. **69**, 101950 (2021)
10. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proceedings of the MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge, vol. 5, p. 12 (2015)
11. Liu, Y., Zhang, X., Zhang, S., He, X.: Part-aware prototype network for few-shot semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 142–158. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_9
12. Oreshkin, B., Rodríguez López, P., Lacoste, A.: Tadam: task dependent adaptive metric for improved few-shot learning. Adv. Neural Inf. Process. Syst. **31** (2018)
13. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: training few-shot medical image segmentation without annotation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12374, pp. 762–780. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58526-6_45
14. Salahuddin, S.A., Hansen, S., Gautam, S., Kampffmeyer, M., Jenssen, R.: A self-guided anomaly detection-inspired few-shot segmentation network. CEUR Workshop Proceedings (2022)
15. Shen, Q., Li, Y., Jin, J., Liu, B.: Q-net: query-informed few-shot medical image segmentation. In: Arai, K. (eds.) Intelligent Systems and Applications. IntelliSys 2023. LNNS, vol. 823, pp. 610–628. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-47724-9_40
16. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Adv. Neural Inf. Process. Syst. **30**, 4077–4087 (2017)
17. Tian, Y., Liang, Y., Chen, Y., Zhang, J., Bian, H.: Multilevel support-assisted prototype optimization network for few-shot medical segmentation of lung lesions. Sci. Rep. **15**(1), 3290 (2025)
18. Tomasetti, L., et al.: Self-supervised few-shot learning for ischemic stroke lesion segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2023)
19. Tsochatzidis, L., Koutla, P., Costaridou, L., Pratikakis, I.: Integrating segmentation information into cnn for breast cancer diagnosis of mammographic masses. Comput. Methods Programs Biomed. **200**, 105913 (2021)
20. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: l2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 1041–1049 (2017)
21. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: few-shot image semantic segmentation with prototype alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9197–9206 (2019)

22. Yang, B., Liu, C., Li, B., Jiao, J., Ye, Q.: Prototype mixture models for few-shot semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12353, pp. 763–778. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58598-3_45
23. Zhao, Z., Gao, J., Luo, Z., Li, S.: Cpnet: cross prototype network for few-shot medical image segmentation. In: Lin, Z., et al. (eds.) Pattern Recognition and Computer Vision. PRCV 2024. LNCS, vol. 15045, pp. 135–149. Springer, Singapore (2025). https://doi.org/10.1007/978-981-97-8499-8_10
24. Zhu, Y., Cheng, Z., Wang, S., Zhang, H.: Learning de-biased prototypes for few-shot medical image segmentation. Pattern Recogn. Lett. **183**, 71–77 (2024)
25. Zhu, Y., Wang, S., Xin, T., Zhang, H.: Few-shot medical image segmentation via a region-enhanced prototypical transformer. In: Greenspan, H., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. MICCAI 2023. LNCS, vol. 14223, pp. 271–280. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-43901-8_26
26. Zhu, Y., Wang, S., Xin, T., Zhang, Z., Zhang, H.: Partition-a-medical-image: extracting multiple representative sub-regions for few-shot medical image segmentation. IEEE Trans. Instrum. Meas. (2024)



Towards Accurate Tumor Budding Detection: A Benchmark Dataset and A Detection Approach Based on Implicit Annotation Standardization and Positive-Negative Feature Coupling

Rui-Qing Sun[1], Zeng Fan[1], Boyang Dai[1], Yiyan Su[2], Qun Hao[1,3(✉)], Chuyang Ye[1(✉)], and Shaohui Zhang[1(✉)]

[1] Beijing Institute of Technology, 100081 Beijing, China
{qhao,chuyang.ye,zhangshaohui}@bit.edu.cn
[2] Shanxi Medical University, 030001 Taiyuan, China
[3] Changchun University of Science and Technology, 130022 Changchun, China

Abstract. The detection of tumor budding on histopathological images provides vital information for treatment planning and prognosis prediction. As manual identification of tumor budding is labor-intensive, automated tumor budding detection is desired. However, unlike other tumor cell detection tasks, tumor budding involves clusters of multiple tumor cells, which is more likely to be confused with other clusters of cells with similar appearances. It becomes challenging for existing cell detection methods to discriminate tumor budding from other cells. Additionally, the lack of public datasets for tumor budding detection hinders further development of accurate tumor budding detection methods. To address these challenges, to the best of our knowledge, we introduce the first publicly available benchmark dataset for tumor budding detection. The dataset consists of 410 images with H&E staining and the corresponding bounding box annotations of 3,968 cases of tumor budding made by experts. Moreover, based on this dataset, we propose a designated approach *Tumor Budding Detection Network* (TBDNet) for tumor budding detection with improved detection performance. On top of standard objection detection backbones, we develop two major components in TBDNet, *Iteratively Distilled Annotation Relocation* (IDAR) and *Rotational Feature Decoupling And Recoupling* (RFDAR). First, as different experts have different standards for budding boundaries in the annotation, the detection model may receive inconsistent knowledge during model training. Therefore, we introduce the IDAR module that implicitly standardizes the annotations. IDAR relocates the annotations via iterative model distillation so that the relocated annotations are consistent for training the detection model. Second, to reduce the interference from cells with similar features, i.e., negative samples, to tumor budding, i.e., positive samples, we develop the RFDAR module. RFDAR enhances feature extraction via positive-negative feature

R.-Q. Sun and Z. Fan—These authors contributed equally.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 662–672, 2026.
https://doi.org/10.1007/978-3-032-04947-6_63

coupling regularized by prior feature distributions, so that it is better capable of distinguishing tumor budding. The results on the benchmark show that our approach outperforms state-of-the-art detection methods by a noticeable margin. All code and data are available at https://github.com/J-F-AN/TumorBuddingDetection.

Keywords: tumor budding detection · computational pathology · benchmark dataset

1 Introduction

Tumor budding refers to the phenomenon where tumor cells detach from the main tumor mass and infiltrate surrounding tissues in small clusters [6]. It has been identified as a key indicator of tumor invasiveness and metastatic potential. The accurate quantification of tumor budding can inform effective treatment, which prolongs patient survival and improves treatment outcome [17]. Traditional quantitative analysis of tumor budding relies on manual detection by pathologists in pathological images, a labor-intensive process. Therefore, there is a pressing need for automated tumor budding detection. Although tumor budding detection has not been extensively studied before, it can be formulated as an object detection problem. With the development of object detection technologies based on *deep learning* (DL) [12,13,20], DL detection models, including generic ones [10,11,18,19,23] or those designed for tumor cell detection [1,2,16], can be applied to tumor budding detection. For example, Bokhorst et al. [2] use a general DL detection model to automatically detect tumor budding in colorectal cancer. Their subsequent work [1] trains the network using a semi-supervised approach, which further improves detection accuracy under data-limited conditions. Piansaddhayanaon et al. [16] propose a module for tumor cell detection aimed at enhancing the performance of existing two-stage detection frameworks, and it can be used for budding detection as well. However, tumor budding detection is generally more challenging than other tumor cell detection tasks, as budding typically presents as clusters of tumor cells, which are prone to being mistaken for other cells or cell clusters with similar morphological characteristics. Therefore, direct application of existing detection methods to tumor budding only achieves suboptimal performance. Moreover, there is currently no publicly available dataset for tumor budding detection, which further hinders the development of accurate budding detection methods.

To address these challenges, we introduce the first publicly available benchmark for tumor budding detection, and it is named *Tumor Budding Detection Dataset* (TBDD). TBDD contains 410 images with H&E staining and 3,968 annotated cases of tumor budding in the form of bounding boxes. Based on this benchmark dataset, we also propose *Tumor Budding Detection Network* (TBDNet), which is a designated approach to tumor budding detection with improved detection performance. There are two major contributions, *Iteratively Distilled Annotation Relocation* (IDAR) and *Rotational Feature Decoupling And Recoupling* (RFDAR) in TBDNet. First, due to the irregular shape of tumor budding,

experts tend to have different standards when determining budding boundaries for annotation. Such a discrepancy can lead to inconsistent knowledge learning during model training, which adversely affects the detection performance. To tackle this issue, we introduce IDAR that implicitly standardizes the bounding box annotations. IDAR uses a teacher-student framework to relocate the boxes, and with iterative model distillation the final student model produces consistently relocated annotations for subsequent model training. Second, to better distinguish between tumor budding, i.e., positive samples, and non-budding targets, i.e., negative samples, with similar features, RFDAR enhances the feature extraction by decoupling and recoupling the features of positive and negative samples, under the regularization of prior feature distributions. Such positive-negative feature coupling encourages the model to focus on the discriminative features for distinguishing tumor budding and non-budding targets. Qualitative and quantitative evaluation results on the benchmark dataset TBDD show that TBDNet outperforms existing detection methods.

2 Method

The contribution of this work is summarized in Fig. 1, which gives an overview of the data curation for TBDD and the proposed TBDNet. Their detailed descriptions are given below.

2.1 Dataset Curation for Tumor Budding Detection

TBDD was collected for colorectal cancer patients, as colorectal cancer is a leading cause of cancer-related morbidity and mortality [22]. Histopathology images were acquired for 410 patients over two years on a KFBIO KF-PRO-400-HI scanner. The samples were processed, sectioned, and stained with H&E; then they were examined with 40x magnification to capture clear, high-resolution whole slide images. For each patient, only an 800×800 patch with tumor budding that is clinically informative was cropped from the whole slide image and annotated. Three experienced pathologists annotated tumor budding areas on the 800×800 images with bounding boxes, each handling a disjoint subset of the images, and all annotations were cross-reviewed according to the criteria of the International Tumor Budding Consensus Conference (ITBCC) to ensure consistency. The released dataset contains both the cropped images and their annotations.

2.2 Iteratively Distilled Annotation Relocation

As tumor budding typically has variable shapes depending on the composition of tumor cells, different experts can determine the sizes and locations of annotation bounding boxes with different standards [9]. This annotation inconsistency hinders effective training of detection models [21]. To address this issue, we propose the IDAR method, which gradually relocates the annotation through iterative

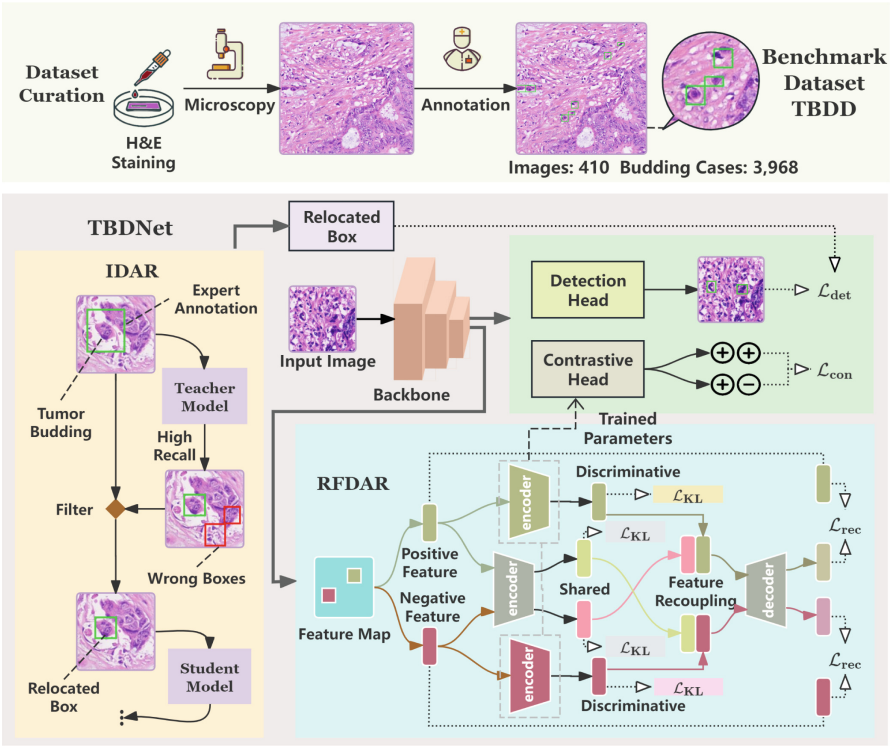


Fig. 1. An overview of the proposed work. First, we introduce a benchmark dataset TBDD for tumor budding detection. Then, we develop TBDNet for tumor budding detection based on TBDD. TBDNet comprises two major components, IDAR and RFDAR. IDAR implicitly standardizes the annotation to avoid inconsistent knowledge learning, whereas RFDAR uses positive-negative feature coupling to enhance the extraction of discriminative features for distinguishing tumor budding from other targets.

knowledge distillation [7] and produces consistent standardized bounding boxes better suited for model learning.

In IDAR, we first train a teacher model with the manually annotated data. The bounding boxes generated by the teacher model are then used as soft labels to train the student model. In the next iteration, the student model serves as the new teacher model to train a new student model. With the iterative distillation, the size and location of the bounding boxes stabilize, which implicitly achieves standardized annotation that both conforms to the original annotation and becomes consistent across samples. Notably, during each iteration, when generating soft-label bounding boxes, to ensure all annotated budding cases are preserved, the teacher model predicts with a low confidence threshold of 0.05. As false positives may be produced with the low threshold, the teacher outputs

are then filtered based on the manual annotations with an *Intersection over Union* (IoU) threshold of 0.2.

2.3 Rotational Feature Decoupling and Recoupling

Then, we train the detection model with the standardized annotation. The tumor budding cell clusters can be visually similar to other cell clusters, where their cytoplasm has similar appearances and their nucleus appearances are different. Such subtle differences can be difficult to learn for existing detection frameworks. Therefore, we design RFDAR to address the issue.

Our method is based on a standard object detection backbone. Here, we choose YOLOv5 [10] due to its ease of deployment and high performance, but other backbones are also applicable. RFDAR uses a contrastive learning strategy by adding a contrastive head to YOLOv5 [10] (see Fig. 1), which encourages discrimination between the positive samples of tumor budding and the negative samples of similar non-tumor-budding cell clusters. However, the features extracted by the backbone network contain both discriminative ones and those that are shared between tumor budding and other cell clusters. This is possibly because the nucleus components are different and the cytoplasm components are similar between positive and negative samples, and naive training of the detection model cannot optimally preserve the discriminative features. Therefore, in conventional contrastive learning, the separation of discriminative features of positive and negative samples is hindered by the shared features.

To address this problem, RFDAR seeks to decompose the features into discriminative ones and shared ones, and the decomposition is learned by coupling the features of positive and negative samples. Then, only the discriminative features are used in contrastive learning to better distinguish positive and negative samples. Specifically, during model training the backbone features are first categorized as features of positive samples and negative samples based on the annotation. As shown in Fig. 1, both positive and negative features are fed into an encoder of discriminative features and an encoder of shared features for feature decoupling. The shared features are then rotated between the positive and negative samples, i.e., the positive shared features are recoupled with the negative discriminative features and the negative shared features are recoupled with the positive discriminative features. Since the shared features are similar between positive and negative samples, the original features should be reconstructed from the recoupled features.

$$\mathcal{L}_{\text{rec}} = \frac{1}{2}\left(\|\mathbf{x}_{\text{orig}} - \mathbf{x}^{(1)}_{\text{recouple}}\|_2^2 + \|\mathbf{x}_{\text{orig}} - \mathbf{x}^{(2)}_{\text{recouple}}\|_2^2\right), \tag{1}$$

Therefore, by minimizing a reconstruction loss $\mathcal{L}_{\text{rec}}$ measured by the mean squared error of the positive and negative features, RFDAR learns how to decompose the raw features into discriminative and shared ones.

To avoid degeneration and improve the stability of the decoupling, we introduce additional regularization for training RFDAR based on the prior feature

distributions. Specifically, from the raw features, we fit the distributions of positive and negative features with a *Gaussian mixture model* (GMM) [4]. As the raw features are high-dimensional, before model fitting we apply *principal component analysis* (PCA) [8] to the features for dimensionality reduction. Then, the first principal component that best reflects the major variation in the samples is used for fitting the GMM. We assume that the GMM comprises three components, the discriminative features of positive samples, the discriminative features of negative samples, and the shared features of the positive and negative samples. With the three distributions, the *KullbackLeibler* (KL) divergence loss $\mathcal{L}_{\text{KL}}$ is computed as regularization for the discriminative feature of positive samples, the discriminative feature of negative samples, and the shared feature with respect to their corresponding prior distributions (see Fig. 1).

The encoders of discriminative and shared features are trained by minimizing the sum of $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{KL}}$ with the other network parameters fixed.

Finally, with the trained RFDAR module (frozen), we perform the training of the detection model aided by contrastive learning. Specifically, the standard detection loss $\mathcal{L}_{\text{det}}$ of YOLOv5 [10] is combined with the following contrastive loss motivated by [3] and [25]:

$$\mathcal{L}_{\text{con}} = \frac{1}{N} \sum_{i=1}^{N} \left[(1 - y_i) \cdot (d_i)^2 + y_i \cdot (\max(0, m - d_i))^2 \right], \tag{2}$$

where y_i denotes the label disagreement ($y_i = 0$ for two samples of the budding class and $y_i = 1$ for samples of different classes) of the i-th sample pair, d_i represents the Euclidean distance between the sample pair, N is the total number of sample pairs, and m is a predefined hyperparameter representing the maximum valid distance between inter-class pairs. Note that the features are normalized by their L2 norm before computing d_i, and we set $m = 1.2$ based on experiments on a validation set [5,24]. Then, the total loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda \cdot \mathcal{L}_{\text{con}}, \tag{3}$$

where λ is a balancing hyperparameter set to 1.5 based on the validation set.

2.4 Implementation Details

Both teacher and student models in IDAR use the YOLOv5 detector as the backbone, with the iteration count empirically fixed to two based on experimental tuning. Each encoder in RFDAR comprises two fully connected layers. It maps the input to a 256-dimensional space with ReLU activation [15], followed by two parallel linear layers that generate the mean and log variance with a dimension of 128. The decoder in RFDAR also comprises two fully connected layers, where the first layer has a channel width of 256 with ReLU activation and the second one projects the output of the first layer back to the original feature space. The contrastive head uses the encoder trained in RFDAR for computing the

contrastive loss. All these auxiliary modules are employed solely during training, and at inference time the model collapses to the vanilla YOLOv5 detector, incurring no additional parameters or latency.

Model training is performed with the default settings of YOLOv5 [10] when applicable. The images are cropped into 384×384 patches with an overlap of 128 pixels in each dimension as model input. The predictions on the input patches are merged for test images.

Table 1. Detection performance of different methods on TBDD. The best results are highlighted in bold.

Category	Method	F1 (%) ↑	P (%) ↑	R (%) ↑
CNN-based (two-stage)	Faster R-CNN (2016) [18]	59.11	60.61	57.69
	RetinaNet (2017) [19]	58.06	55.75	60.58
	ReCasNet (2023) [16]	53.26	56.98	49.99
CNN-based (one-stage)	YOLO11 (2024) [11]	57.07	58.97	55.29
	MambaYOLO (2024) [23]	56.46	58.80	54.30
Transformer-based	SwinT (2021) [14]	40.48	45.11	36.71
	Deformable DETR (2021) [26]	45.37	41.67	49.79
Ours	**TBDNet**	**63.29**	**63.59**	**62.98**

3 Experiments

3.1 Experimental Setup and Detection Performance

The 410 images in TBDD were split into a training set of 276 images, a validation set of 94 images, and a test set of 40 images. We compared TBDNet with several representative object detection methods, including CNN-based two-stage detectors Faster R-CNN [18], RetinaNet [19], and ReCasNet [16], CNN-based one-stage detectors YOLO11 [11] and MambaYOLO [23], and Transformer-based approaches Deformable DETR [26] and SwinT [14]. All these methods applied their default settings and used the same experimental settings as TBDNet for fair comparison.

The detection performance was quantitatively evaluated with the F1-score (F1), precision (P), and recall (R). The results are presented in Table 1. TBDNet achieves better performance than all competing methods. The CNN-based two-stage detectors tend to perform better among the competing methods. Compared with the second best method Faster R-CNN, TBDNet improves the F1-score by 4.18%. Note that our method is based on YOLOv5, and it outperforms (by over 6%) more recent versions of YOLO detectors, YOLO11 and MambaYOLO, whereas the baseline YOLOv5 has similar performance to the two competitors (see the ablation study later in Table 2).

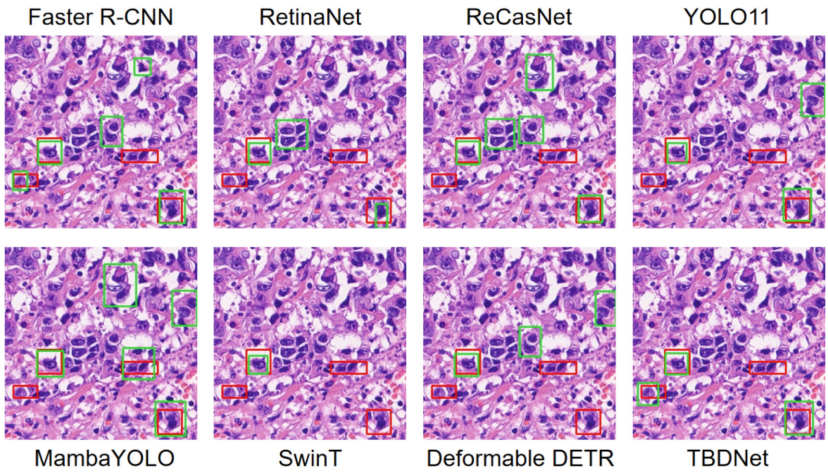


Fig. 2. Visualization examples of detection results. The green box is the output of the model, and the red box is the annotation. (Color figure online)

Visualization examples of the detection results are shown in Fig. 2. Compared with the competing methods, TBDNet produces more true positive and fewer false positive budding detection results. This observation agrees with the better precision and recall values of TBDNet in Table 1.

3.2 Ablation Study

To further validate the effectiveness of each module in our method, we performed an ablation study. The results are shown in Table 2. First, compared to the YOLOv5 baseline model, one iteration of IDAR improves the F1-score, and it is further increased after the second iteration. This confirms the benefit of IDAR. Second, the contrastive head itself improves the F1-score after the application of IDAR. When the full RFDAR is also applied (with both $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{KL}}$), the best performance is achieved, which shows the benefit of RFDAR. Note that if only $\mathcal{L}_{\text{rec}}$ or $\mathcal{L}_{\text{KL}}$ is used in RFDAR, the results are worse than the use of the contrastive head, because $\mathcal{L}_{\text{rec}}$ mainly increases recall by encouraging the detection of positive cases, while $\mathcal{L}_{\text{KL}}$ tightens the posterior distribution to improve precision. Using either loss alone therefore fails to balance sensitivity and specificity.

Table 2. Ablation study with different configurations. The baseline uses YOLOv5. Iter represents the number of iterations in IDAR. C-Head represents the use of the contrastive head. The best results are highlighted in bold.

Baseline	Iter	C-Head	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{KL}}$	F1 (%) ↑	P (%) ↑	R (%) ↑
✓	–	–	–	–	57.42	59.18	55.77
✓	1	–	–	–	58.63	57.67	59.62
✓	2	–	–	–	60.77	60.48	61.06
✓	2	✓	–	–	61.23	62.94	59.62
✓	2	✓	✓	–	59.36	56.52	62.50
✓	2	✓	–	✓	57.00	57.28	56.73
✓	2	✓	✓	✓	**63.29**	**63.59**	**62.98**

4 Conclusion

We have curated and released a tumor budding detection dataset TBDD and proposed TBDNet for better performance of tumor budding detection. TBDD comprises 410 H&E images with 3,968 cases of annotated tumor budding. In TBDNet, we have made two major contributions IDAR and RFDAR. IDAR implicitly standardizes the annotation by relocating the annotation bounding boxes via iterative model distillation. This avoids knowledge inconsistency during model training due to the different standards of different experts for budding boundaries. RFDAR better distinguishes tumor budding from other similar cells or cell clusters by extracting more discriminative features via positive-negative feature coupling. Experimental results show that TBDNet outperforms existing methods for the challenging tumor budding detection task by a noticeable margin.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62275020).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bokhorst, J.M., et al.: Semi-supervised learning to automate tumor bud detection in cytokeratin-stained whole-slide images of colorectal cancer. Cancers **15**(7), 2079 (2023)
2. Bokhorst, J.M., Rijstenberg, L., Goudkade, D., Nagtegaal, I., van der Laak, J., Ciompi, F.: Automatic detection of tumor budding in colorectal carcinoma with deep learning. In: Stoyanov, D., et al. (eds.) Computational Pathology and Ophthalmic Medical Image Analysis. OMIA COMPAY 2018 2018. LNCS, vol. 11039, pp. 130–138. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00949-6_16

3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 539–546. IEEE (2005)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc.: Ser. B (Methodol.) **39**(1), 1–22 (1977)
5. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 1735–1742. IEEE (2006)
6. Hase, K., Shatney, C., Johnson, D., Trollope, M., Vierra, M.: Prognostic value of tumor "budding" in patients with colorectal cancer. Diseases Colon Rectum **36**(7), 627–635 (1993)
7. Hinton, G.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
8. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**(6), 417 (1933)
9. Jiao, L., et al.: A survey of deep learning-based object detection. IEEE Access **7**, 128837–128868 (2019)
10. Jocher, G.: Ultralytics YOLOv5 (2020). https://doi.org/10.5281/zenodo.3908559, https://github.com/ultralytics/yolov5
11. Jocher, G., Qiu, J.: Ultralytics YOLO11 (2024). https://github.com/ultralytics/ultralytics
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25** (2012)
13. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
15. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)
16. Piansaddhayanaon, C., Santisukwongchote, S., Shuangshoti, S., Tao, Q., Sriswasdi, S., Chuangsuwanich, E.: ReCasNet: improving consistency within the two-stage mitosis detection framework. Artif. Intell. Med. **135**, 102462 (2023)
17. Prall, F., Nizze, H., Barten, M.: Tumour budding as prognostic factor in stage I/II colorectal carcinoma. Histopathology **47**(1), 17–24 (2005)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)
19. Ross, T.Y., Dollár, G.: Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2980–2988 (2017)
20. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
21. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: a survey. IEEE Trans. Neural Netw. Learn. Syst. **34**(11), 8135–8153 (2022)
22. Sung, H., et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J. Clin. **71**(3), 209–249 (2021)

23. Wang, Z., Li, C., Xu, H., Zhu, X.: Mamba YOLO: SSMs-Based YOLO for object detection. arXiv preprint arXiv:2406.05835 (2024)
24. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
25. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9299–9306 (2019)
26. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2020)



Unisyn: A Generative Foundation Model for Universal Medical Image Synthesis Across MRI, CT and PET

Yulin Wang[1], Honglin Xiong[1], Kaicong Sun[1], Jiameng Liu[1], Xin Lin[1], Ziyi Chen[1], Yuanzhe He[1], Qian Wang[1,2], and Dinggang Shen[1,2,3(✉)]

[1] School of Biomedical Engineering and State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, 201210 Shanghai, China
[2] Shanghai Clinical Research and Trial Center, 201210 Shanghai, China
[3] Shanghai United Imaging Intelligence Co., Ltd., 200230 Shanghai, China
dgshen@shanghaitech.edu.cn

Abstract. Multi-modal brain imaging with MRI, CT, and PET has significantly advanced our understanding of cognition and neurodisease by providing complementary information. However, constraints on scan time and cost often result in missing critical high-quality sequences. Existing cross-modality synthesis methods are typically task- or modality-specific, leading to performance degradation when applied to heterogeneous real-world imaging data. Here, we propose UniSyn, a unified framework capable of synthesizing target imaging modalities with specific acquisition parameters from any available ones, guided by metadata. UniSyn first learns robust metadata representations through image-text alignment on large-scale multimodal neuroimaging datasets. We then introduce a cross-modality synthesis framework that leverages learned metadata representations to guide the generation of metadata-specified target images. To enhance interpretable metadata-driven control over image synthesis across diverse protocols, we design a dual-parameter arithmetic operation that explicitly integrates source and target metadata into the image translation process. Extensive experiments on multi-institutional brain imaging datasets demonstrate that UniSyn surpasses the existing cross-modality synthesis approaches in both quantitative fidelity and clinical relevance, enabling the generation of missing imaging counterparts tailored to specific clinical and research needs.

Keywords: Multi-modal medical image · Brain image synthesis · Vision-language model · Metadata-injecting prompt

1 Introduction

Neuroimaging is essential for diagnosing and studying neurological disorders. Different modalites such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Positron Emission Tomography (PET) offer complementary insights into brain structure, function, and metabolism. For instance,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 673–682, 2026.
https://doi.org/10.1007/978-3-032-04947-6_64

MRI provides high soft-tissue contrast without ionizing radiation, CT is optimal for detecting hemorrhages or calcifications, and PET enables metabolic assessments crucial for neurodegenerative diseases. However, acquiring multiple modalities for every patient is often impractical due to financial constraints, radiation exposure, and scanning duration limitations. The absence of a required modality can impair diagnostic accuracy and research applications, highlighting the need for a robust cross-modality synthesis framework capable of generating missing imaging counterparts.

Existing cross-modality synthesis methods predominantly focus on specific translation tasks, such as synthesizing PET from MRI [1,2], CT from MRI [3], or inter-sequence generation [4–6], isotope tracer conversion [7], and dose modulation within the same modality [8,9]. However, these approaches suffer from key limitations: (1) They lack generalizability across diverse imaging modalities and are constrained to fixed source-target pairs, limiting their applicability in real-world clinical scenarios. (2) Most methods operate in a purely image-driven manner, neglecting valuable non-imaging metadata such as clinical reports and imaging protocols that could enhance synthesis quality and interpretability. Previous work [10] leverages textual imaging parameters of target modality to synthesize desired MRI sequence from available ones, albeit its satisfactory synthesis performance, it is limited to inter-conversion between MRI sequences and relies on cross-attention mechanisms [11] that are sensitive to hyperparameter configurations, limiting robustness and interpretability of feature space manipulation. To address these challenges, we propose UniSyn, a unified generative framework that enables universal cross-modality brain image synthesis, guided by source and target textual metadata. UniSyn consists of two key components: (1) a metadata prompt learning network, NeuroCLIP, pre-trained on large-scale multimodal datasets, to learn a textual representation aligned with neuroimaging data, and (2) a cross-modality synthesis network, which generates anatomically and contrast-preserving target images guided by learned text embeddings. By explicitly disentangling modality-specific features from input images and integrating them with target text embeddings, UniSyn achieves improved generalization and synthesis fidelity. Extensive experiments on multi-institutional brain imaging datasets demonstrate that UniSyn outperforms existing methods in both quantitative synthesis fidelity and qualitative clinical relevance, representing a significant step toward a generalized medical image translation paradigm with broad applications in neuroimaging-based diagnosis and research.

2 Method

The overall framework of Unisyn is shown in Fig. 1. Specifically, we achieve text-guided image synthesis through two stages: the metadata prompt learning network (NeuroCLIP) and the generative foundation model (GFM).

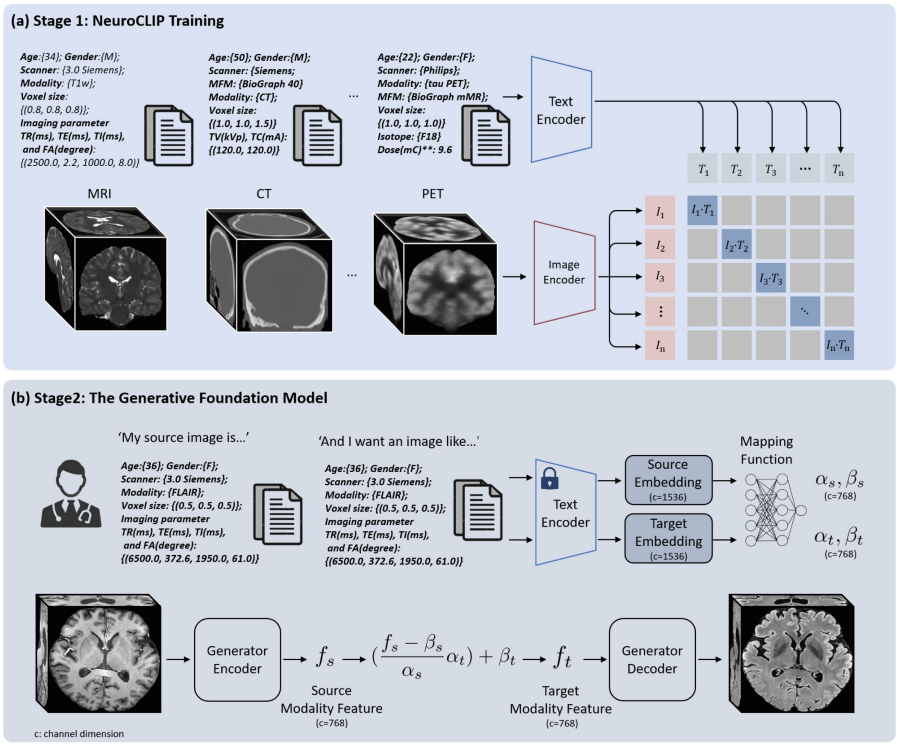


Fig. 1. Overview of the Unisyn framework: (a) The training of NeuroCLIP model; (b) The generative foundation model

2.1 NeuroCLIP

Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in understanding and processing natural language across general domains. Models like GPT-4 [12] and BERT [13] have achieved unprecedented performance in various natural language processing tasks. However, these general-purpose LLMs often struggle with domain-specific technical language, particularly in highly specialized medical fields.

Even domain-adapted models like BiomedCLIP [14], which are specifically trained on biomedical data, show limitations in comprehending specialized neuroimaging parameters such as TR (Repetition Time), TE (Echo Time), and flip angle. This gap in technical comprehension poses a significant challenge for accurate medical image synthesis and analysis.

Therefore, we propose NeuroCLIP, a specialized text encoder designed specifically for neuroimaging metadata comprehension. In Stage 1, to extract effective semantic information from textual imaging parameters and demographic attributes associated with each image, we leverage the contrastive learning

strategy to pretrain a text encoder that is specifically designed for 3D brain medical images' metadata understanding.

Construction of Image and Text Prompt Dataset. The appearance of brain medical images exhibits a strong correlation with both biological attributes (e.g., age and gender) and imaging parameters. To effectively leverage these factors as guidance of GFM, we prioritize imaging parameters that most influence image intensity and contrast, and design structured textual prompts for different imaging modalities under the guidance of senior radiologists. Image samples along with their corresponding textual descriptions used for training the NeuroCLIP are presented in Fig. 1(a).

When specific imaging parameters are unavailable, "None" is used as a placeholder to maintain a consistent prompt structure. However, source and target modality information is always explicitly provided to ensure accurate modality translation.

Contrastive Learning Pre-training. To enable text representations to effectively guide cross-modality image synthesis and ensure that the generated images accurately match the specified parameters, it is essential for the text encoder to comprehensively capture the relevant imaging metadata. To achieve this, the image and text encoders are jointly trained on paired image-text samples using the contrastive learning strategy. We categorize the dataset based on imaging parameters and set the batch size accordingly. During each forward pass, a batch B consists of one sample from each class, and the embedding distance between paired image-text samples is minimized, while the distance between non-paired samples is maximized by computing contrastive loss.

The model architecture is built upon [10], with the encoded token length extended from 90 to 224, thereby improving the text encoder's ability to process longer text prompts and support diverse downstream tasks.

Upon completion of the Stage 1 training, the pretrained text encoder is frozen and subsequently utilized in Stage 2.

2.2 The Generative Foundation Model

Stage 2 focuses on generating the target image I_t from any available source image I_s by leveraging both source T_s and target text prompts T_t as guidance. As illustrated in Fig. 1(b), the synthesis process begins by encoding I_s, T_s, and T_t into corresponding feature embeddings, T_s and T_t are then used to guide the image feature transformation process to obtain the target image feature f_t. Finally, f_t is reconstructed to the target image $\hat{I}_t$ by the image decoder.

Text-Guided Representation Learning. The representation learning process begins with encoding source text prompts into feature vectors. Inspired by [15], we decompose medical image volumes into two fundamental components:

content features that capture anatomical structures, and contrast features that encode modality-specific characteristics.

To effectively model the modality-specific contrast characteristics, we introduce a novel mapping mechanism that transforms text features into two key parameters: a scaling factor (α) and a bias term (β). Specifically, the text prompt is first encoded into a 1536-dimensional feature representation, where the first 768 dimensions are assigned to α and the remaining 768 dimensions to β. These parameters are designed to capture the essential aspects of cross-modality variations: α models contrast differences between imaging modalities, while β accounts for baseline intensity shifts arising from diverse acquisition protocols.

Our framework implements a systematic transformation pipeline for cross-modality synthesis: First, we normalize the source image feature f_s (768 dimensions) to obtain f_0 by applying: $f_0 = (f_s - \beta_s)/\alpha_s$. This normalization step isolates the modality-independent anatomical representation. Subsequently, we synthesize the target feature f_t by incorporating the target modality characteristics: $f_t = \alpha_t f_0 + \beta_t$, where α_t and β_t are the scaling and bias parameters specific to the target modality. Finally, the transformed feature f_t is processed through a decoder to generate the synthesized target image $\hat{I}_t$. This formulation enables explicit and interpretable modulation of the image feature space, facilitating effective cross-modality synthesis while preserving anatomical integrity.

Model Architecture

Image Encoder: The image encoder consists of 24 residual blocks (ResBlocks), with residual connections between each block to facilitate gradient propagation. Each ResBlock consists of two sequential 3×3 convolutional layers, followed by a ReLU activation layer. The encoder is designed with a channel dimension of 1024, ensuring sufficient capacity for feature extraction.

Image Decoder: The decoder is a four-layer fully connected network with ReLU activation. The dimensions of each layer are 2048, 1024, 1024, and 8, respectively. The final layer with 8 output channels matches the input channel number of the image encoder, facilitating the image synthesis from processed feature maps.

Loss Functions. To strengthen the constraint on general feature representation learning, we incorporate a cyclic synthesis strategy during training. Specifically, within each iteration, the source image and target image are utilized to generate each other, and their encoded features F_x and F_y are enforced to be consistent. We impose a pixel-level similarity as the supervisory signal.

$$\mathcal{L}_f = \sum_i \|F_x - F_y\|_1 \tag{1}$$

We also introduce a synthesis loss as a supervisory signal for cross-contrast translation by directly comparing the synthesized images with their corresponding

ground truth counterparts. The synthesis loss is then computed as the L1-norm of the residual error:

$$\mathcal{L}_s = \sum_i \|\hat{X}_i - X_i\|_1 \tag{2}$$

Combining the two loss terms above, we derive the overall loss as:

$$\mathcal{L}_{total} = \lambda_f \mathcal{L}_f + \lambda_s \mathcal{L}_s \tag{3}$$

3 Experiments and Results

3.1 Datasets and Data Preprocessing

Description of the GFM Dataset

Huashan Dataset. The Huashan (HS) dataset consists of 1,002 participants for the Alzheimer's Disease Study, recruited from both the Universal Medical clinic and Shanghai Sixth People's Hospital, China. From this cohort, we selected a subset of 847 individuals who all underwent T1-weighted (T1w) MRI scans. Among these participants, the availability of additional imaging modalities—CT scans, amyloid-beta (Aβ)-PET scans, Fluorodeoxyglucose (FDG)-PET scans, and other MRI sequences, including T2-weighted (T2w) and Fluid-Attenuated Inversion Recovery (FLAIR)—varied on a per-subject basis.

Zhongshan Dataset. The Zhongshan (ZS) dataset comprises 114 participants from Zhongshan Hospital, Fudan University, enrolled for the neuro-degradation study. Each subject includes T1w, T2w, and FLAIR MRI sequences, CT, and FDG-PET scans acquired using United Imaging Healthcare (UIH) scanning systems.

Description of the NeuroCLIP Dataset. For NeuroCLIP pretraining, except for the above two datasets, we use a large-scale dataset containing 47,841 3D scans collected from multiple centers, covering various imaging parameters. These datasets include T1w, T2w, FLAIR, PD, SWI, T2Star, T1CE MRI sequences, PET imaging with FDG, AV45, and TAU deposition, and CT.

For image preprocessing, all imaging modalities are spatially registered to the corresponding subject's T1w image, preserving its original spatial resolution. Subsequently, skull stripping is performed on all images. The dataset is then split into training, validation, and test sets using a 7:1:2 ratio.

3.2 Implementation Details

In our experiments, both the NeuroCLIP and GFM are trained on a single NVIDIA A100 GPU with 40 GB memory. For NeuroCLIP, both image and text encoders are trained from scratch for 100 epochs using the Adam optimizer [16]. For the GFM, its image encoder and decoder are trained from scratch, with an input and output patch size of $8 \times 64 \times 64$. GFM is trained for 300 epochs using

the Adam optimizer, with a batch size of 8, an initial learning rate of 10^{-4}, and a weight decay by 0.5 every 100 epochs. We empirically set $\lambda_f = 0.1$, $\lambda_s = 1.0$ in our experiments. To comprehensively evaluate model performance on image synthesis, we employed Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) as quantitative metrics.

Table 1. Quantitative comparison across different tasks.

		ZS T1-CT	ZS CT-PET	ZS T1-PET	HS T1-CT	HS T1-FLAIR	HS T1-PET
PTNet	PSNR	25.78 ± 1.19	24.63 ± 0.38	25.89 ± 0.51	24.91 ± 2.22	27.21 ± 3.01	23.89 ± 2.32
	SSIM	0.918 ± 0.033	0.885 ± 0.020	0.906 ± 0.027	0.905 ± 0.027	0.940 ± 0.048	0.889 ± 0.034
BME-X	PSNR	24.81 ± 1.21	23.56 ± 0.51	24.71 ± 0.64	23.85 ± 2.44	25.83 ± 3.13	22.53 ± 2.37
	SSIM	0.902 ± 0.030	0.872 ± 0.026	0.896 ± 0.029	0.894 ± 0.040	0.933 ± 0.049	0.871 ± 0.037
TUMSyn	PSNR	27.56 ± 1.29	24.90 ± 0.49	26.98 ± 0.43	26.33 ± 2.28	28.14 ± 3.07	24.78 ± 2.18
	SSIM	0.930 ± 0.034	0.893 ± 0.021	0.913 ± 0.027	0.915 ± 0.033	0.958 ± 0.042	0.894 ± 0.028
Unisyn	PSNR	**28.04 ± 1.36**	**25.75 ± 0.53**	**27.45 ± 0.79**	**26.78 ± 2.32**	**29.05 ± 2.99**	**25.47 ± 2.14**
	SSIM	**0.941 ± 0.034**	**0.909 ± 0.028**	**0.926 ± 0.032**	**0.927 ± 0.035**	**0.965 ± 0.039**	**0.907 ± 0.027**

3.3 Comparison with State-of-the-Art Methods

We conducted comprehensive experiments on the HS and ZS datasets to evaluate the effectiveness of our framework against three competitive state-of-the-art (SOTA) medical image synthesis models: 1) PTNet [5], an advanced 3D transformer-based model for infant brain MRI sequences synthesis; 2) BME-X [17], a unified model for the motion correction, resolution enhancement, denoising and harmonization of MR images; and 3) TUMSyn [10], which utilizes textual metadata as prompt for any MRI sequences synthesis. All methods are in their default configuration.

Quantitative Comparison Analysis. Table 1 presents the quantitative evaluation of all methods across six clinically significant tasks. For PET image synthesis tasks, we synthesized FDG-PET images for the ZS dataset and AV45-PET images for the HS dataset. Our proposed framework consistently achieves the highest PSNR and SSIM scores across all tasks, outperforming three SOTA methods. This performance enhancement highlights that explicit integration of both source and target textual metadata, along with the joint training strategy, can collectively empower the model to capture robust shared features across varying imaging parameters, thereby mitigating noise and artifacts introduced during cross-modality transformations.

Qualitative Comparison Analysis. To qualitatively assess the effectiveness of our proposed framework, we conducted visual comparisons across four representative tasks: T1w-to-CT, T1w-to-PET, CT-to-PET, and T1w-to-FLAIR. As

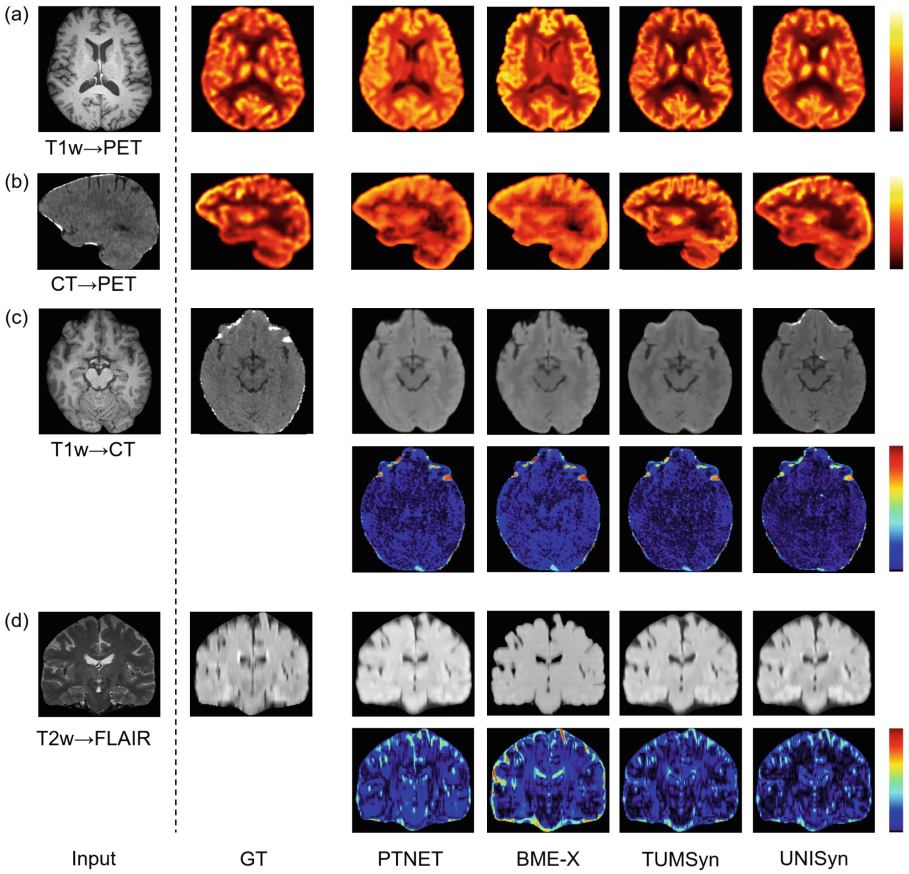


Fig. 2. Results of four cross-modality synthesis tasks from different methods.

illustrated in Fig. 2, the synthesized images generated by UniSyn closely resemble the ground truth (GT), demonstrating superior anatomical detail preservation and contrast consistency compared to three SOTA methods. These findings align with the quantitative evaluation results. The same conclusion is provided by the error maps (second row of each task), which highlight the reduced discrepancies in UniSyn's outputs.

3.4 Ablation Study

To substantiate the effectiveness of our text-guided representation learning and the jointly training strategy, we conduct ablation experiments by replacing the baseline task-specific model, which employs a cross-attention mechanism, with our proposed arithmetic operation. Additionally, we compare the unified training strategy that incorporates all imaging modalities and tasks.

Table 2. Comparison of baseline and Unisyn models

Task	Baseline			Unisyn (specific task)			Unisyn	
	PSNR	SSIM	Inference time (s)	PSNR	SSIM	Inference time (s)	PSNR	SSIM
CT-T1w	25.53 ± 2.09	0.879 ± 0.018	72.8	26.01 ± 2.03	0.893 ± 0.018	50.2	**26.54 ± 2.19**	**0.901 ± 0.021**
T1w-T2w	29.39 ± 2.16	0.938 ± 0.025	81.1	30.43 ± 2.17	0.951 ± 0.027	58.5	**30.74 ± 2.24**	**0.958 ± 0.028**

As detailed in Table 2, the baseline method, which relies on cross-attention mechanisms for target text prompt and source image fusion (Baseline), our arithmetic operation (Unisyn (specific task)) achieves mean improvements of 1.9% in PSNR and 1.6% in SSIM across both tasks. Furthermore, applying the joint training strategy further enhances 0.4 dB PSNR and 0.008 SSIM.

These findings indicate that employing arithmetic operations for image-text fusion significantly reduces the model's learning burden regarding complex interactions between these modalities. Concurrently, the integration of textual metadata facilitates unified training across heterogeneous imaging modalities, subsequently enhancing the robustness of synthetic image generation.

4 Conclusion

In this study, we propose UniSyn, a universal framework for the synthesis of diverse neuroimaging modalities from any available ones, guided by textual metadata. By initially pre-training a domain-specific text encoder, UniSyn effectively captures the factors that differentially influence image contrast and anatomical content from metadata, enabling unified, parameter-customized image synthesis. To seamlessly integrate imaging and non-imaging data, we introduce an arithmetic operation that enhances the precision and robustness of source-target image feature mapping. Extensive experiments on heterogeneous datasets demonstrate UniSyn's ability to generate high-fidelity 3D brain MRI, PET, and CT volumes, highlighting its potential to address the challenge of missing multimodal neuroimaging data in clinical and research settings.

Acknowledgments. This study was funded in part by National Natural Science Foundation of China (grant numbers 82441023, U23A20295, 62131015, 82394432), the China Ministry of Science and Technology (S20240085, STI2030-Major Projects-2022ZD0209000, STI2030-Major Projects-2022ZD0213100), Shanghai Municipal Central Guided Local Science and Technology Development Fund (No. YDZX20233100001001), The Key R&D Program of Guangdong Province, China (grant number 2023B0303040001), and HPC Platform of ShanghaiTech University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Hussein, R., et al.: Turning brain mri into diagnostic pet: 15o-water pet cbf synthesis from multi-contrast mri via attention-based encoder-decoder networks. Med. Image Anal. **93**, 103072 (2024)
2. Ou, Z., Pan, Y., Xie, F., Guo, Q., Shen, D.: Image-and-label conditioning latent diffusion model: synthesizing aβ-pet from mri for detecting amyloid status. IEEE J. Biomed. Health Inform. (2024)
3. Pan, S., et al.: Synthetic ct generation from mri using 3d transformer-based denoising diffusion model. Med. Phys. **51**(4), 2538–2548 (2024)
4. Wang, Y., et al.: Deep learning-based 3d mri contrast-enhanced synthesis from a 2d noncontrast t2flair sequence. Med. Phys. **49**(7), 4478–4493 (2022)
5. Zhang, X., et al.: Ptnet3d: a 3D high-resolution longitudinal infant brain mri synthesizer based on transformers. IEEE Trans. Med. Imaging **41**(10), 2925–2940 (2022)
6. Liu, J., Pasumarthi, S., Duffy, B., Gong, E., Datta, K., Zaharchuk, G.: One model to synthesize them all: multi-contrast multi-scale transformer for missing data imputation. IEEE Trans. Med. Imaging (2023)
7. Zhou, B., et al.: Synthesizing multi-tracer PET images for Alzheimer's disease patients using a 3D unified anatomy-aware cyclic adversarial network. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 34–43. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_4
8. Xue, Y., Bi, L., Peng, Y., Fulham, M., Feng, D.D., Kim, J.: Pet synthesis via self-supervised adaptive residual estimation generative adversarial network. IEEE Trans. Radiat. Plasma Med. Sci. (2023)
9. Jiang, C., Pan, Y., Cui, Z., Nie, D., Shen, D.: Semi-supervised standard-dose pet image generation via region-adaptive normalization and structural consistency constraint. IEEE Trans. Med. Imaging **42**(10), 2974–2987 (2023)
10. Wang, Y., et al.: Toward general text-guided multimodal brain mri synthesis for diagnosis and medical image analysis. Cell Rep. Med. (2025)
11. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
12. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (long and short papers), pp. 4171–4186 (2019)
14. Zhang, S., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
15. Xiong, H., et al.: Contrast representation learning from imaging parameters for magnetic resonance image synthesis. In: Linguraru, M.G., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024. LNCS, vol. 15007, pp. 181–190. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72104-5_18
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Sun, Y., Wang, L., Li, G., Lin, W., Wang, L.: A foundation model for enhancing magnetic resonance images and downstream segmentation, registration and diagnostic tasks. Nat. Biomed. Eng. 1–18 (2024)

Unpaired Multi-site Brain MRI Harmonization with Image Style-Guided Latent Diffusion

Mengqi Wu[1,2], Minhui Yu[1,2], Weili Lin[1], Pew-Thian Yap[1], and Mingxia Liu[1(✉)]

[1] Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
mingxia_liu@med.unc.edu

[2] Lampe Joint Department of Biomedical Engineering, University of North Carolina at Chapel Hill and North Carolina State University, Chapel Hill, NC 27599, USA

Abstract. Multi-site brain MRI heterogeneity caused by differences in scanner field strengths, acquisition protocols, and software versions poses a significant challenge for consistent analysis. Image-level harmonization, leveraging advanced learning methods, has attracted increasing attention. However, existing methods often rely on paired data (*e.g.*, human traveling phantoms) for training, which are not always available. Some methods perform MRI harmonization by transferring target-style features to source images but require explicitly learning disentangled image styles (*e.g.*, contrast) via encoder-decoder networks, which increases computational complexity. This paper presents an unpaired MRI harmonization (UMH) framework based on a new image style-guided diffusion model. UMH operates in two stages: (1) a *coarse harmonizer* that aligns multi-site MRIs to a unified domain via a conditional latent diffusion model while preserving anatomical content; and (2) a *fine harmonizer* that adapts coarsely harmonized images to a specific target using style embeddings derived from a pre-trained Contrastive Language-Image Pre-training (CLIP) encoder, which captures semantic style differences between the original MRIs and their coarsely-aligned counterparts, eliminating the need for paired data. By leveraging rich semantic style representations of CLIP, UMH avoids learning image styles explicitly, thereby reducing computation costs. We evaluate UMH on 4,123 MRIs from three distinct multi-site datasets, with results suggesting its superiority over several state-of-the-art (SOTA) methods across image-level comparison, downstream classification, and brain tissue segmentation tasks.

Keywords: Brain MRI Harmonization · Style Translation · Diffusion

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 683–693, 2026.
https://doi.org/10.1007/978-3-032-04947-6_65

1 Introduction

Brain MRI analysis using learning-based methods, particularly deep learning, often suffers from limited data at individual acquisition sites, resulting in suboptimal performance and poor generalizability across sites. To address this, multi-site MRI pooling is increasingly used in neuroimaging studies to expand sample size, enhance cohort diversity, and improve statistical power [1,22,25]. However, variations in scanner vendors, scanning sequences, and field strength introduce site-specific non-biological style differences (*e.g.*, intensity, contrast, and signal-to-noise ratio), which can hinder model training [12,13,19,28].

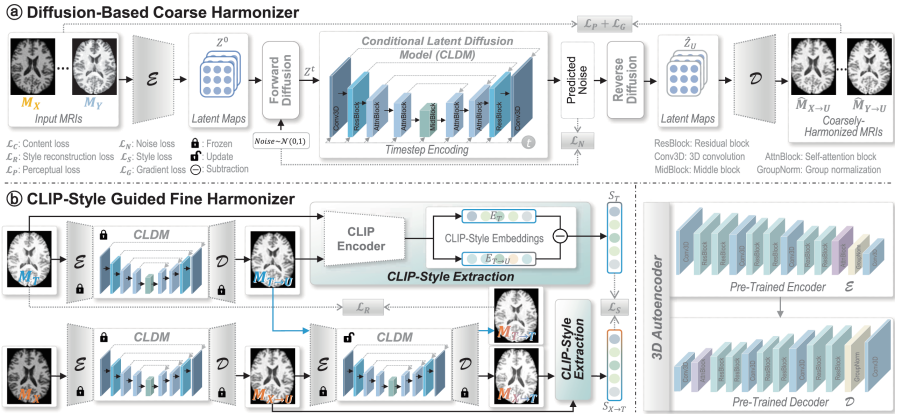


Fig. 1. Proposed UMH framework, containing (a) a diffusion-based coarse harmonizer (DCH) that aligns multi-site MRIs into a unified domain while preserving anatomical content, and (b) a CLIP-style guided fine harmonizer that aligns coarsely-harmonized MRIs to a target style using a disentangled CLIP-style loss. This dual-stage process achieves efficient MRI harmonization without explicit content-style disentanglement.

Data harmonization techniques have been developed to reduce site effects in pre-extracted MRI features or directly normalize raw images. Existing feature-level harmonization methods, such as ComBat [11] and ComBat-GAM [21], utilize the Empirical Bayes framework that models and adjusts the site effect as multiplicative and additive errors in data batches. However, these methods heavily rely on the extracted features [1,4]. Existing image-level harmonization methods often use generative adversarial networks (GANs) for cross-domain image translation [6,16,18], which can be time-consuming and unstable to train. Some approaches use multiple encoder-decoder networks to learn disentangled image style and content representations [4,8,27,31], which is computationally heavy. In addition, they often require paired MRIs, such as traveling phantom data from the same subjects, which may not always be available in practice.

To this end, we propose an unpaired 3D brain MRI harmonization (UMH) framework with a conditional latent diffusion model, guided by semantic style

embeddings derived from the contrastive language-image pre-training (CLIP) encoder [29]. As shown in Fig. 1, UMH operates in two stages. In the first stage, a conditional latent diffusion model (CLDM) is designed to coarsely align multi-site input MRIs into a unified domain while preserving anatomical content information. The second stage involves fine-tuning the coarse harmonizer to align the coarsely-harmonized MRIs to a specific target style guided by a CLIP-based style loss, which leverages the difference of CLIP-style embeddings from the original MRIs and their coarsely-harmonized counterparts to model disentangled style information without requiring paired MRIs. This dual-stage process allows our UMH to leverage the stable training and generative power of diffusion models [20] and the rich semantic representations of CLIP, achieving efficient MRI style transfer without the need for separate encoders and decoders to learn content and style information explicitly. The UMH is trained and evaluated on three multi-site datasets with a total of 4,123 T1-weighted (T1w) MRIs through three tasks. Experimental results demonstrate the superiority of UMH over several state-of-the-art (SOTA) methods in aligning multi-site MRI styles while preserving critical biological and anatomical features.

2 Proposed Method

Problem Formulation. Our goal is to align the style (*e.g.*, intensity and contrast) of source MRIs with a target style while preserving their content (*i.e.*, anatomical structures). We utilize a two-stage approach. **(1) Coarse harmonization**: To reduce computational cost, we perform harmonization in a low-dimensional latent space using a 3D autoencoder. The encoder $\boldsymbol{E}$ compresses an MRI M into a latent map $Z = \boldsymbol{E}(M) \in \mathbb{R}^{c \times h \times w \times d}$, while the decoder $\boldsymbol{D}$ reconstruct the MRI from the latent map Z. Here c, w, h, and d stand for channel, width, height, and depth dimensions, respectively. To coarsely remove site-specific style variations while maintaining anatomical content, we train a conditional latent diffusion model (CLDM) $\boldsymbol{\Phi}$, which maps the latent representations of arbitrary MRIs to a unified latent domain U. The coarsely-harmonized MRIs are then reconstructed as: $M_U = \boldsymbol{D}(Z_U\{C, S_U\}) = \boldsymbol{D} \circ \boldsymbol{\Phi}(Z\{C, S\})$, where C and S represent content and style information, respectively. Two novel image-level constraints are designed to ensure image content preservation. **(2) Fine harmonization**: The CLDM is fine-tuned to align coarsely-harmonized MRIs to a specific target style. Given MR images from a source domain X and a target domain T, we first map them to the unified domain U: for $i \in \{X, T\}$, $M_{i \to U} = \boldsymbol{D} \circ \boldsymbol{\Phi} \circ \boldsymbol{E}(M_i)$, generating coarsely-harmonized images. We then fine-tune CLDM to translate the coarsely-harmonized source latent map to match the target style: $Z_{X \to T}\{C_X, S_T\} = \boldsymbol{\Phi}(Z_U\{C_X, S_U\})$. This translation is guided by a disentangled CLIP-style loss, which captures style embedding differences between each original MRI and its coarsely-harmonized counterpart. Finally, the harmonized source image is reconstructed by decoding the translated latent map $Z_{X \to T}$: $M_{X \to T} = \boldsymbol{D}(Z_{X \to T}\{C, S_T\})$. This two-stage process allows unseen MRIs from a new site to be harmonized *by only fine-tuning the second stage.*

Diffusion-based Coarse Harmonizer (DCH). The DCH is trained to project multi-site MRIs into a coarsely-aligned unified latent domain by leveraging encoder $\boldsymbol{E}$ and CLDM (see Fig. 1). To begin with, we first transform an input MRI M to a low-dimensional latent code $Z^0 = \boldsymbol{E}(M)$ through the pre-trained encoder $\boldsymbol{E}$. The CLDM $\boldsymbol{\Phi}$ is then trained to iteratively reconstruct Z^0 via forward diffusion process (FDP) and reverse-diffusion (RDP) operations, governed by a Markov chain with a total of T timesteps. During FDP, noise is gradually added into Z^0 to create a noisy latent map Z^t at each timestep t:

$$Z^t = \sqrt{\bar{\alpha}_t} Z^0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{1}$$

where ϵ is the sampled noise, $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$, $\alpha_t := 1 - \beta_t$, and β_t follows a predefined variance schedule. To suppress the influence of image style and provide a content condition, the original latent code Z^0 undergoes instance normalization (IN): $\mathrm{IN}(Z^0) = (Z^0 - \mu(Z^0))/\sigma(Z^0)$, where $\mu(\cdot)$ and $\sigma(\cdot)$ compute the mean and standard deviation across channels, effectively removing style information [15, 26]. The resulting $\mathrm{IN}(Z^0)$ can serve as the image content condition. The CLDM, implemented as a time-conditioned 3D U-Net, takes the noisy latent map Z^t at timestep t as input and $\mathrm{IN}(Z^0)$ as conditions to predict the noise ϵ_θ, which is compared against noise ϵ added during training via the *noise loss*:

$$\mathcal{L}_N = \|\epsilon - \epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0))\|_2^2. \tag{2}$$

During RDP, the CLDM employs a DDIM [23] sampling strategy, which iteratively denoises Z^t over T_r timesteps ($t = T_r : 0$) to recover a coarsely-aligned image Z_U during inference, through the following:

$$Z_U^{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{Z}_U + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0))), \quad \hat{Z}_U = \tfrac{1}{\sqrt{\bar{\alpha}_t}}(Z^t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0))), \tag{3}$$

where $\hat{Z}_U$ is the one-step estimate of Z_U^0. Typically, Z_U^t is iteratively denoised until Z_U^0 is obtained. During training, we use this one-step estimate and decode it to a coarsely-aligned MRI $\hat{M}_U = \boldsymbol{D}(\hat{Z}_U)$ for efficiency in this work.

We further employ two novel *image-level constraints* to enforce MR image content preservation. This is achieved by incorporating a gradient loss $\mathcal{L}_G$ and a perceptual loss $\mathcal{L}_P$ in CLDM. Specifically, the gradient loss is designed to measure the difference between the gradient maps of $\hat{M}_U$ and M:

$$G(M) = \frac{1}{3}(\nabla_h M + \nabla_w M + \nabla_d M), \quad \mathcal{L}_G = \|G(\hat{M}_U) - G(M)\|_2^2, \tag{4}$$

where ∇ denotes the gradient operation. The perceptual loss compares feature maps extracted from a pre-trained MedicalNet ResNet-50 [7]:

$$\mathcal{L}_P = \sum\nolimits_l \lambda_l \left\| \psi_l(M) - \psi_l(\hat{M}_U) \right\|_1, \tag{5}$$

where ψ_l denote feature maps extracted from the l-th layer of the ResNet-50 and λ_l controls the contribution of each layer. These losses defined at the image level ensure content and semantic consistency during training, effectively guiding

CLDM to generate style-agnostic reconstruction while preserving anatomical integrity. The hybrid loss for training our DCH is defined as $\mathcal{L}_C = \mathcal{L}_N + \mathcal{L}_G + \mathcal{L}_P$.

CLIP-Style Guided Fine Harmonizer. In this stage, we fine-tune the DCH to align a coarsely-aligned source MRI $M_{X\to U}$ into a translated image $M_{X\to T}$ with a style of a target site T. We leverage a pre-trained CLIP encoder [29] to implicitly extract style embeddings and develop a *hybrid disentangled CLIP-style loss*, eliminating the need for paired data or explicit style definitions.

Given a source MRI M_X and an unpaired target MRI M_T, we first map them to the coarsely-aligned unified domain through FDP defined in Eq. (1) and RDP in Eq. (3) while fixing the weight of the trained DCH:

$$M_{X\to U} = \boldsymbol{D} \circ \boldsymbol{\Phi} \circ \boldsymbol{E}(M_X), \quad M_{T\to U} = \boldsymbol{D} \circ \boldsymbol{\Phi} \circ \boldsymbol{E}(M_T). \tag{6}$$

The fine harmonizer is then trained to adapt $M_{X\to U}$ to match the style of M_T through other FDP and RDP processes. Unlike FDP in DCH, where random noise is added to Z^0 through Eq. (1), we iteratively add learned noise from the coarse harmonizer over T_f iterations ($t = 0 : T_f$) to get the noisy latent Z^{T_f}:

$$Z_X^0 = Z_{X\to U} = \boldsymbol{E}(M_{X\to U}), \quad Z_X^{t+1} = \sqrt{\bar{\alpha}_{t+1}}\hat{Z}_X^0 + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0)), \tag{7}$$

where $\hat{Z}_X^0$ is the one-step estimate similar to $\hat{Z}_U$, see Eq. (3). RDP is then used to iteratively denoise $Z_X^{T_f}$ for $t = T_r : 0$ times to yield the final translated feature:

$$Z_{X\to T}^{T_r} = Z_X^{T_f}, \quad Z_{X\to T}^{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{Z}_{X\to T}^0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(Z^t, t, \mathrm{IN}(Z^0)), \tag{8}$$

where $\hat{Z}_{X\to T}^0$ is the one-step estimate. After T_r iterative timesteps, we decode the translated latent map $Z_{X\to T}^0$ to obtain the translated MRI: $M_{X\to T} = \boldsymbol{D}(Z_{X\to T}^0)$.

To align source MRIs with target style, we extract embeddings for both target and coarsely-harmonized source images via a pre-trained CLIP encoder Ψ:

$$S_T = \Psi(M_T) - \Psi(M_{T\to U}), \quad S_{X\to T} = \Psi(M_{X\to T}) - \Psi(M_{X\to U}), \tag{9}$$

Since $M_{T\to U}$ is the coarsely-harmonized unified version of the target MRI M_T (with the same content), the difference between their CLIP-space embeddings (*i.e.*, S_T) captures the target style information. Similarly, $S_{X\to T}$ captures the style of the harmonized source image. We define the *style translation loss* as:

$$\mathcal{L}_S = \|S_T - S_{X\to T}\|_1 + (1 - (S_T \cdot S_{X\to T})/(\|S_T\|\|S_{X\to T}\|)), \tag{10}$$

where the 1st term is the l_1 distance in the CLIP-embedding space and the 2nd term quantifies the directional discrepancy between two style embeddings.

To ensure style consistency, we further design a *style reconstruction loss* by minimizing style embeddings of each target MRI and its harmonized counterpart: $\mathcal{L}_R = \|S_T - S_{T\to T}\|_1$. The hybrid disentangled CLIP-style loss is defined as:

$$\mathcal{L}_F = \mathcal{L}_S + \mathcal{L}_R. \tag{11}$$

By leveraging CLIP's semantic-rich embeddings, our fine harmonizer effectively translates source MRIs to the target style *without requiring explicit image*

style and content disentanglement learning, ensuring that anatomical content remains unchanged. The two-stage training strategy used in the proposed UMH allows the coarse harmonizer to be employed universally, requiring only the fine-tuning stage for harmonizing unseen MRI data, providing superior flexibility.

Implementation. We use the 3D AutoencoderKL from MONAI [5], with three groups of upsampling/downsampling 3D convolutional layers with residual blocks ({32, 64, 128} channels). CLDM is implemented as a time-conditioned 3D U-Net with a symmetric architecture: 2 upsampling/downsampling layers, 2 residual blocks, 4 self-attention blocks, and a middle block ({32, 64, 64, 64} channels). Both models are trained using the Adam optimizer with default settings and an initial learning rate of 1×10^{-4}. The variance scheduler β is empirically set to increase linearly from 0.0015 to 0.0195. We empirically apply $T = 1,000$ noise steps for the first coarse harmonization stage and choose $T_f = 25, T_r = 35$ for the second fine harmonization stage through hyperparameter grid search.

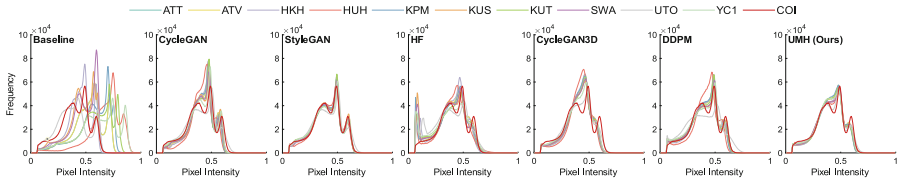


Fig. 2. Histograms of 22 SRPBS test MRIs across 11 sites, with COI as the target.

Table 1. Comparison between source site MRIs and corresponding target site (COI) MRIs with matching subjects in the SRPBS test set (2 subjects across 11 sites).

Method	SSIM ↑	PSNR ↑	PCC ↑	WD ↓
Baseline	0.879 ± 0.033	22.020 ± 4.282	0.979 ± 0.008	0.041 ± 0.026
CycleGAN [6]	0.868 ± 0.026	21.410 ± 1.652	0.973 ± 0.009	0.028 ± 0.011
StyleGAN [16]	0.893 ± 0.028	23.498 ± 1.756	0.975 ± 0.009	$\mathbf{0.005 \pm 0.002}$
HF [2]	0.890 ± 0.024	23.411 ± 1.427	0.976 ± 0.011	0.016 ± 0.006
CycleGAN3D [30]	$\mathbf{0.937 \pm 0.025}$	26.906 ± 1.771	0.988 ± 0.006	0.013 ± 0.004
DDPM [10]	0.869 ± 0.124	30.021 ± 10.027	0.968 ± 0.032	0.015 ± 0.009
UMH (Ours)	0.933 ± 0.097	$\mathbf{31.263 \pm 2.025}$	$\mathbf{0.989 \pm 0.005}$	$\mathbf{0.005 \pm 0.003}$

3 Experiments

Materials. Three datasets are utilized: (1) OpenBHB [9], with T1w MRIs from 3,984 healthy subjects across 58 sites; (2) SRPBS [24], with T1w MRIs from 9

traveling subjects across 11 sites; and (3) DWI-THP [17] with T1w MRIs from 5 subjects scanned across 8 sites. The OpenBHB dataset is divided into a training set (3,227 MRIs) and a validation set (757 MRIs) for training the 3D autoencoder and CLDM. The SRPBS and DWI-THP datasets are used for fine-tuning and evaluation, with data splits detailed in the experimental section.

Competing Methods. We compare UMH with five SOTA image-level MRI harmonization methods: CycleGAN [6], StyleGAN [16], CycleGAN-3D [30], Harmonizing Flow (HF) [2], and DDPM [10]. We ensured consistent training data and hyperparameters across all methods for a fair comparison.

Task 1: Histogram and Voxel-Level Comparison. This experiment evaluates harmonization outcomes using SRPBS with ground-truth traveling phantom data. We train on 77 MRIs (7 subjects across 11 sites) and test on 22 MRIs (2 subjects across 11 sites). Each method harmonizes MRIs from 10 source sites to a target site, selected as COI due to its highest mean peak-signal-to-noise ratio (PSNR). Performance is assessed via histogram comparisons and voxel-level metrics: PSNR, structural similarity index (SSIM), Wasserstein distance (WD), and Pearson correlation coefficient (PCC) with raw source MRIs as Baseline. The results in Fig. 2 show significant site-wise intensity variations in raw MRIs. Our method effectively aligns the source histograms to the target's, performing comparably to StarGAN without explicit style learning. Quantitative results in Table 1 suggest that our UMH yields the highest PSNR and PCC, indicating superior voxel-level agreement, and the lowest WD, confirming effective style alignment. It also achieves the second-best SSIM, demonstrating strong anatomical preservation. Sample visualization in Fig. 3 (a) shows that UMH harmonized MRI is closer to the target. These results highlight UMH's ability to harmonize MRIs while maintaining high image quality and anatomical fidelity.

Task 2: Brain Tissue Segmentation. We further evaluate anatomical preservation via a brain tissue segmentation task. Each method is trained on 24 MRIs (3 subjects across 8 sites) and harmonizes 16 MRIs from 2 additional subjects in the DWI-THP dataset with site CCF as the target. We use FreeSurfer [3] to generate gray matter (GM) and white matter (WM) segmentation maps for original and harmonized MRIs. Segmentation quality is assessed using the Anatomical Preservation (AP) score [19], and Dice Similarity Coefficient (DSC). The results in Table 2 show that our method achieves the highest performance for GM and mean scores and the second-highest for WM. Figure 3 (b) further shows that UMH yields fewer segmentation errors in the WM and GM tissue boundaries.

Task 3: Site Classification and Brain Age Prediction. We evaluate UMH in reducing site-related variations while preserving anatomical information through site classification and brain age prediction. Each method is trained on OpenBHB training data and applied to harmonize the validation data with Site 17 selected as the target site. We extract features from these harmonized MRIs using ResNet18 [14] with its final layer removed. A logistic regression, trained on 70% of these deep features and tested on 30%, performs multi-class site classification, and a ridge regressor predicts brain age. Both tasks are repeated 5 times

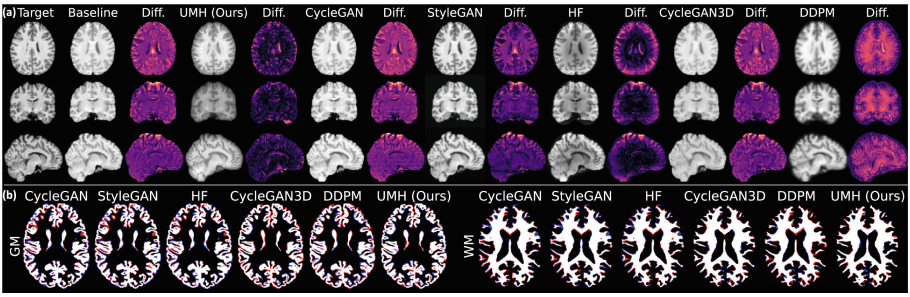


Fig. 3. (a) harmonization results and difference (Diff.) and (b) segmentation maps with *White*: accurate segmentation; *Red*: under-segmentation; *Blue*: over-segmentation. (Color figure online)

Table 2. Results of AP and DSC metrics on GM and WM segmentation in DWI-THP.

Method	Anatomical Preservation (AP) ↑			Dice Similarity Coefficient (DSC) ↑		
	GM	WM	Mean	GM	WM	Mean
CycleGAN [6]	0.986 ± 0.017	$\mathbf{0.994 \pm 0.004}$	0.990 ± 0.009	0.880 ± 0.046	0.920 ± 0.030	0.900 ± 0.038
StyleGAN [16]	0.992 ± 0.006	$\mathbf{0.994 \pm 0.005}$	$\mathbf{0.993 \pm 0.004}$	0.886 ± 0.043	0.924 ± 0.029	0.905 ± 0.036
HF [2]	0.985 ± 0.016	0.990 ± 0.009	0.987 ± 0.010	0.879 ± 0.046	0.918 ± 0.029	0.899 ± 0.038
CycleGAN3D [30]	0.988 ± 0.019	0.992 ± 0.006	0.990 ± 0.010	0.886 ± 0.055	$\mathbf{0.932 \pm 0.032}$	0.909 ± 0.043
DDPM [10]	0.992 ± 0.005	0.989 ± 0.009	0.990 ± 0.007	0.829 ± 0.168	0.847 ± 0.154	0.838 ± 0.161
UMH (Ours)	$\mathbf{0.993 \pm 0.010}$	0.992 ± 0.006	$\mathbf{0.993 \pm 0.006}$	$\mathbf{0.898 \pm 0.045}$	0.926 ± 0.030	$\mathbf{0.912 \pm 0.038}$

for random data partition, with mean ± standard deviation results reported in Table 3. Lower site classification results indicate better removal of site-related variations, while lower age prediction error suggests superior anatomical feature preservation. Table 3 shows that UMH yields the lowest site classification performance in BAC, F1, and Recall, effectively removing site-related features and maintaining faithful anatomical integrity with slightly lower mean absolute error (MAE).

Table 3. Site classification and age prediction results on harmonized OpenBHB MRIs.

Method	Site Classification				Age Prediction	
	Balanced Accuracy (BAC) ↓	F1 ↓	Precision (PRE) ↓	Recall ↓	MAE ↓	MSE ↓
Baseline	0.343 ± 0.024	0.663 ± 0.023	0.757 ± 0.018	0.732 ± 0.019	5.295 ± 0.260	47.446 ± 1.405
CycleGAN [6]	0.425 ± 0.016	0.695 ± 0.027	0.770 ± 0.030	0.739 ± 0.020	6.625 ± 0.264	78.951 ± 10.541
StyleGAN [16]	0.258 ± 0.022	0.593 ± 0.012	0.662 ± 0.015	0.651 ± 0.015	7.314 ± 0.494	85.716 ± 12.905
HF [2]	0.342 ± 0.011	0.665 ± 0.020	0.736 ± 0.021	0.723 ± 0.021	5.835 ± 0.221	57.009 ± 3.849
CycleGAN3D [30]	0.324 ± 0.029	0.656 ± 0.019	0.751 ± 0.019	0.723 ± 0.017	5.901 ± 0.360	$\mathbf{33.348 \pm 9.671}$
DDPM [10]	0.166 ± 0.016	0.560 ± 0.013	$\mathbf{0.545 \pm 0.007}$	0.535 ± 0.020	5.333 ± 0.258	48.548 ± 6.796
UMH (Ours)	$\mathbf{0.129 \pm 0.013}$	$\mathbf{0.510 \pm 0.035}$	0.625 ± 0.024	$\mathbf{0.522 \pm 0.042}$	$\mathbf{5.240 \pm 0.141}$	52.022 ± 5.052

Ablation Study. We validate three key components by comparing UMH with its three degraded variants: UMH-I (no image-level constraints during coarse

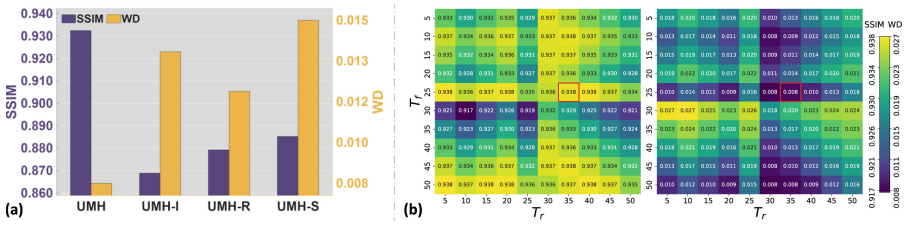


Fig. 4. Results of (a) UMH and its variants, and (b) UMH with different T_f and T_r.

harmonizer training), UMH-R (without style reconstruction loss), and UMH-S (without style translation loss). Results in Fig. 4 (a) show that UMH-I achieves the lowest SSIM, indicating our image-level constraints help preserve structural details. Additionally, UMH-S yields the highest WD, reflecting poor style alignment without style translation loss. UMH-R's suboptimal performance further underscores the importance of style reconstruction loss during fine-tuning.

Parameter Analysis. We conduct a hyperparameter grid search on T_f (forward diffusion steps) and T_r (reverse diffusion steps), evaluating volume-level metrics on the SRPBS test set. Results in Fig. 4 (b) show that our model performs better with larger T_r than T_f, with peak performance when $T_f = 25$ and $T_r = 35$.

Computational Cost Comparison. The proposed UMH employs a lightweight latent-diffusion U-Net and a 3D autoencoder with 3.0M and 3.3M trainable parameters, respectively, totaling 6.3M—smaller than CycleGAN (28.3M), StyleGAN (161.3M), CycleGAN3D (22.6M), and DDPM (10.3M), and comparable to HF (5.7M). Training UMH on the SRPBS cohort takes about 6.5 h (3.0h for Stage 1 on OpenBHB; 2.5 h for Stage 2 fine-tuning), which is faster than CycleGAN (16.2 h), StyleGAN (10.5 h), HF (13.4 h), and CycleGAN3D (12.4 h), and comparable to DDPM (5.5 h). All models were trained on an H100 GPU.

4 Conclusion

We propose an unpaired MRI harmonization framework using a latent diffusion model, which aligns MRIs into a unified latent space and translates them into the target style via disentangled CLIP-based style losses. Leveraging the semantic-rich CLIP embeddings, our method enables effective volume-level harmonization without paired MRIs or auxiliary style encoders. Extensive experiments across three multi-site datasets demonstrate UMH's superiority in style alignment, site variation removal, and anatomical preservation.

Acknowledgments. This study was supported in part by NIH grants (Nos. AG073297, AG082938, EB035160, and NS134849).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. An, L., et al.: Goal-specific brain MRI harmonization. Neuroimage **263**, 119570 (2022)
2. Beizaee, F., Desrosiers, C., Lodygensky, G.A., Dolz, J.: Harmonizing flows: unsupervised mr harmonization based on normalizing flows. In: Frangi, A., de Bruijne, M., Wassermann, D., Navab, N. (eds.) Information Processing in Medical Imaging. IPMI 2023. LNCS, vol. 13939, pp. 347–359. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-34048-2_27
3. Billot, B., Magdamo, C., Cheng, Y., Arnold, S.E., Das, S., Iglesias, J.E.: Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. Proc. Natl. Acad. Sci. **120**(9), e2216399120 (2023)
4. Cackowski, S., Barbier, E.L., Dojat, M., Christen, T.: Imunity: a generalizable VAE-GAN solution for multicenter MR image harmonization. Med. Image Anal. **88**, 102799 (2023)
5. Cardoso, M.J., et al.: Monai: an open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
6. Chang, X., et al.: Self-supervised learning for multi-center magnetic resonance imaging harmonization without traveling phantoms. Phys. Med. Biol. **67**(14), 145004 (2022)
7. Chen, S., Ma, K., Zheng, Y.: Med3D: transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
8. Dewey, B.E., et al.: A disentangled latent space for cross-site MRI harmonization. In: Martel, A.L., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. LNCS, vol 12267. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59728-3_70
9. Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., Duchesnay, E.: OpenBHB: a large-scale multi-site brain MRI data-set for age prediction and debiasing. Neuroimage **263**, 119637 (2022)
10. Durrer, A., et al.: Diffusion models for contrast harmonization of magnetic resonance images. arXiv preprint arXiv:2303.08189 (2023)
11. Fortin, J.P., et al.: Harmonization of cortical thickness measurements across scanners and sites. Neuroimage **167**, 104–120 (2018)
12. Gadewar, S.P., et al.: Synthesizing study-specific controls using generative models on open access datasets for harmonized multi-study analyses. arXiv preprint arXiv:2403.00093 (2024)
13. Glocker, B., Robinson, R., Castro, D.C., Dou, Q., Konukoglu, E.: Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects. arXiv preprint arXiv:1910.04597 (2019)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
16. Liu, M., Maiti, P., Thomopoulos, S., Zhu, A., Chai, Y., Kim, H., Jahanshad, N.: Style transfer using generative adversarial networks for multi-site MRI harmonization. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. LNCS, vol. 12903, pp. 313–322. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_30

17. Magnotta, V.A., et al.: Multicenter reliability of diffusion tensor imaging. Brain Connectivity **2**(6), 345–355 (2012)
18. Modanwal, G., Vellal, A., Buda, M., Mazurowski, M.A.: MRI image harmonization using cycle-consistent generative adversarial network. In: Computer-Aided Diagnosis, vol. 11314, pp. 259–264. SPIE (2020)
19. Parida, A., Jiang, Z., Packer, R.J., Avery, R.A., Anwar, S.M., Linguraru, M.G.: Quantitative metrics for benchmarking medical image harmonization. In: IEEE ISBI, pp. 1–5. IEEE (2024)
20. Pfaff, L., et al.: No-new-denoiser: a critical analysis of diffusion models for medical image denoising. In: Linguraru, M.G., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024. LNCS, vol. 15010, pp. 568–578. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72117-5_53
21. Pomponio, R., et al.: Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage **208**, 116450 (2020)
22. Schnack, H.G., et al.: Mapping reliability in multicenter MRI: voxel-based morphometry and cortical thickness. Hum. Brain Mapp. **31**(12), 1967–1982 (2010)
23. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
24. Tanaka, S., et al.: A multi-site, multi-disorder resting-state magnetic resonance image database. Scientific Data **8**(1), 227 (2021)
25. Tofts, P., Collins, D.: Multicentre imaging measurements for oncology and in the brain. Br. J. Radiol. **84**, S213–S226 (2011)
26. Ulyanov, D.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
27. Wu, M., Zhang, L., Yap, P.T., Zhu, H., Liu, M.: Disentangled latent energy-based style translation: an image-level structural MRI harmonization framework. Neural Netw. **184**, 107039 (2025)
28. Yu, M., Guan, H., Fang, Y., Yue, L., Liu, M.: Domain-prior-induced structural MRI adaptation for clinical progression prediction of subjective cognitive decline. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. LNCS, vol. 13431, pp. 24–33. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16431-6_3
29. Zhang, S., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2024)
30. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
31. Zuo, L., et al.: Unsupervised MR harmonization by learning disentangled representations using information bottleneck theory. Neuroimage **243**, 118569 (2021)

Author Index

A
Aftabi, Hamidreza 448
Ahn, Jung Su 386
Ahn, Minkyu 235
Ahn, Sung Jun 140
Alus, Or 24
Amruthalingam, Ludovic 312
An, Jonghoon 57
Atlaw, Meklit 598

B
Bae, Kyongtae Tyler 376
Bai, Yang 609
Balakrishnan, Guha 194
Breininger, Katharina 172
Bustin, Héloïse 418

C
Cai, Kunyan 533
Cao, Chao 13
Cascarano, Pasquale 77
Chan, Raymond H. 77
Chang, Chengyuan 224
Chen, Bailiang 67
Chen, Boyu 365
Chen, Chu 77
Chen, Geng 598
Chen, Jiahui 224
Chen, Minheng 13
Chen, Nanguang 501
Chen, Ning 161
Chen, Rusi 407
Chen, Siqi 257
Chen, Tong 13
Chen, Xuejin 257
Chen, Ziyi 673
Cheng, Kun 35
Cho, Young-Rae 386
Choi, Heeyoul 235
Chu, Xiaowen 544
Chun, Se Young 396
Chung, Minyoung 57
Cui, Hengfei 598
Cui, Jian 428
Cui, Kangning 77

D
Dai, Boyang 662
Dai, Jingguo 428
de Wilde, Bram 630
Deng, Xuedong 344
Ding, Amanda 448
Dong, Caixia 257
Dou, Hao 428
Dou, Haoran 3, 407
Duan, Jinming 3

E
Endo, Hidenori 355
Essofi, Abdelmajid 204

F
Fan, Xuwei 161
Fan, Zeng 662
Fels, Sidney 448
Feng, Qianjin 120
Finck, Marten J. 523
Fleishman, Greg M. 323
Fletcher, P. Thomas 323
Frangi, Alejandro F. 3
Fu, Kaiwen 224
Fu, Yingji 501

G
Gao, Mengze 481
Gao, Xinru 344
Gao, Zhibin 161
Glüer, Claus-C. 523
Gong, Jing 279
Gonzalez-Jimenez, Alvaro 312
Gottfrois, Philippe 312

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
J. C. Gee et al. (Eds.): MICCAI 2025, LNCS 15962, pp. 695–699, 2026.
https://doi.org/10.1007/978-3-032-04947-6

Gröger, Fabian 312
Guo, Jun 120

H
Han, Ji Yong 577
Han, Luyi 183
Han, Xu 333, 566
Han, Yuchi 555
Hansen, Stine 651
Hao, Qun 662
Hardan, Shahad 204
He, Min 533
He, Yuanzhe 673
Hennemuth, Anja 418
Hodgson, Antony 448
Homma, Noriyasu 355
Hong, Charmgil 235
Hong, Gil-sun 214
Hong, Mingxuan 301
Hu, Qinjie 224
Hu, Yibo 512
Huang, Hailiang 459
Huang, Jinghan 3
Huang, Lianfen 161
Huang, Wenqi 481
Huang, Yanchao 333, 566
Huang, Yuhao 344, 407

J
Jeong, Hwihun 396
Jia, Dan 407
Jiang, Han 619
Jiang, Haotian 598
Jiang, Jimao 130
Jiang, Tianzi 428
Jin, Pengfei 99
Jin, Yueming 301
Jo, Wooyoung 214
Jordan, Jakob 418

K
Kampffmeyer, Michael 651
Kanamori, Masayuki 355
Kang, MungSoo 24
Ke, Wei 183
Kee, Youngwook 24
Kellman, Peter 555
Kim, Dahee 577
Kim, Harim 235
Kim, Hyeongji 651
Kim, Jihwan 214
Kim, Namkug 214
Kim, Sekeun 99
Kim, So Hyun 57
Kim, Sujeong 577
Kim, Sunjung 577
Konukoglu, Ender 46
Koohi-Moghadam, Mohamad 376
Koser, Niklas C. 523
Kuang, Xiaodong 459
Kwak, Ki Hoon 386

L
Lai, Jiahao 257
Lajoinie, Guillaume 630
Lam, Chan-Tong 183
Lam, Chi Kin 183
Lee, Dong Hyun 57
Lee, Hyunki 214
Lee, Hyunyeol 140
Lee, Ji Young 140
Lee, Jongho 396
Lee, Joonsang 214
Li, Hao 428
Li, Hui 587
Li, Lei 67
Li, Lin 544
Li, Mingqian 470
Li, Qian 301
Li, Qing 609
Li, Quanzheng 99
Li, Ran 344
Li, Shuo 110
Li, Xiang 99
Li, Yang 257
Li, Yaxuan 619
Li, Yuxing 246
Li, Ziyu 609
Liang, Huanwen 344
Liang, Yaodong 470
Liao, Congyu 481
Liao, Zhijun 150
Lim, Sang-Heon 577
Lin, Weili 683
Lin, Xin 673
Lin, Yimeng 481
Lionetti, Simone 312
Liu, Haoyang 355
Liu, Huafeng 459

Liu, Jiameng 673
Liu, Kun 224
Liu, Mingxia 683
Liu, Tianming 13
Liu, Yanjun 344
Liu, Yanzhen 587
Lloyd, John E. 448
Long, Yuxuan 459
Loo, Junn Yong 640
Lu, Hao 587
Lu, Lijun 333, 566
Lu, Wantong 566
Lu, Yijun 566
Lv, Sangyin 438
Lyu, Yanjun 13

M
Ma, Qiongxiong 470
Ma, Yifei 619
Meyer, Tom 418
Moon, Taejin 57
Mu, Tingting 3
Mu, Wei 290

N
Nagai, Arata 355
Nailon, William Henry 35
Nam, Yoojin 214
Nandakumar, Karthik 204
Navarini, Alexander A. 312
Ni, Dong 344, 407
Nikouei Mahani, Mohammad-Ali 376
Niu, Xiaolong 333, 566
Noman, Fuad 640

O
Odille, Freddy 67
Oh, Joonseo 214
Oh, Yujin 99
Ombao, Hernando 640
Ondruschka, Benjamin 523

P
Park, Joon Hyung 214
Pei, Yuru 130
Peng, Qin 279
Phan, Raphaël C.-W. 640
Piccolomini, Elena Loli 77
Pirk, Sören 523
Pouly, Marc 312
Prisman, Eitan 448
Pu, Cailing 438

Q
Qi, Fei 224
Qin, Shihua 619
Qiu, Anqi 501
Qu, Liangqiong 533

R
Rao, Fan 459
Rietberg, Max T. 630
Rueckert, Daniel 481

S
Sack, Ingolf 418
Safdari, Reza 376
Sagl, Benedikt 448
Salcudean, Septimiu E. 301
Sang, Yudi 587
Schillinger, Moritz 172
Schlereth, Maja 172
Schulthess, Nico 46
Seo, Jung Woo 386
Setsompop, Kawin 481
Shen, Bingzhi 257
Shen, Cong 150
Shen, Daiyun 301
Shen, Dinggang 673
Shen, Shaowei 161
Sheng, Xicheng 67
Shi, Daqian 365
Simonetti, Orlando 555
Solebo, Ameenat 365
Son, Wonjun 140
Song, Hongning 407
Song, Sifan 99
Su, Li 13
Su, Yiyan 662
Sun, Hao 333, 566
Sun, Jianqi 512
Sun, Kaicong 673
Sun, Rui-Qing 662
Sun, Yue 183

T
Tan, Tao 183, 533
Tan, Yee-Fan 640

Tang, Dong 544
Tang, Jijun 150
Tang, Wei 77
Tang, Xinyu 640
Tao, Guowei 344
Tao, Qian 88, 555
Tao, Xing 407
Taratynova, Darya 204
Taylor, Paul 365
Tehlan, Kartikay 491
Tew, Hwa Hui 640
Tian, Jie 290
Ting, Chee-Ming 640
Tivnan, Matthew 99
Tong, Tong 183
Tzschätzsch, Heiko 418

V
Veeraraghavan, Ashok 194
von Brackel, Felix N. 523
Vyas, Kushal 194

W
Walczak, Lars 418
Wang, Chen 428
Wang, Dongfang 268
Wang, Guisheng 35
Wang, Hong 459
Wang, Jinliang 120
Wang, Nan 481
Wang, Qian 673
Wang, Shuai 533
Wang, Xiaotian 224
Wang, Yu 587
Wang, Yuhan 235
Wang, Yulin 673
Wen, Xuyun 598
Wendler, Thomas 491
Wolterink, Jelmer M. 630
Won, Jongjun 214
Woo, Jonghye 619
Wu, Dufan 99
Wu, Hao 110
Wu, Hubing 333, 566
Wu, Jinge 365
Wu, Mengqi 683
Wu, Weitong 279
Wu, Xia 609
Wu, Xinbao 587
Wu, Xuanbin 566
Wu, Yun 344
Wu, Zijian 301

X
Xia, Yong 598
Xie, Xinghe 183
Xing, Fangxu 619
Xing, Zeyu 428
Xing, Zhaohu 279
Xiong, Honglin 673
Xu, Jingxian 344
Xu, Lisa X. 512
Xu, Weixin 290
Xue, Hui 555

Y
Yan, Chenggang 533
Yan, Miaoning 470
Yan, Yuanting 438
Yan, Zhiqian 470
Yang, Jian 268
Yang, Jie 161
Yang, Ronghao 407
Yang, Shuojue 301
Yang, Su 577
Yang, Tongyun 555
Yang, Wei 120
Yang, Xiaofei 544
Yang, Xin 344, 407
Yang, Yuanting 407
Yao, Jiezhi 407
Yap, Pew-Thian 683
Yaqub, Mohammad 204
Ye, Chuyang 246, 662
Ye, Zanting 333, 566
Yi, Won-Jin 577
Yibulayimu, Sutuke 587
Yoo, Yereen 214
Yoon, Siyeop 99
Yu, Chengjin 438

Yu, Fei 544
Yu, Minhui 683
Yu, Xiaowei 13
Yum, Jieun 214
Yun, Heejin 577

Z
Zakeri, Arezoo 3
Zeng, Yuwen 355
Zhai, Penghua 290
Zhan, Yihong 161
Zhang, Dong 438
Zhang, Guoqing 257
Zhang, Hao 438
Zhang, Ji 407
Zhang, Jing 13
Zhang, Lu 13
Zhang, Qi 512
Zhang, Qingmao 470
Zhang, Shaohui 662
Zhang, Sihong 257
Zhang, Wen 120
Zhang, Xiaoyong 355
Zhang, Xin 428
Zhang, Yang 67
Zhang, Yanyan 110
Zhang, Yi 88, 555
Zhang, Yizhe 268
Zhang, Yuanji 344
Zhang, Yue 120
Zhang, Yuhan 344, 407
Zhang, Yukun 35
Zhang, Yulin 459
Zhang, Zhiyu 150
Zhao, Chenyu 120
Zhao, Chunpeng 587
Zhao, Sheng 344
Zhao, Yidong 88, 555
Zhao, Ziqi 512
Zheng, Jian 183
Zheng, Yefeng 459
Zheng, Zhong 428
Zhou, Qing 407
Zhou, Tao 268
Zhou, Wentong 355
Zhou, Yongsong 407
Zhou, Yuyin 235
Zhou, Zherui 3
Zhu, Dajiang 13
Zhu, Gang 587
Zhu, Lei 279
Zhu, Pengli 501
Zhu, Qikui 110
Zhu, Wentao 459
Zhu, Zhiyuan 609
Zhuang, Xiahai 67
Zhuang, Yan 13