Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging

Grewal, M.

DOI
10.4233/uuid:7fa49f2e-371e-4e9e-a0aa-9362733a2905

Publication date
2025

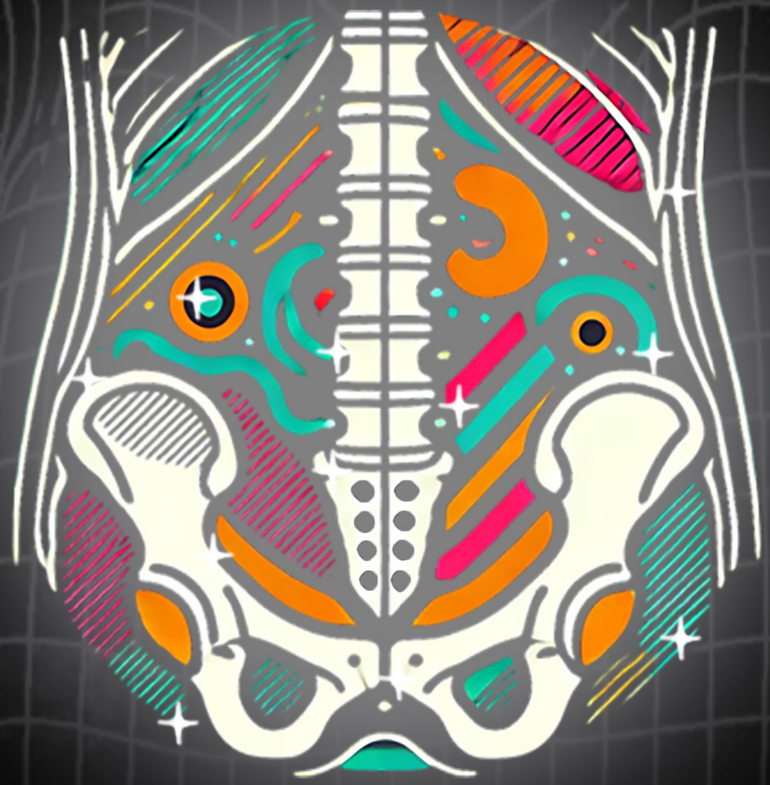
Document Version
Final published version

Citation (APA)
Grewal, M. (2025). *Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging*. [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:7fa49f2e-371e-4e9e-a0aa-9362733a2905

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

DEEP LEARNING FOR LANDMARK DETECTION, SEGMENTATION, AND MULTI-OBJECTIVE DEFORMABLE REGISTRATION IN MEDICAL IMAGING

Monika Grewal



Deep learning for landmark detection, segmentation, and multi-objective deformable registration in medical imaging - Monika Grewal

Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging

DEEP LEARNING FOR LANDMARK DETECTION, SEGMENTATION, AND MULTI-OBJECTIVE DEFORMABLE REGISTRATION IN MEDICAL IMAGING

Proefschrift

ter verkrijging van de graad van doctor

aan de Technische Universiteit Delft,

op gezag van de Rector Magnificus prof. dr. ir. Tim van der Hagen,

voorzitter van het College voor Promoties,

in het openbaar te verdedigen op Friday 11 April 2025 om 12:30 uur

door

Monika GREWAL

Master of Technology in Electronics & Communication Engineering,
Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India,

born in Delhi, India.

Dit proefschrift is goedgekeurd door de

Promotor:	prof. dr. P.A.N. Bosman
Copromotors:	dr. T. Alderliesten
	dr. G. H. Westerveld, MD

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
prof. dr. P.A.N. Bosman	Technische Universiteit Delft, NL; Centrum Wiskunde & Informatica, NL
dr. T. Alderliesten	Leids Universitair Medisch Centrum, NL; Technische Universiteit Delft, NL
dr. G. H. Westerveld, MD	Erasmus Medisch Centrum, NL
Onafhankelijke leden:	
prof. dr. M.B. van Herk	University of Manchester, UK
prof. dr. K. Krawiec	Pozman University of Technology, Poland
prof. dr. R.A. Nout	Technische Universiteit Delft, NL; Erasmus Medisch Centrum, NL
prof. dr. ir. M. Staring	Leids Universitair Medisch Centrum, NL
prof. dr. M. T. J. Spaan,	Technische Universiteit Delft, NL (reservelid)

TU Delft Delft University of Technology
CWI Centrum Wiskunde & Informatica
LUMC Leids Universitair Medisch Centrum
NWO
Rijksdienst voor Ondernemend Nederland
Elekta
Xamnia
SIKS

This research is part of the research programme Open Technology Programme with project number 15586, which is financed by the Dutch Research Council (NWO), Elekta, and Xomnia. Further, the work is co-funded by the public-private partnership allowance for top consortia for knowledge and innovation (TKIs) from the Dutch Ministry of Economic Affairs. The research has been carried out in collaboration with the Amsterdam UMC, location Academic Medical Center (AMC) and Leiden University Medical Center (LUMC), Netherlands.

We thank Jan Wiersma (Amsterdam UMC location AMC, The Netherlands), Jeroen de Vries (Amsterdam UMC location AMC, The Netherlands), and Bart van de Poel (Xomnia B.V., The Netherlands) for their contributions in obtaining the data, data curation, and data cleaning, respectively, in the initial stage of the project work in Chapter 4. We thank W. Visser-Groot and S.M. de Boer (Leiden University Medical Center, The Netherlands) for their contributions to the work in Chapter 6.

SIKS Dissertation Series No. 2025-16
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: deep learning, deformable image registration, cervical cancer, radiation treatment, organs at risk segmentation, landmark detection, multi-objective optimization, multi-objective learning
Printed by: www.proefschriftenprinten.nl
Front & Back: Monika Grewal, ChatGPT 4o, Brijesh Bhardwaj

ISBN/EAN: 978-94-6518-026-7

An electronic version of this dissertation is available at http://repository.tudelft.nl/.
The source code associated with this thesis is available in 4TU.ResearchData at:
Chapter 2: 10.4121/2df9d7ab-52bf-48f8-bcd3-fa821698e1ee
Chapter 3: 10.4121/16e31e65-7d5c-41e0-a2cc-fda7ed390082
Chapter 4: 10.4121/c4c2fb0a-7185-46ed-b4fb-856eba9c04c0
Chapter 5: 10.4121/d1299711-5e02-4b15-8d0f-82e9af98051b
Chapter 6: 10.4121/2d554736-63df-4dea-9387-5c24f19a0729

Contents

SUMMARY

Cervical cancer affects about half a million women globally every year. The treatment of cervical cancer with the aim of healing mainly consists of surgery, radiation treatment, or a combination of radiation treatment with chemotherapy or hyperthermia. Radiation treatment is a type of treatment wherein a high dose of ionizing radiation is used to kill the tumor cells. The radiation dose is usually delivered in the form of External Beam Radiation Treatment (EBRT) with a linear accelerator followed by internal radiation treatment (brachytherapy) during which a small radioactive source is passed through an applicator and needles that are placed temporarily nearby the cervix. EBRT typically spans several weeks with daily sessions (often referred to as fractions), whereas brachytherapy typically consists of three or four fractions based on one to three implantations. The aim of the radiation treatment is to provide effective radiation to kill the tumor cells while sparing the nearby healthy tissue or Organs At Risk (OARs) as much as possible. This is achieved by treatment planning following the contouring of target volumes and OARs, on medical imaging scans, which typically are Computed Tomography (CT) and/or Magnetic Resonance Imaging (MRI).

Deformable Image Registration (DIR) refers to aligning a source image to a target image by finding a Deformation Vector Field (DVF), which maps each voxel in the target image to a voxel in the source image. DIR can streamline the radiation treatment workflow by the automatic transfer of contours from one scan to another scan. Further applications of DIR in cervical cancer radiation treatment include image fusion (i.e., overlaying information from different imaging modalities to aid in delineation), dose accumulation, and online (e.g., just prior to, or during treatment) adaptation of radiation treatment. These applications can potentially provide benefits in terms of time management and quality of treatment.

Despite its potential benefits in radiation treatment, DIR is seldom used in clinical practice. **We identify three main reasons for the limited use of DIR in clinical practice.** First, traditional optimization methods for DIR take up to several hours for an entire pelvic scan, rendering them impractical to use in scenarios where time is critical (e.g., online adaptation of radiation treatment). Another major hindrance for clinical use of DIR is the underlying challenges in aligning two scans that differ due to factors like organ (bladder/rectum) filling, changes in patient anatomy due to e.g., weight change, surgery, or response to the treatment, and content mismatch due to e.g., presence or absence of gas pockets. The last factor affecting clinical adoption is the need for patient-specific adaptation of DIR algorithms (e.g., the need to set different hyperparameters each time to get a usable result for each patient) and quality assurance of DIR performance, which is challenging to automate because of the absence of the underlying ground truth.

To address the first issue mentioned in the above paragraph, we use deep learning - a prevalent technique in modern-day artificial intelligence. A deep artificial neural network can be trained to perform DIR using pairs of existing medical imaging scans. After training, the artificial neural network can predict the DVF for an unseen pair of scans within a couple of seconds, potentially making DIR also accessible for use cases where time is critical.

To improve DIR in the presence of underlying challenges, we present methods for generating additional guidance from images, which can be used to achieve a performance gain in DIR. First work in this direction is a deep learning approach for automatic detection of landmark correspondences in a pair of CT scans. In Chapter 2, we develop a novel deep learning approach, in which a neural network is trained to identify salient locations in a pair of two-dimensional medical images as corresponding landmarks. The neural network also provides a matching probability for each pair of landmarks in the given pair of images. In Chapter 3, we extend the approach in Chapter 2 to work on three-dimensional scans. Further, we use our trained deep neural network to find corresponding landmarks that can be used for additional guidance in Elastix - a state-of-the-art optimization-based approach and publicly available software for DIR. We demonstrate on a test dataset that the additional guidance from the corresponding landmarks obtained by our deep neural network helps improve the performance of DIR.

Further, in Chapter 4, we present a novel semi-supervised approach for training a deep neural network for automatic segmentation of medical images in case of partially annotated dataset. We apply this approach for automatic segmentation of four OARs (bowel bag, bladder, hips, and rectum) in cervical cancer radiation treatment, which can be utilized to provide additional guidance to DIR. The proposed approach makes efficient use of clinically available data. We show that with the proposed approach, state-of-the-art performance can be achieved even with a conventionally used baseline neural network for organ segmentation, UNet. Moreover, we demonstrate that the contours generated from segmentation masks provided by the trained neural network are clinically acceptable.

To address the challenge regarding patient-specific adaptation and quality assurance, we take a Multi-Objective (MO) perspective to DIR. MO optimization refers to optimization of two or more conflicting objectives simultaneously. This is typically done by finding a set of outputs corresponding to different trade-offs between conflicting objectives, such that no output is better than any other output in any objective without a simultaneous detriment in at least one other objective. This set of outputs can then be presented to decision-makers to make an a posteriori choice of the trade-off between conflicting objectives. In this thesis, we use the concept of MO optimization in DIR with a motivation to enable a posteriori decision making by clinicians. With an MO DIR approach, the clinicians are provided with a set of possible DIR outputs, each corresponding to a different trade-off between pre-determined performance metrics. The clinicians can evaluate the set of possiblities and choose the most appropriate output while also considering patient-specific criteria that were not part of DIR.

To this end, we develop a novel approach, which enables a posteriori MO decision making with deep neural networks in Chapter 5. Next, we use the approach developed in Chapter 5 for MO learning of DIR in Chapter 6. Specifically, we train a neural network multi-objectively to minimize three losses: normalized cross correlation loss, deformation smoothness loss, and Dice loss between the segmentation masks of OARs. We demonstrate in a proof-of-principle study that the proposed MO learning approach has potential benefits as compared to linear scalarization of different loss terms using weights sampled from a grid of possible weights.

In essence, the work in this thesis focuses on the use of deep learning to improve DIR, specifically for the use case of cervical cancer radiation treatment. In doing so, the work in this thesis is focused on two main directions: 1) developing methods for generating additional guidance for use in DIR, 2) enabling MO decision making with deep learning and applying it to DIR. In Chapter 7, we outline future directions of research inspired by the findings in this PhD thesis.

SAMENVATTING

Baarmoederhalskanker treft jaarlijks ongeveer een half miljoen vrouwen wereldwijd. De behandeling van baarmoederhalskanker met het doel om te genezen bestaat uit chirurgie, radiotherapie (ook wel bestraling genoemd), of een combinatie van bestraling met chemotherapie of hyperthermie. Radiotherapie is een behandelingsvorm waarbij een hoge dosis ioniserende straling wordt gebruikt om tumorcellen te doden. De stralingsdosis wordt normaliter toegediend in de vorm van uitwendige radiotherapie waarbij gebruik gemaakt wordt van een lineaire versneller, gevolgd door inwendige radiotherapie (brachytherapie) waarbij gebruik gemaakt wordt van een kleine radioactieve bron die door het lichaam geleid wordt via een applicator en interstitiële naalden, welke tijdelijk in de buurt van de baarmoederhals worden geïmplanteerd. Uitwendige radiotherapie duurt doorgaans meerdere weken met dagelijkse sessies (vaak fracties genoemd), terwijl brachytherapie meestal uit drie of vier fracties bestaat, met behulp van één tot drie implantaties. Het doel van de radiotherapie is om voldoende straling toe te dienen om de tumorcellen te doden, terwijl het omliggende gezonde weefsel of de risico organen zoveel mogelijk gespaard blijven. Hiertoe wordt een bestralingsplan gemaakt op basis van intekeningen van het doelgebied en de risico organen op beelden die gemaakt zijn met behulp van Computer Tomografie (CT), eventueel in combinatie met Magnetic Resonance Imaging (MRI) beelden.

Deformeerbare beeldregistratie (Deformable Image Registration (DIR)) verwijst naar het transformeren van een bronbeeld naar een doelbeeld door een vectorveld te vinden, dat elke voxel in het doelbeeld koppelt aan een voxel in het bronbeeld. DIR kan de workflow van radiotherapie vereenvoudigen door, onder andere, het propageren van contouren van de ene scan naar de andere scan te automatiseren. Andere mogelijke toepassingen van DIR binnen de radiotherapie voor baarmoederhalskanker zijn beeldfusie (d.w.z. het over elkaar leggen van beelden van verschillende modaliteiten ter ondersteuning van het maken van intekeningen), dosisaccumulatie en het online aanpassen van bestralingsplannen (bijv. net voor of tijdens de dosisafgifte). Deze toepassingen kunnen mogelijk voordelen bieden op het gebied van tijd efficiëntie en de kwaliteit van de behandeling.

Ondanks de potentiële voordelen van DIR in de radiotherapie, wordt het zelden gebruikt in de klinische praktijk. **We identificeren drie hoofdredenen voor het beperkte gebruik van DIR in de klinische praktijk.** Ten eerste nemen traditionele optimalisatiemethoden voor DIR tot enkele uren in beslag voor een volledige bekken-scan, wat DIR onpraktisch maakt in situaties waar tijd cruciaal is (bijv. online aanpassing van bestralingsplannen). Een andere belangrijke belemmering voor het klinisch gebruik van DIR zijn mogelijke onderliggende uitdagingen bij het registreren van twee scans die verschillen door factoren zoals orgaanvulling (blaas/rectum), veranderingen in de

anatomie van de patiënt door bijvoorbeeld gewichtsverandering, operatie, of reactie op de behandeling, en inhoudelijke verschillen zoals de aanwezigheid of afwezigheid van gasophopingen. De laatste factor die de klinische adoptie beïnvloedt, is de noodzaak om patiëntspecifieke aanpassingen te maken aan DIR-algoritmes (bijv. de noodzaak om elke keer verschillende hyperparameters in te stellen om een bruikbaar resultaat te verkrijgen voor een patiënt) en de kwaliteitsborging van de DIR-prestaties, wat moeilijk te automatiseren is door het ontbreken van een onderliggende 'ground truth'.

Om het eerste probleem uit de bovenstaande paragraaf aan te pakken, maken we gebruik van deep learning - een veelgebruikte techniek in het gebied van moderne kunstmatige intelligentie. Een diep kunstmatig neuraal netwerk kan worden getraind om DIR uit te voeren met behulp van bestaande paren van medische scans. Na training kan het neurale netwerk binnen enkele seconden het vectorveld van de vervorming voorspellen voor een ongezien paar scans, waardoor DIR potentieel ook geschikt zal zijn om in te zetten voor toepassingen waar tijd een cruciale factor is.

Om DIR te verbeteren in het geval van de aanwezigheid van onderliggende uitdagingen, presenteren we methoden voor het extraheren van extra begeleiding uit de beelden, wat kan worden gebruikt om een prestatieverbetering in DIR te bereiken. Het eerste werk in deze richting is een aanpak die op deep learning is gebaseerd voor automatische detectie van overeenkomende herkenningspunten in een verzameling van twee CT-scans. In Hoofdstuk 2 ontwikkelen we een nieuwe deep learning-aanpak, waarbij een neuraal netwerk wordt getraind om herkenningspunten in een set van twee tweedimensionale medische beelden te identificeren als corresponderende herkenningspunten. Het neurale netwerk biedt ook een waarschijnlijkheidsschatting voor elk paar herkenningspunten in de gegeven set van twee beelden. In Hoofdstuk 3 breiden we de aanpak uit Hoofdstuk 2 uit naar driedimensionale scans. Daarnaast gebruiken we ons getrainde diepe neurale netwerk om corresponderende herkenningspunten te vinden die kunnen worden gebruikt voor extra begeleiding in Elastix - een toonaangevende aanpak voor DIR die op optimalisatie gebaseerd is en in de vorm van software publiekelijk beschikbaar is. We tonen aan, op een testdataset, dat de extra begeleiding van de corresponderende herkenningspunten, verkregen door ons diepe neurale netwerk, helpt om de prestaties van DIR te verbeteren.

Verder presenteren we in Hoofdstuk 4 een nieuwe semi-gesuperviseerde aanpak voor het trainen van een diep neuraal netwerk voor de automatische segmentatie van medische beelden in het geval van gedeeltelijk geannoteerde datasets. We passen deze aanpak toe voor de automatische segmentatie van vier risico organen (darmzak, blaas, heupen en rectum) bij radiotherapie voor baarmoederhalskanker, wat kan worden gebruikt om extra begeleiding te bieden aan DIR. De voorgestelde aanpak maakt efficiënt gebruik van de klinisch beschikbare data. We tonen aan dat met deze aanpak state-of-the-art prestaties kunnen worden behaald, zelfs met een veelgebruikt standaard neuraal netwerk voor orgaandetectie, UNet. Bovendien laten we zien dat de contouren, gegenereerd uit de segmentatiemaskers die door het getrainde neurale netwerk worden geleverd, klinisch acceptabel zijn.

Om de uitdaging van de noodzaak om patiëntspecifieke aanpassingen te moeten doen en kwaliteitsborging aan te pakken, hanteren we een Multi-Objectief (MO) perspectief op DIR. MO-optimalisatie verwijst naar het gelijktijdig optimaliseren van twee of meer conflicterende doelen. Dit wordt meestal gedaan door een verzameling resultaten te vinden die verschillende afwegingen tussen conflicterende doelen vertegenwoordigen, waarbij geen enkel resultaat beter is in één doel dan een ander resultaat zonder dat dit ten koste gaat van één of meer andere doelen. Deze verzameling resultaten kan vervolgens worden voorgelegd aan behandelaars, die achteraf een keuze kunnen maken over de afweging tussen conflicterende doelen. In dit proefschrift gebruiken we het concept van MO-optimalisatie in DIR met als doel behandelaars in staat te stellen om achteraf beslissingen te nemen. Met een MO DIR-aanpak worden behandelaars voorzien van een verzameling van mogelijke DIR-resultaten, elk met een andere afweging tussen vooraf bepaalde prestatie-indicatoren. De behandelaars kunnen deze mogelijkheden evalueren en het meest geschikte resultaat kiezen, rekening houdend met patiëntspecifieke criteria die geen deel uitmaakten van de DIR.

Hiertoe ontwikkelen we een nieuwe aanpak die a posteriori MO-besluitvorming met diepe neurale netwerken mogelijk maakt in Hoofdstuk 5. Vervolgens gebruiken we de in Hoofdstuk 5 ontwikkelde aanpak voor het MO-leren van DIR in Hoofdstuk 6. Concreet trainen we een neuraal netwerk multi-objectief om drie verliezen te minimaliseren: een verlies dat betrekking heeft op de genormaliseerde kruiscorrelatie van de intensiteitswaarden van de voxels, de gladheid van het vectorveld dat de vervorming beschrijft, en de Dice waarden tussen de segmentatiemaskers van de risico organen. In een proof-of-principle studie tonen we aan dat de voorgestelde MO-leeraanpak potentiële voordelen biedt in vergelijking met het gebruik van een gewogen som van verschillende verliesfuncties met gewichten die worden gesampled uit een rooster.

Samengevat, het werk in dit proefschrift richt zich op het gebruik van deep learning om DIR te verbeteren, specifiek voor de toepassing van radiotherapie bij baarmoederhalskanker. Hierbij ligt de focus van dit werk op twee hoofdrichtingen: 1) het ontwikkelen van methoden voor het genereren van extra begeleiding voor gebruik in DIR, 2) het mogelijk maken van MO-besluitvorming met deep learning en dit toepassen op DIR. In Hoofdstuk 7 schetsen we toekomstige onderzoekslijnen geïnspireerd door de bevindingen in dit proefschrift.

1

INTRODUCTION

Deep learning – a prevalent artificial intelligence technique today, is transforming many application domains (e.g., natural language processing [46, 20], text-to-image generation [47, 40], and computer vision [52]). On the other hand, medical imaging techniques (e.g., Computed Tomography (CT), and Magnetic Resonance Imaging (MRI)), have revolutionized the medical field through advancements in diagnostics, treatment planning, and image-guided surgery. This thesis concerns the above-mentioned two fields: deep learning and medical imaging. Our specific focus is on the task of deformable image registration in cervical cancer radiation treatment. In this first chapter, we provide a background on the topics related to this thesis and further elaborate our motivation.

1.1. BACKGROUND

1.1.1. DEEP LEARNING

Deep learning is a subset of machine learning that focuses on building and training Artificial Neural Networks (ANNs) to learn from large amounts of data. Deep learning has gained significant attention and popularity due to its ability to automatically discover and learn complex patterns directly from raw data, without the need for explicit feature engineering. Deep Neural Networks (DNNs) consist of multiple processing layers, allowing them to learn hierarchical representations of input data. Each layer in a DNN captures increasingly abstract information from the input data enabling the DNN to perform human-like cognitive tasks. In recent years, deep learning has shown tremendous generalization ability to unseen data in computer vision tasks involving image classification, and image segmentation. Moreover, with advances in computer hardware, DNNs can analyze an image of 512 pixels × 512 pixels within a fraction of a second as opposed to several minutes (sometimes hours depending on the task) it takes for traditional image analysis algorithms. These advantages provide a strong motivation to use deep learning for analysis of medical images.

1

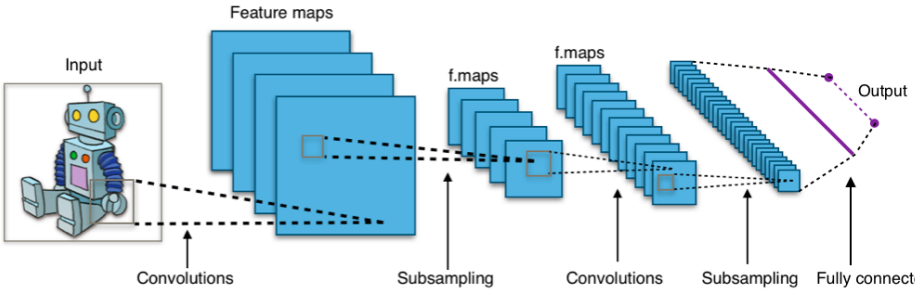


Figure 1.1: A typical fully connected convolutional neural network by Aphex34 used under (CC-BY-SA-4.0). The input is processed by 4 convolution kernels (each with a small field of view, shown with a gray square) in the first layer to compute 4 feature maps. This is followed by spatial sub-sampling to reduce the size of the feature maps. This is repeated in the subsequent layers to extract more complex features from the input image.

The fundamental building block of an ANN is an artificial neuron, often simply referred to as a neuron. It is a computational unit that mimics the behavior of biological neurons in the human brain. An artificial neuron receives inputs, processes these as a weighted sum using their corresponding weights, applies an activation function, and produces an output. Mathematically, this operation can be represented as:

$$\text{output} = \sigma(\sum_{i=1}^{n}(x_i \times w_i))$$

where x_i is the ith input, w_i is the weight associated with the ith input, n is the total number of inputs, and σ is the activation function. The activation function gives the artificial neurons the capability to model non-linearity between the inputs and output. A simple activation function can be activating a neuron (i.e., a non-zero output) only if the weighted sum is above a certain threshold. Common activation functions include sigmoid, hyperbolic tangent (tanh), and Rectified Linear Unit (ReLU) [43]. Often, one of the inputs (x_0) is a constant (value +1) and the weight term associated with it (w_0) is called a bias ($w_0 = b$). The bias term is used to shift the activation to either positive or negative direction, which is useful in modeling the relationship between the inputs and output better. A typical DNN consists of multiple layers, with each layer consisting of multiple neurons. Generally, the neurons in each subsequent layer take the outputs from all the neurons in the previous layer as input. In other words, the neurons in each subsequent layer are fully connected with the neurons in the previous layer.

Often, DNNs consist of a convolution kernel [33] as a fundamental building block, especially when developed for image processing tasks. Consequently, these DNNs are called Convolutional Neural Networks (CNNs). In Figure 1.1, a typical CNN architecture is depicted. CNNs have the following key features:

- **Local receptive field**, which means that the convolution kernel computes output corresponding to a small region (typically 3×3) in the input image. This allows extraction of translation-invariant features from different parts of an image.

- **Weight sharing**, which means that the same convolution kernel is used in different parts of the image. This allows for the processing of large images without requiring additional weights.
- **Spatial or temporal sub-sampling**, which means that after every couple of convolutional layers, the outputs (or feature maps) are downsampled by subsampling in a sliding window. This allows combining of features from previous layers to extract complex feature maps in each subsequent layer.

These features combined make CNNs an appropriate tool for image processing. A detailed understanding of CNNs, and how they work can be found in [19] and [44]. Since in this thesis we work with medical images, we will mainly use CNNs.

1.1.2. DEEP NEURAL NETWORK TRAINING

The development of a deep learning model involves training a DNN to predict outputs corresponding to a cognitive task for given inputs. The training of a DNN consists of three main steps: forward propagation, computing a loss and gradients, and backward propagation. During the forward propagation, the input data is fed into the neural network, and the outputs of each neuron in each layer are computed sequentially from the input layer to the output layer. After the forward propagation, the output of the neural network is compared to the true output (i.e., label) corresponding to the input data and a loss is computed. The quantitative value of the loss indicates how different the computed output from the neural network is from the true output. A simple example of loss is the mean squared error between the neural network output and the true label. In this phase, the gradient of the loss function with respect to each weight in the neural network is also computed. This is done using the chain rule for computing derivatives, which states that the derivative of $f(g(x))$ can be computed as $\frac{\partial f(g(x))}{\partial x} = \frac{\partial f(g(x))}{\partial g(x)} \cdot \frac{\partial g(x)}{\partial x}$. This allows us to compute the gradients of the loss function with respect to the weights in each layer in a backward direction, starting from the output layer to the input layer. This algorithm is called backward propagation [19]. In each training iteration, the weights of the neural network are updated by subtracting a scaled gradient (controlled by the learning rate) from the current weights, i.e., by performing gradient descent. Some of the popular gradient descent methods are Adam [29], and RMSprop [56]. This process is repeated iteratively for a fixed number of iterations or until the loss converges to a predefined satisfactory level.

After the training, a DNN can be used to compute (often referred to as predict) output for an unseen input, which was not used during the training. This phase is called testing or inference.

1.1.3. TYPES OF LEARNING PARADIGMS IN DEEP LEARNING

As described above, the training of DNNs requires pairs of inputs and corresponding outputs (or labels). To achieve good generalization to unseen data during inference, a typical DNN needs to be trained with thousands (or even millions) of pairs of inputs and labels (also referred to as data samples) representative of the task.

Depending upon how the labels corresponding to an input are generated and used in the training of a DNN, deep learning can be divided into the following types[1].

- **Supervised learning**: supervised learning refers to training the neural network with pairs of inputs and corresponding labels. This is the predominant type of learning technique used for developing deep learning algorithms. However, it turns out to be infeasible in many real-world scenarios because the labels are generally generated by manual annotation, which is time-consuming, expensive, and often prone to inter/intra-observer variance.
- **Semi-supervised learning**: semi-supervised learning refers to a scenario where the labels are available only for a portion of the training dataset. This type of learning is often used in scenarios where obtaining large labeled datasets is expensive or time-consuming. For example, an NN can be trained for organ segmentation from CT scans where the annotations for the organ contours are available only for a small subset of all the transverse slices [73].
- **Self-supervised learning**: the motivation for self-supervised learning also comes from the challenges involved in obtaining labels for training data. A typical form of this training involves the following steps: pre-train an NN on simulated data, use the trained NN to generate labels on the real-world data and train the NN using the resulting labeled real-world dataset. This type of training makes use of constraints on the predictions to generate reliable labels for the data using an NN. In recent years, this technique has been used to obtain good performance on different computer vision tasks without using manually annotated data [14, 67].
- **Unsupervised learning**: unsupervised learning refers to learning without labels, meaning that the NN has no access to explicit labels for each input. The objective of unsupervised learning is to learn patterns, or representations of the input data [5]. Common applications of unsupervised learning include clustering, anomaly detection, and dimensionality reduction.

1.1.4. DEFORMABLE IMAGE REGISTRATION (DIR)

Image registration is the task of aligning the content of two images by estimating a transformation that maps one image to the coordinate space of the other image. In the most common formulation of registration, one of the images is considered fixed and the other image is moved to align with the fixed image. The fixed image is also referred to as the 'target' or 'reference' image. The moving image is also referred to as the 'source' image. The transformation can be linear, affine, or non-linear. The linear and affine transformations are global and can be defined by a transformation matrix consisting of a few parameters. For example, the transformation between the image shown in Figure 1.2 (a) and (b) can be defined by a single parameter (angle of rotation in the clockwise or counterclockwise direction).

On the other hand, the image shown in Figure 1.2 (c) exhibits local changes with respect to the image shown in Figure 1.2 (a). Consequently, the transformation

[1] We only mention the deep learning paradigms relevant to this thesis.

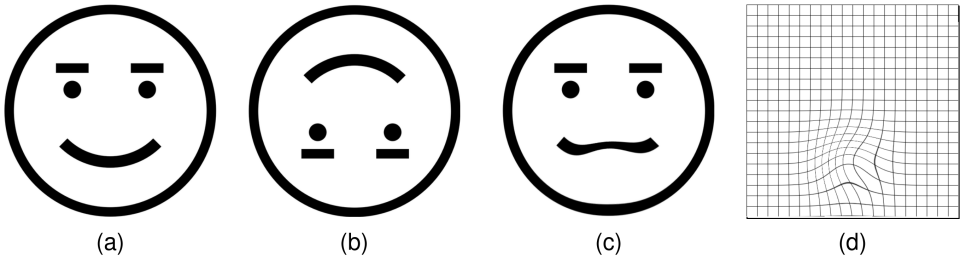


Figure 1.2: Transformation examples. (a) source image. (b)-(c) target images. The source image can be aligned with the target image in (b) by a rotation transformation of 180deg in the clockwise (or counterclockwise) direction. The source image in (a) needs to be transformed by a deformation vector field shown in (d) in order to align with the target image in (c).

describing these local changes should consist of deformation vectors defining displacements along the horizontal and vertical directions at each pixel of the image shown in Figure 1.2 (a). The task of aligning two images by finding such a transformation characterized by a Deformation Vector Field (DVF) is called Deformable Image Registration (DIR). In Figure 1.2 (d), the DVF aligning 1.2 (a) to 1.2 (c) is visualized through a deformed Cartesian grid.

DIR can be considered as an optimization task, wherein, typically a parameterized DVF is optimized iteratively by gradient-based approaches (e.g., Elastix [57, 38]) or Evolutionary Algorithm (EA) based approaches [1]. The parameters are continuously updated such that a metric representing image similarity is improved. Optimizing only for maximizing image similarity may yield a highly irregular or sometimes physically infeasible DVF. Therefore, the use of an additional objective corresponding to penalizing the deformation magnitude or irregularity is imperative to DIR. A typical DIR formulation optimizes the following objective.

$$Objective = ImageSimilarityObjective + \alpha DeformationPenaltyObjective \tag{1.1}$$

Here, *ImageSimilarityObjective* is a term associated with maximizing a specific image similarity metric, *DeformationPenaltyObjective* is associated with penalizing irregular deformations, and α is used to control the relative contribution of the two objectives towards the final objective value.

1.1.5. CERVICAL CANCER RADIATION TREATMENT

Radiation treatment is a type of treatment of cancer, which involves killing cancer cells by exposing them to high doses of ionizing radiation. The focus of this thesis is on radiation treatment for cervical cancer i.e., the growth of cancer cells or tumors in the cervix. Cervical cancer is the fourth leading cause of cancer death in women globally with an estimated 570,000 new cases and 311,000 deaths in 2018 [7]. Radiation treatment plays a crucial role in the management of cervical cancer, either as the

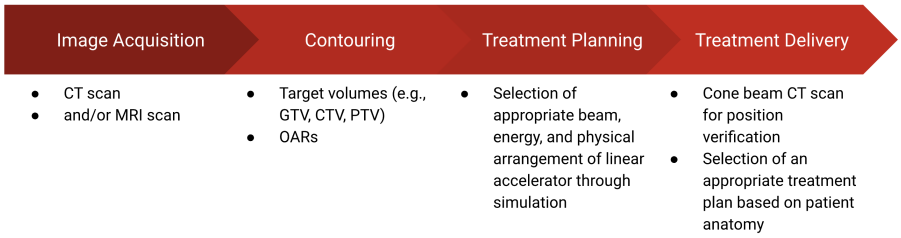


Figure 1.3: Different stages of the EBRT workflow. GTV: Gross Tumor volume, CTV: Clinical Target Volume, PTV: Planning Target Volume

primary treatment or in combination with surgery and/or chemotherapy. In the next paragraphs, we describe a typical radiation treatment workflow for locally advanced (FIGO - The International Federation of Gynecology and Obstetrics, stage IB3 and higher) cervical cancer.

The radiation treatment can be delivered by External Beam Radiation Treatment (EBRT) and brachytherapy. In EBRT, the patient lies on a treatment table, and the target volume is irradiated using highly focused radiation beams that originate from a machine called a linear accelerator (LINAC). In internal radiation treatment or brachytherapy, the source of radiation is brought into the body using an applicator inserted to the vaginal cavity and cervix/uterus. The treatment approach is determined by a radiation oncologist upon thorough evaluation, and considering factors such as the stage of the cancer, the location and size of the tumor, patient's overall health, and any previous treatments received.

The **different stages of the EBRT workflow** are described in Figure 1.3. Following diagnosis and subsequent consultation by the radiation oncologist, the patient undergoes **imaging** for treatment planning. The imaging primarily consists of Computed Tomography (CT) scans but sometimes additional Magnetic Resonance Imaging (MRI) scans are also acquired for better visualization of the target due to better soft-tissue contrast in the pelvis with MRI. Typically, different scans are acquired corresponding to different bladder filling levels. This is followed by **contouring** of the target volumes (i.e., Gross Tumor Volume (GTV), Clinical Target Volume (CTV), and Planning Target Volume (PTV)), and Organs At Risk (OARs) on the acquired scans. These contours are used for **treatment planning**, which involves simulation of radiation treatment for optimization of different parameters, e.g., beam, energy, and the physical arrangement of the LINAC such that the prescribed dose can be delivered to the tumor location while sparing the OARs. The **treatment delivery** for EBRT is typically done in 23-25 daily sessions, referred to as fractions. Each fraction lasts a few minutes. Before each fraction, a Cone Beam Computed Tomography (CBCT) scan is acquired to inspect the patient's internal anatomy. Based on the anatomy (e.g., bladder filling), an appropriate treatment plan is selected from the library of plans[2] made in the previous stage to deliver the radiation. In some cases where a patient's internal anatomy has

[2] Creating a library of treatment plans corresponding to different bladder fillings is a standard practice in the Netherlands Radiotherapy departments, but not worldwide.

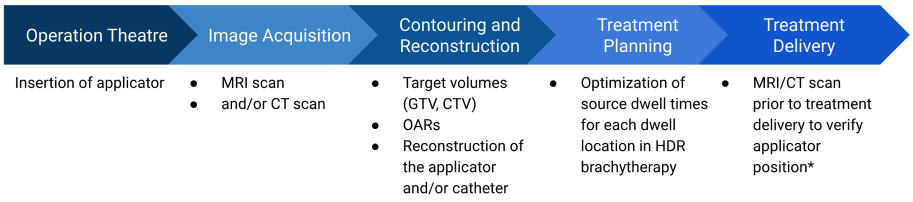


Figure 1.4: Different stages of the Brachytherapy workflow. *Re-imaging is performed only in case of multiple fractions with one applicator.

changed considerably from the beginning of treatment (when the treatment plan was made), the radiation oncology team adapts the original treatment plan.

In the treatment of locally advanced cancer, a brachytherapy boost is given after EBRT. Various forms of brachytherapy exist, including the permanent placement of low-strength radioactive sources (seeds) inside the body (normally indicated for prostate cancer), which is referred to as Low Dose Rate (LDR) brachytherapy as well as temporary Pulsed Dose Rate (PDR) or High Dose Rate (HDR) brachytherapy. Temporary brachytherapy involves the placement of applicators and/or hollow needles through which a lower-strength or a higher-strength radioactive source in case of PDR and HDR, respectively, can be guided. The latter is typically applied in several fractions. The **different stages of the temporary brachytherapy workflow** are described in Figure 1.4. In the first stage, the patient goes into the **operation theatre** for the insertion of an intra-cavitary applicator (and potentially also interstitial needles) into the cervix/uterus. The **image acquisition** stage typically involves making an MRI scan, which is followed by reconstruction of the applicator and **contouring** of target volumes and OARs. **Treatment planning** is done while the patient is waiting in bed on the nursery clinical board. It involves the optimization of the source dwell positions and dwell times in case of HDR and PDR brachytherapy. In case of multiple fractions within one application, re-imaging by CT or MRI is done prior to **treatment delivery** to check the position of the applicator and OARs [39].

1.1.6. DIR IN CERVICAL CANCER RADIATION TREATMENT

DIR can reduce manual workload in different stages of radiation treatment [9, 45, 55]. There are the following main use cases of DIR in cervical cancer radiation treatment.

- **Image fusion.** Sometimes during delineation of target volumes, the radiation oncologists refer to a different imaging modality (e.g., MRI) because different modalities exhibit different types of contrast between different tissue types. DIR can be employed for the fusion of images from different modalities to aid the radiation oncologists in the delineation of the target volumes [61].
- **Automatic contouring and contour propagation [10].** DIR is also useful for atlas based contouring of organs at risks [22]. With DIR, automatic contour propagation can be done between one planning CT and another planning CT in which the anatomy of the patient is different (e.g., a different bladder filling).

- **Adaptive radiation treatment [8, 53].** Through DIR, the contours for target volumes can be propagated from the planning CT onto the CBCT acquired on the day of radiation delivery to gauge the changes with respect to the planning CT. Based on this the radiation oncology team can decide if re-planning is required in case of substantial changes that may have occurred.
- **Dose accumulation [41, 11].** By registering images taken at different time points pertaining to when radiation treatment took place, radiation oncologists can accumulate the radiation dose delivered to the tumor and surrounding healthy tissues over the course of treatment. Such an information can further be used to better predict the chances of (severe) toxicity in different OARs.

1.1.7. EXISTING WORK

There has been extensive work on DIR of medical images [59, 31, 18]. Among optimization based approaches, SimpleITK [35, 70, 4], and Elastix [57, 38] are widely used for DIR including with the aim to be applied in radiation treatment. SimpleITK provides an open-source toolkit for medical image segmentation and registration along with other tools (e.g., reading and writing medical images in different formats) for processing medical images. Elastix provides a collection of algorithms specifically for registration (both rigid as well as deformable) in a modular design. Further, Bondar et al. [6] proposed a symmetric nonrigid registration method to handle large organ deformations in cervical cancer patients. Pirpinia et al. [50] proposed an approach for prone-to-supine breast MRI registration, another DIR scenario that involves large deformations.

Multiple deep learning based algorithms have been proposed for DIR in medical imaging. Heinrich et al. [23] proposed a deep learning approach for DIR by combining uncertainty estimates from supervoxel belief propagation. Eppenhof et al. [15] proposed a supervised learning approach for DIR by using random transformations as ground truth. Fan et al. [16] proposed to make use of ground-truth guidance using deformation fields obtained by an existing registration method to train a deep neural network for DIR. Yoo et al. [72] proposed a novel deep learning algorithm that combined a spatial transformer for image deformation and a convolutional autoencoder for unsupervised feature learning. Vos et al. [66] proposed a neural network for end-to-end DIR of two-dimensional (2D) images. The proposed neural network consisted of a CNN regressor to predict displacement on a grid of control points for a B-spline transformer. Balakrishnan et al. [3] proposed a deep neural network for unsupervised learning of DIR by making use of additional guidance from organ contours. Sokooti et al. [58] proposed a multi-scale three-dimensional (3D) CNN for DIR. De Vos et al. [13] and Stergios et al. [60] proposed deep neural networks for performing affine registration as well as DIR.

Other deep learning methods for DIR in medical imaging have focused on different aspects e.g., large diffeomorphic deformations [69], generative neural networks for DIR of cross-modality images [36, 62, 68], large motion specifically for lung registration [25, 24], and weak label supervision [26]. Deep learning based DIR methods have also been proposed for specific applications, e.g., Rigaud et al. [54] developed a DIR method for dose mapping between EBRT and brachytherapy images of cervical cancer.

Several commercial softwares e.g., MIMVista [28], Velocity [64], Raystation [32], and Mirada RTx [12] provide DIR capability for medical images. In the commercial softwares, DIR is also integrated with its different applications in the radiation treatment workflow e.g., automatic contouring, contour propagation, and multi-modal image fusion.

1.1.8. CHALLENGES IN DIR IN CERVICAL CANCER RADIATION TREATMENT

Cervical cancer radiation treatment involves CT and MRI scans of the pelvic anatomy. A coronal CT scan slice representing female abdominal and pelvic anatomy is shown in Figure 1.5. The pelvic anatomy contains organs e.g., small and large bowel, bladder, and rectum, which can demonstrate large inter- and intra-patient anatomical differences. Moreover, physical phenomena such as bladder filling, gas pockets in the bowel, tumor shrinkage, and insertion of an applicator in case of brachytherapy, pose challenges to DIR such as large deformations and content mismatch. Another factor contributing to low performance of DIR algorithms is low-contrast and homogeneous tissue regions [71]. Due to these challenges, image similarity is not always sufficient to guide the deformations. Consequently, the DIR methods often get stuck in local minima, producing sub-optimal DIR solutions.

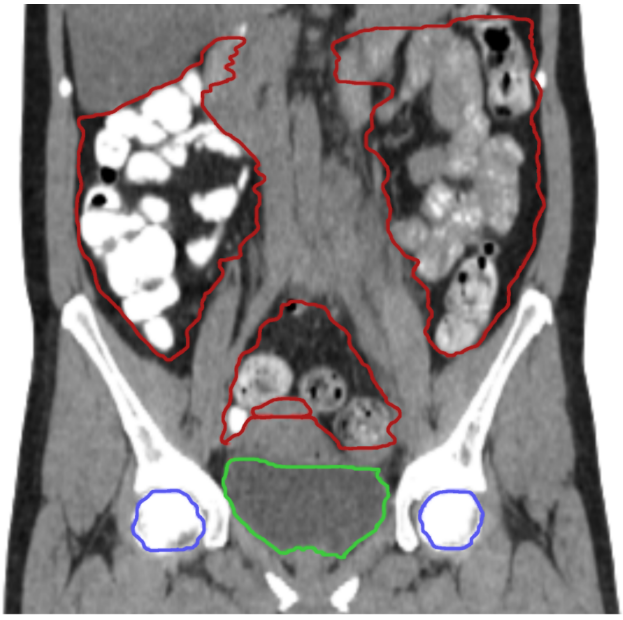


Figure 1.5: Coronal slice of a female abdomen and pelvic CT scan. The contours correspond to major organs visible in the scan. The contours in red, green, and blue correspond to bowel, bladder, and hips, respectively.

Apart from the abovementioned challenges, there exist two more key challenges. The first challenge is that the pelvic scans are large. A typical pelvic CT scan consists of $512 \times 512 \times 150$ voxels. This means optimization of $\approx$ 100 million parameters to obtain a high-resolution DVF, which is time-consuming. This hinders the application of DIR in use-cases where time is critical e.g., online adaptive radiotherapy. The second key challenge is that a DIR algorithm created for one application is not suitable for all applications [30]. Earlier studies have also reported that a DIR algorithm needs to

be adapted specific to each patient or image pair [50, 49]. Furthermore, the lack of knowledge on the actual deformation that occurred between the two images makes quality-assurance at each registration level difficult [9].

1.1.9. ADDITIONAL GUIDANCE CAN HELP DIR

Because the optimization problem associated with DIR can be complex and may exhibit many local optima, especially in cases where the field of view includes many different anatomical structures, a DIR optimization algorithm may get stuck in local minima. To steer the optimization algorithm more effectively, the optimization algorithm can be given additional information regarding what can be observed in the pair of images. This additional information may be seen as additional guidance to the optimization algorithm [1]. Many earlier research works have indicated the beneficial effect of using additional guidance in the improvement of DIR [65, 17, 51, 1, 21]. The following two types of additional guidance are most commonly used.

- **Organs At Risk (OARs) Contours:** DIR optimization may be constrained to find a DVF that sufficiently matches the contours of different structures in the given pair of images. Typically, in the case of images for radiation treatment, these structures can be the OARs, because they are contoured as a part of radiation treatment planning. Using the constraint of matching the OARs contours can potentially increase the potential of finding large deformations without getting stuck in a local minima.
- **Corresponding Landmarks:** Alternatively, information about anatomical landmarks or corresponding keypoints[3] in both the source and target image may be provided. The optimization algorithm may be constrained to find a DVF that sufficiently aligns the corresponding landmarks in the source and target images.

Similar to Equation 1.1, traditionally, taking additional guidance into account, is established by introducing additional terms into the objective function as given in the equation below.

$$\begin{aligned} \mathit{Objective} = \mathit{ImageSimilarityObjective} + \alpha\, \mathit{DeformationPenaltyObjective} + \\ \beta\, \mathit{AdditionalGuidanceObjective} \end{aligned} \tag{1.2}$$

Here, *AdditionalGuidanceObjective* aims to maximize similarity based on additional guidance information (i.e., landmarks or contours), β is a weight term that controls the relative contribution of the additional guidance objective toward the final objective value.

1.1.10. DIR IS MULTI-OBJECTIVE

Multi-objective (MO) optimization refers to the subfield of optimization where two or more conflicting objectives need to be optimized simultaneously. Many real-world problems are multi-objective, for example, designing an e-commerce recommendation

[3] Corresponding keypoints refer to points on the source and target images that represent the same anatomical locations.

algorithm [34], optimization of radiation treatment plans [37, 42], and optimization of rocket engine pumps [48]. In MO optimization, the aim is to find a set of Pareto optimal solutions corresponding to diverse trade-offs between the conflicting objectives. A solution is considered Pareto optimal if none of the objectives can be improved without a simultaneous detriment in performance with respect to at least one of the other objectives [63]. The set consisting of all Pareto optimal solutions is called the Pareto set and its mapping to the objective space is called the Pareto front. Sometimes, the decision-makers know the desired trade-off between the conflicting objectives beforehand. However, more often the preference between the possible trade-offs becomes clear only after the best possible options become known. This is called a posteriori decision-making [27]. In such a scenario, the Pareto set can be presented to the decision-makers to make an informed decision a posteriori, based on their preference and potentially taking into consideration additional objectives that were not part of the optimization.

DIR is inherently a multi-objective problem, because it involves optimizing multiple objectives, which may be conflicting [2, 1]. The objective of the deformation penalty inherently conflicts with the objective of maximizing image similarity. Further, ideally, additional guidance should always improve DIR performance. However, in practice, the additional guidance may be less reliable due to inter/intra-observer variability, and manual error. If the additional guidance is generated through automatic methods, it may be erroneous due to the limitations of the algorithm used for automatic generation. Therefore, in practice, the additional guidance may conflict with the image similarity objective.

Moreover, too much importance on the additional guidance objective may cause overfitting to the regions where additional guidance is available, resulting in deteriorating registration performance in other regions [3]. Therefore, it is important to tune the weight factor β carefully. Manually tuning the weight terms α and β to balance the individual contribution of different (conflicting) objectives towards the final objective value is non-trivial and time-consuming. Further, previous research has shown that the optimal weights may be different for each pair of source and target images [50]. In this scenario, modeling DIR with an MO perspective to find multiple solutions corresponding to diverse trade-offs between different objectives is an intuitive choice. This gives an automatic way to tune the relative contribution of different objectives.

1.2. MAIN CONTRIBUTIONS

In this thesis, we address the abovementioned challenges in DIR and develop algorithms to improve DIR. We use three main components: deep learning, additional guidance, and a multi-objective perspective. The use of deep learning is motivated by good generalization capability of deep neural networks on new unseen data. Another motivation to develop deep learning based algorithms is their fast inference time, which makes them applicable in time-critical scenarios as well.

Further, we develop deep learning methods for the automatic generation of additional guidance to address the issue of poor performance of DIR methods due to complex physical phenomena, or low intensity-contrast. To address the challenge of overcoming the need to make patient-specific adaptations to the DIR method, we take a multi-objective perspective to DIR. Another motivation to take an MO perspective to DIR is its potential for faster clinical adoption. By taking an MO perspective with a posteriori decision-making, the final decision is left to the clinical experts. It allows the clinical experts to a posteriori evaluate the different trade-offs on the DIR solutions and select the most appropriate DIR solution for the application at hand while taking into consideration other clinical parameters that were not explicitly modeled.

The main contributions of this thesis are the following.

1. In Chapter 2 and Chapter 4, we describe deep learning based algorithms for the automated detection of corresponding landmarks and segmentation of OARs, which can be used to provide additional guidance to DIR. In Chapter 2, we propose a self-supervised learning approach to automatically detect corresponding landmarks in a pair of two-dimensional (2D) CT images. In Chapter 4, we propose an iterative teacher-student training approach for deep neural networks to learn to segment OARs on CT scans using a partially annotated large dataset.

2. In Chapter 3, we extend the the approach to automatically detecting corresponding landmarks from Chapter 2 to work on three-dimensional (3D) images. Further, we use the identified corresponding landmarks to provide additional guidance to DIR. We investigate how the spatial density and matching accuracy of corresponding landmarks affect the performance of DIR of intra-patient CT scans. Further, we investigate the generalization capability of the trained model on cross-modality data without retraining.

3. In Chapter 5, we develop an approach for making a posteriori decision-making in a multi-objective context possible with deep learning. We refer to this approach as 'MO learning'. Briefly stated, with this approach, multiple neural networks (or a single multi-headed neural network) can be trained multi-objectively on a set of conflicting losses. After training, the neural networks provide multiple outputs for a given input, which each represent a different trade off between the conflicting objectives.

4. In Chapter 6, we use the MO learning approach described in Chapter 5 for the task of DIR. We investigate the added benefits of the so-developed MO DIR approach as compared to a scenario where only a single DIR output is provided in the case of DIR of MRI scans from two different fractions of brachytherapy for cervical cancer.

BIBLIOGRAPHY

[1] Tanja Alderliesten, Peter A. N. Bosman, and Arjan Bel. "Getting the most out of additional guidance information in deformable image registration by leveraging multi-objective optimization". In: *Medical Imaging 2015: Image Processing*. Vol. 9413. Proc. SPIE. International Society for Optics and Photonics. 2015, 94131R.

[2] Tanja Alderliesten, Jan-Jakob Sonke, and Peter A. N. Bosman. "Multi-objective optimization for deformable image registration: proof of concept". In: *Medical Imaging 2012: Image Processing*. Vol. 8314. SPIE. 2012, pp. 594–600.

[3] Guha Balakrishnan et al. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 38.8 (2019), pp. 1788–1800. DOI: 10.1109/TMI.2019.2897538.

[4] Richard Beare, Bradley Lowekamp, and Ziv Yaniv. "Image segmentation, registration and characterization in R with SimpleITK". In: *Journal of Statistical Software* 86.8 (2018), pp. 1–35.

[5] Yoshua Bengio. "Deep learning of representations for unsupervised and transfer learning". In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. JMLR Workshop and Conference Proceedings. 2012, pp. 17–36.

[6] Luiza Bondar et al. "A symmetric nonrigid registration method to handle large organ deformations in cervical cancer patients". In: *Medical Physics* 37.7, Part 1 (2010), pp. 3760–3772.

[7] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians* 68.6 (2018), pp. 394–424. DOI: https://doi.org/10.3322/caac.21492. eprint: https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21492. URL: https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21492.

[8] Kristy K. Brock. "Adaptive Radiotherapy: Moving Into the Future". In: *Seminars in Radiation Oncology* 29.3 (2019). Adaptive Radiotherapy and Automation, pp. 181–184. ISSN: 1053-4296. DOI: https://doi.org/10.1016/j.semradonc.2019.02.011. URL: https://www.sciencedirect.com/science/article/pii/S1053429619300207.

[9] Kristy K. Brock et al. "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132". In: *Medical Physics* 44.7 (2017), e43–e76. DOI: https://doi.org/10.1002/mp.12256. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12256. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12256.

[10] Christina Hunter Chapman et al. “Deformable image registration–based contour propagation yields clinically acceptable plans for MRI-based cervical cancer brachytherapy planning”. In: *Brachytherapy* 17.2 (2018), pp. 360–367.

[11] Indrin J Chetty and Mihaela Rosu-Bubulac. “Deformable registration for dose accumulation”. In: *Seminars in Radiation Oncology*. Vol. 29. 3. Elsevier. 2019, pp. 198–208.

[12] Canon Medical Systems Corporation. *MIRADA RTx*. URL: https://www.mirada-medical.com/miradartx.

[13] Bob D De Vos et al. “A deep learning framework for unsupervised affine and deformable image registration”. In: *Medical Image Analysis* 52 (2019), pp. 128–143.

[14] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. “Superpoint: Self-supervised interest point detection and description”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 224–236.

[15] Koen A. J. Eppenhof et al. “Deformable image registration using convolutional neural networks”. In: *Medical Imaging 2018: Image Processing*. Ed. by Elsa D. Angelini and Bennett A. Landman. Vol. 10574. International Society for Optics and Photonics. SPIE, 2018, 105740S. DOI: 10.1117/12.2292443. URL: https://doi.org/10.1117/12.2292443.

[16] Jingfan Fan et al. “BIRNet: Brain image registration using dual-supervised fully convolutional networks”. In: *Medical Image Analysis* 54 (2019), pp. 193–206. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2019.03.006. URL: https://www.sciencedirect.com/science/article/pii/S1361841519300283.

[17] Zeinab Ghassabi et al. “An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors”. In: *EURASIP Journal on Image and Video Processing* 2013.1 (2013), p. 25. ISSN: 1687-5281. DOI: 10.1186/1687-5281-2013-25. URL: https://doi.org/10.1186/1687-5281-2013-25.

[18] Soumya Ghose et al. “A review of segmentation and deformable registration methods applied to adaptive cervical cancer radiation therapy treatment planning”. In: *Artificial Intelligence in Medicine* 64.2 (2015), pp. 75–87.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[20] Grammarly Inc. *Grammarly*. 2023. URL: https://app.grammarly.com/.

[21] Dong Han et al. “Robust anatomical landmark detection with application to MR brain image registration”. In: *Computerized Medical Imaging and Graphics* 46 (2015), pp. 277–290. ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.compmedimag.2015.09.002. URL: http://www.sciencedirect.com/science/article/pii/S089561111500124X.

[22] Xiao Han et al. “Atlas-based auto-segmentation of head and neck CT images”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008: 11th International Conference, New York, NY, USA, September 6-10, 2008, Proceedings, Part II 11*. Springer. 2008, pp. 434–441.

[23] Mattias P Heinrich et al. “Deformable image registration by combining uncertainty estimates from supervoxel belief propagation”. In: *Medical Image Analysis* 27 (2016), pp. 57–71.

[24] Alessa Hering, Bram van Ginneken, and Stefan Heldmann. "mlVIRNET: Multilevel Variational Image Registration Network". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Cham: Springer International Publishing, 2019, pp. 257–265. ISBN: 978-3-030-32226-7.

[25] Alessa Hering and Stefan Heldmann. "Unsupervised learning for large motion thoracic CT follow-up registration". In: *Medical Imaging 2019: Image Processing*. Vol. 10949. SPIE. 2019, pp. 331–337.

[26] Alessa Hering et al. "Memory-efficient 2.5 D convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans". In: *International Journal of Computer Assisted Radiology and Surgery* 14 (2019), pp. 1901–1912.

[27] C-L Hwang and Abu Syed Md Masud. *Multiple objective decision making—methods and applications: a state-of-the-art survey*. Vol. 164. Springer Science & Business Media, 2012.

[28] MIM Software Inc. *MIM Maestro*. URL: https://www.mimsoftware.com/radiationoncology/maestro.

[29] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations*. 2015. URL: http://arxiv.org/abs/1412.6980.

[30] Neil Kirby et al. "The need for application-based adaptation of deformable image registration". In: *Medical Physics* 40.1 (2013), p. 011702.

[31] Stefan Klein, Marius Staring, and Josien P. W. Pluim. "Evaluation of Optimization Methods for Nonrigid Medical Image Registration Using Mutual Information and B-Splines". In: *IEEE Transactions on Image Processing* 16.12 (2007), pp. 2879–2890. DOI: 10.1109/TIP.2007.909412.

[32] RaySearch Laboratories. *RayStation*. URL: https://www.raysearchlabs.com/raystation/.

[33] Yann LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[34] Xiao Lin et al. "A Pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation". In: *Proceedings of the 13th ACM Conference on Recommender Systems*. 2019, pp. 20–28.

[35] Bradley C Lowekamp et al. "The design of SimpleITK". In: *Frontiers in Neuroinformatics* 7 (2013), p. 45.

[36] Dwarikanath Mahapatra et al. "Deformable medical image registration using generative adversarial networks". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1449–1453.

[37] Stefanus C Maree et al. "Evaluation of bi-objective treatment planning for high-dose-rate prostate brachytherapy—A retrospective observer study". In: *Brachytherapy* 18.3 (2019), pp. 396–403.

[38] K. Marstal et al. "SimpleElastix: A user-friendly, multi-lingual library for medical image registration". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 574–582.

[39] Geetha Menon et al. "Brachytherapy workflow for locally advanced cervical cancer: A survey of Canadian Medical Physicists". In: *Brachytherapy* 21.4 (2022),

pp. 405–414. ISSN: 1538-4721. DOI: https://doi.org/10.1016/j.brachy.2022.03.003. URL: https://www.sciencedirect.com/science/article/pii/S1538472122000393.

[40] Midjourney. *Midjourney*. Version 5. 2023. URL: https://www.midjourney.com/home.

[41] R. Mohammadi et al. "Evaluation of deformable image registration algorithm for determination of accumulated dose for brachytherapy of cervical cancer patients". In: *Journal of Contemporary Brachytherapy* 11 (5 2019), pp. 469–478.

[42] B. Müller et al. "Multicriteria plan optimization in the hands of physicians: a pilot study in prostate cancer and brain tumors". In: *Radiation Oncology* 12 (2017).

[43] Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 807–814.

[44] Andrew Ng. *Lecture Supplementary Material, CS230 Deep Learning*. https://cs230.stanford.edu/. 2023.

[45] Seungjong Oh and Siyong Kim. "Deformable image registration in radiation therapy". In: *Radiation Oncology* 35.2 (2017), p. 101.

[46] OpenAI. *ChatGPT*. Version 3.5. 2023. URL: https://chat.openai.com/.

[47] OpenAI. *DALL-E*. Version 2. 2023. URL: https://openai.com/dall-e-2.

[48] Akira Oyama and Meng-Sing Liou. "Multiobjective optimization of rocket engine pumps using evolutionary algorithm". In: *Journal of Propulsion and Power* 18.3 (2002), pp. 528–535.

[49] Chiara Paganelli et al. ""Patient-specific validation of deformable image registration in radiation therapy: Overview and caveats"". In: *Medical Physics* 45.10 (2018), e908–e922. DOI: https://doi.org/10.1002/mp.13162. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13162. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13162.

[50] Kleopatra Pirpinia et al. "The feasibility of manual parameter tuning for deformable breast MR image registration from a multi-objective optimization perspective". In: *Physics in Medicine & Biology* 62.14 (2017), p. 5723.

[51] Thomas Polzin et al. "Combining automatic landmark detection and variational methods for lung CT registration". In: *Fifth International Workshop on Pulmonary Image Analysis*. 2013, pp. 85–96.

[52] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 779–788.

[53] Bastien Rigaud et al. "Automatic Segmentation Using Deep Learning to Enable Online Dose Optimization During Adaptive Radiation Therapy of Cervical Cancer". In: *International Journal of Radiation Oncology*Biology*Physics* 109.4 (2021), pp. 1096–1110. ISSN: 0360-3016. DOI: https://doi.org/10.1016/j.ijrobp.2020.10.038. URL: https://www.sciencedirect.com/science/article/pii/S0360301620344849.

[54] Bastien Rigaud et al. "Deformable image registration for dose mapping between external beam radiotherapy and brachytherapy images of cervical cancer". In: *Physics in Medicine & Biology* 64.11 (2019), p. 115023.

[55] Bastien Rigaud et al. “Deformable image registration for radiation therapy: principle, methods, applications and evaluation”. In: *Acta Oncologica* 58.9 (2019). PMID: 31155990, pp. 1225–1237. DOI: 10.1080/0284186X.2019.1620331. eprint: https://doi.org/10.1080/0284186X.2019.1620331. URL: https://doi.org/10.1080/0284186X.2019.1620331.

[56] Sebastian Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016).

[57] Denis P Shamonin et al. “Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer’s disease”. In: *Frontiers in Neuroinformatics* 7.50 (2014), pp. 1–15.

[58] Hessam Sokooti et al. “Nonrigid image registration using multi-scale 3D convolutional neural networks”. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*. Ed. by Maxime Descoteaux et al. Cham: Springer International Publishing, 2017, pp. 232–239. ISBN: 978-3-319-66182-7.

[59] Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. “Deformable Medical Image Registration: A Survey”. In: *IEEE Transactions on Medical Imaging* 32.7 (2013), pp. 1153–1190. DOI: 10.1109/TMI.2013.2265603.

[60] Christodoulidis Stergios et al. “Linear and deformable image registration with 3D convolutional neural networks”. In: *Image Analysis for Moving Organ, Breast, and Thoracic Images: Third International Workshop, RAMBO 2018, Fourth International Workshop, BIA 2018, and First International Workshop, TIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 3*. Springer. 2018, pp. 13–22.

[61] Lauren M. Tait et al. “The use of MRI deformable image registration for CT-based brachytherapy in locally advanced cervical cancer”. In: *Brachytherapy* 15.3 (2016), pp. 333–340. ISSN: 1538-4721. DOI: https://doi.org/10.1016/j.brachy.2016.01.002. URL: https://www.sciencedirect.com/science/article/pii/S1538472116000040.

[62] Christine Tanner et al. “Generative adversarial networks for MR-CT deformable image registration”. In: *arXiv preprint arXiv:1807.07349* (2018).

[63] David A Van Veldhuizen and Gary B Lamont. “Multiobjective evolutionary algorithms: Analyzing the state-of-the-art”. In: *Evolutionary Computation* 8.2 (2000), pp. 125–147.

[64] Inc. Varian Medical Systems. *Velocity*. Version 4.1. URL: https://www.varian.com/products/software/information-systems/velocity.

[65] Eliana M. Vásquez Osorio et al. “Accurate CT/MR vessel-guided nonrigid registration of largely deformed livers”. In: *Medical Physics* 39.5 (2012), pp. 2463–2477. DOI: 10.1118/1.3701779. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.3701779. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3701779.

[66] Bob D de Vos et al. “End-to-end unsupervised deformable image registration with a convolutional neural network”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2017, pp. 204–212.

[67] Ke Yan et al. “SAM: Self-Supervised Learning of Pixel-Wise Anatomical Embeddings in Radiological Images”. In: *IEEE Transactions on Medical Imaging* 41.10 (2022), pp. 2658–2669. DOI: 10.1109/TMI.2022.3169003.

[68] Qianye Yang et al. “Mri cross-modality neuroimage-to-neuroimage translation”. In: *arXiv preprint arXiv:1801.06940* (2018).

[69] Xiao Yang et al. “Quicksilver: Fast predictive image registration – A deep learning approach”. In: *NeuroImage* 158 (2017), pp. 378–396. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2017.07.008. URL: https://www.sciencedirect.com/science/article/pii/S1053811917305761.

[70] Ziv Yaniv et al. “SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research”. In: *Journal of Digital Imaging* 31.3 (2018), pp. 290–303.

[71] U. J. Yeo et al. “Performance of 12 DIR algorithms in low-contrast regions for mass and density conserving deformation”. In: *Medical Physics* 40.10 (2013), p. 101701. DOI: https://doi.org/10.1118/1.4819945. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.4819945. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.4819945.

[72] Inwan Yoo et al. “ssEMnet: Serial-Section Electron Microscopy Image Registration Using a Spatial Transformer Network with Learned Features”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by M. Jorge Cardoso et al. Cham: Springer International Publishing, 2017, pp. 249–257. ISBN: 978-3-319-67558-9.

[73] Hao Zheng et al. “Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 802–812.

2

AUTOMATIC LANDMARKS CORRESPONDENCE DETECTION IN 2D

Anatomical landmark correspondences in medical images can provide additional guidance information for the alignment of two images, which, in turn, is crucial for many medical applications. However, manual landmark annotation is labor-intensive. Therefore, we propose an end-to-end deep learning approach to automatically detect landmark correspondences in pairs of two-dimensional (2D) images. Our approach consists of a Siamese neural network, which is trained to identify salient locations in images as landmarks and predict matching probabilities for landmark pairs from two different images. We trained our approach on 2D transverse slices from 168 lower abdominal Computed Tomography (CT) scans. We tested the approach on 22,206 pairs of 2D slices with varying levels of intensity, affine, and elastic transformations. The proposed approach finds an average of 639, 466, and 370 landmark matches per image pair for intensity, affine, and elastic transformations, respectively, with spatial matching errors of at most 1 mm. Further, more than 99% of the landmark pairs are within a spatial matching error of 2 mm, 4 mm, and 8 mm for image pairs with intensity, affine, and elastic transformations, respectively. To investigate the utility of our developed approach in a clinical setting, we also tested our approach on pairs of transverse slices selected from follow-up CT scans of three patients. Visual inspection of the results revealed landmark matches in both bony anatomical regions as well as in soft tissues lacking prominent intensity gradients.

The content of this chapter is based on the following publication: **Grewal, M.**, Deist, T. M., Wiersma, J., Bosman, P. A. N., & Alderliesten, T. (2020, March). *An End-to-End Deep Learning Approach for Landmark Detection and Matching in Medical Images.* In Medical Imaging 2020: Image Processing (Vol. 11313, pp. 548-557). SPIE.

2.1. INTRODUCTION

Deformable Image Registration (DIR) can be extremely valuable in workflows related to image-guided diagnostics and treatment planning. However, DIR in medical imaging can be challenging due to large anatomical variations between images. This is particularly the case in the lower abdomen, where internal structures can undergo large deformations between two scans of a patient due to physical conditions like the presence of gas pockets and bladder filling. Such scenarios are particularly challenging for intensity-based registration, as there are many local optima to overcome. Landmark correspondences between images can provide additional guidance information to the DIR methods [1, 11] and increase the probabilty of finding the right transformation by adding landmark matches as an additional constraint or objective in the optimization. Since the manual annotation of anatomical landmarks is labor-intensive and requires expertise, developing methods for finding landmark correspondences automatically has great potential benefits.

The existing methods [25, 24, 19, 8, 4] for obtaining landmark correspondences in medical images are based on large and time-consuming pipelines that involve identifying landmark locations followed by matching local feature descriptors [10] within a restricted neighborhood. These methods rely upon multiple pre- and post-processing steps, multi-resolution search, and manual checking to achieve robustness; each step adding more heuristics and empirical hyperparameters to an already complex pipeline. Further, existing methods for landmark detection that restrict the definition of landmarks to certain intensity gradient patterns specific to the underlying data set or anatomical region may not be easily adaptable to other contexts [12]. Generalizing the definition of landmarks and reducing the number of heuristics would allow for faster adaptation of automated methods for different clinical settings. In addition, faster execution times for landmark detection and matching could benefit their clinical application.

Recently, deep Convolutional Neural Networks (CNNs) have shown promising results for classification and segmentation tasks in medical imaging due to their capability of learning discriminant feature descriptors from raw images [18, 9, 6]. There exist a few deep learning approaches for finding landmarks in medical images [2, 22]. However, in these approaches a neural network is trained in a supervised manner to learn a small number of manually annotated landmarks. It is to be noted that a high density of landmark correspondences is desirable to effectively provide additional guidance to the DIR methods. In a supervised setting, it means annotating thousands of landmarks per CT scan, which is intractable in terms of required manual efforts. On the other hand, many deep learning approaches have been developed for automatically finding object landmarks in natural images [20, 26, 7, 5] that do not require manual annotations. Some of these approaches focus on discovering a limited number of landmarks in an image dataset. Whereas, others either fine-tune a pre-trained network or make use of incremental training in a self-supervised fashion.

Our proposed approach is based on the above-mentioned approaches developed for natural images and tailored to meet the specific requirements relating to the

medical images. We propose a two-headed Siamese neural network that based on a pair of images simultaneously predicts the landmarks and their feature descriptors corresponding to each image. These are then sent to another module to predict their matching probabilities. We train the neural network from scratch and gradients are back-propagated from end-to-end. To the best of our knowledge, this is first endeavour to develop an end-to-end deep learning approach for finding landmark correspondences in medical images. Our approach has the following distinct advantages compared to existing methods for finding landmark correspondences:

- Our approach is end-to-end deep learning based; therefore, the need for data pre- and post-processing during inference is avoided. In addition, the proposed approach is faster at run-time and has fewer hyperparameters than traditional approaches.
- We do not impose any prior on the definition of a landmark in an image. Instead, we train the network in a way that the landmarks represent salient regions in the image that can be found repeatedly despite potential intensity variations, and deformations.
- The proposed approach does not require manual annotations for training and learns from data in a self-supervised manner.
- Our approach improves over the existing approaches for natural images by avoiding the need for pre-training, or incremental fine-tuning of the neural network.

2.2. DATA

In total 222 lower abdominal Computed Tomography (CT) scans of female patients acquired for radiation treatment planning purposes were retrospectively included: 168 scans (24,923 two-dimensional (2D) slices) were used for training and 54 scans (7,402 2D slices) were used for testing. For a separate set of three patients, one original scan along with a follow-up CT scan was included. The scans of these three patients were used for testing the approach in a clinical setting. All CT scans had an in-plane resolution from 0.91 mm × 0.91 mm to 1.31 mm × 1.31 mm. All the 2D slices were resampled to 1 mm × 1 mm in-plane resolution.

2.3. APPROACH

In Figure 2.1, the different modules of our approach are illustrated along with the data flow between them. Our approach comprises a Siamese architecture consisting of *CNN branches* with shared weights. The outputs of the CNN branches are sent to a module named *Sampling Layer* followed by another module named *Feature Descriptor Matching Module*. The network takes two images I_1 and I_2 as inputs and predicts K_1 and K_2 landmarks in I_1 and I_2, respectively. In addition, the network predicts matching probabilities ($\hat{c}_{i,j}$) for each landmark $i \in \{1,2,...,K_1\}$ in I_1 to a landmark $j \in \{1,2,...,K_2\}$ in I_2. In the following paragraphs, a description of each module is provided.

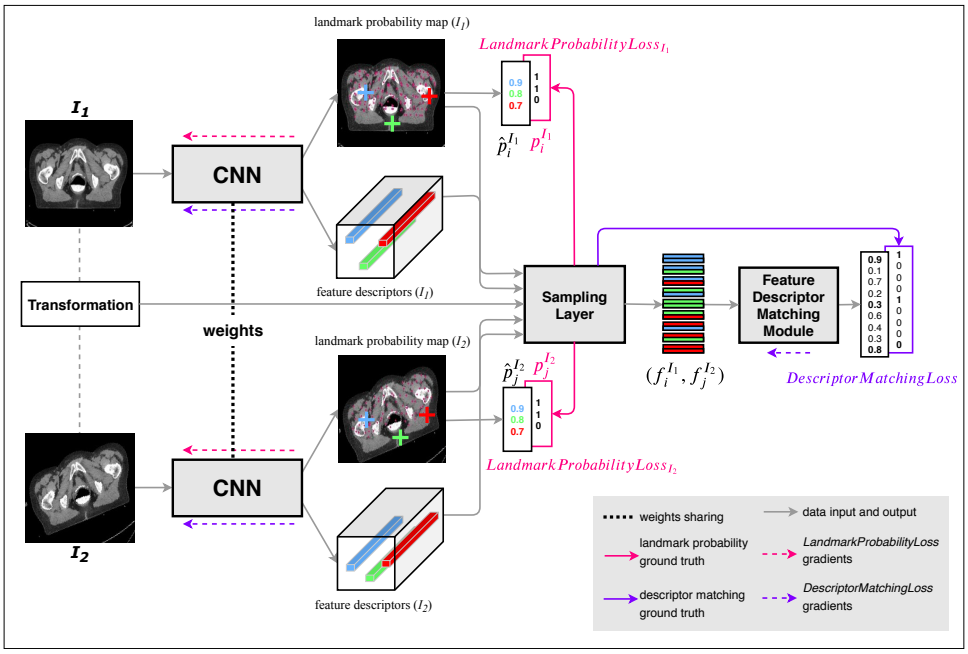


Figure 2.1: **Schematic representation of our approach.** The weights are shared between two branches of the Siamese neural network. The transformation is required only during training for calculating the ground truths. Abbreviations of the data input and output at various stages follow the description in the text.

CNN BRANCHES

The CNN branches of the Siamese neural network have shared weights and consist of an encoder-decoder type network similar to the U-Net[18] architecture. The only difference from the original implementation is that the number of convolutional filters in each layer is reduced by a factor of four to avoid overfitting. The implemented architecture contains 16, 32, 64, 128, and 256 convolutional filters in successive downsampling blocks respectively. The CNN branches give two outputs for each input image: a landmark probability map, and feature descriptors. The landmark probability map is computed at the end of the upsampling path after applying the sigmoid non-linearity and the feature descriptors are computed by concatenation of feature maps from the last two downsampling blocks. The feature maps from different downsampling blocks intrinsically allow for feature matching at multiple resolutions and abstraction levels.

SAMPLING LAYER

The sampling layer is a parameter-free module of the network. It performs the following tasks:

1. It samples K_1 and K_2 landmark locations in I_1 and I_2, respectively, which correspond to the highest probability score locations in the predicted landmark probability maps.

2. It extracts predicted landmark probabilities $\hat{p}_i^{I_1}$, and $\hat{p}_j^{I_2}$ corresponding to K_1 and K_2 locations in landmark probability maps of image I_1 and I_2.
3. It extracts feature descriptors $f_i^{I_1}$ and $f_j^{I_2}$ corresponding to the sampled landmark locations in I_1 and I_2, respectively, and creates feature descriptor pairs $(f_i^{I_1}, f_j^{I_2})$ for each $i \in \{1,2,...,K_1\}$ and $j \in \{1,2,...,K_2\}$.
4. During training, it generates the ground truths for landmark probabilities and feature descriptor matching probabilities on-the-fly as mentioned in Georgakis et al [7]. Briefly, the sampled landmark locations of I_2 are projected onto I_1 based on the known transformation between the images. A landmark location i in I_1 is decided to be matching to a landmark location j in I_2 if the Euclidean distance between i and the projection of j on image I_1 is less than a predefined pixel threshold ($thresh_{pixels}$).

FEATURE DESCRIPTOR MATCHING MODULE

All the feature descriptor pairs $(f_i^{I_1}, f_j^{I_2})$ are fed to the feature descriptor matching module. The feature descriptor matching module consists of a single fully connected layer that predicts the matching probability for each feature descriptor pair.

2.3.1. TRAINING

Training image pairs were generated on-the-fly by sampling a reference image randomly and generating the target image by transforming the reference image with a known transformation (randomly simulated brightness or contrast jitter, rotation, scaling, shearing, or elastic transformation). During training, the ground truths for landmark probabilities and feature descriptor matching probabilities are generated in the sampling layer as described above. We trained the network by minimizing a multi-task loss defined as follows:

$$Loss = LandmarkProbabilityLoss_{I_1} + LandmarkProbabilityLoss_{I_2} + DescriptorMatchingLoss \quad (2.1)$$

The $LandmarkProbabilityLoss_{I_n}$ for the probabilities of landmarks in image $I_n, n \in \{1,2\}$ is defined as:

$$LandmarkProbabilityLoss_{I_n} = \frac{1}{K_n} \sum_{i=1}^{K_n} \left((1 - \hat{p}_i^{I_n}) + CrossEntropy(\hat{p}_i^{I_n}, p_i^{I_n}) \right) \quad (2.2)$$

where $CrossEntropy$ is the cross entropy loss between predicted landmark probabilities $\hat{p}_i^{I_n}$ and ground truths $p_i^{I_n}$. The term $(1 - \hat{p}_i^{I_n})$ in Equation 2.2 encourages high probability scores at all the sampled landmark locations, whereas the cross entropy loss term forces low probability scores at the landmark locations that do not have a correspondence in the other image. As a consequence, the network is forced to predict high landmark probabilities only at the salient locations that have correspondence in the other image as well.

Hinge loss is widely used for learning discriminant landmark descriptors between matching and non-matching landmark pairs. We observed that a positive margin for the matching pairs in the hinge loss encourages the network to focus on hard positive examples (i.e., non-trivial landmark matches).

Therefore, we defined $DescriptorMatchingLoss$ (Equation 2.3) as a linear combination of hinge loss with a positive margin m_{pos} on the L2-norm of feature descriptor pairs and cross entropy loss on matching probabilities predicted by the feature descriptor matching module.

$$\begin{aligned} DescriptorMatchingLoss = \sum_{i=1,j=1}^{K_1,K_2} \Bigg(& \frac{c_{i,j}\, max(0, ||f_i^{I_1} - f_j^{I_2}||^2 - m_{pos})}{K_{pos}} \\ & + \frac{(1 - c_{i,j})\, max(0, m_{neg} - ||f_i^{I_1} - f_j^{I_2}||^2)}{K_{neg}} \\ & + \frac{WeightedCrossEntropy(\hat{c}_{i,j}, c_{i,j})}{(K_{pos} + K_{neg})} \Bigg) \end{aligned} \tag{2.3}$$

where $\hat{c}_{i,j}$, and $c_{i,j}$ are the predicted and the ground truth matching probabilities, respectively, for the feature descriptor pair $(f_i^{I_1}, f_j^{I_2})$; K_{pos} and K_{neg} are the number of matching (positive class) and non-matching (negative class) feature descriptor pairs; m_{pos} and m_{neg} are the margins for the L2-norm of matching and non-matching feature descriptor pairs. $WeightedCrossEntropy$ is the binary cross entropy loss where the loss corresponding to positive class is weighted by the frequency of negative examples and vice versa. The gradients are back-propagated from end-to-end as indicated by the dashed arrows in Figure 2.1.

2.3.2. CONSTRAINING LANDMARK LOCATIONS

A naive implementation of the approach may find all the landmarks clustered in a single anatomical region, which is not desirable. Therefore, to learn landmarks in all anatomical regions during training, we sample the landmarks on a coarse grid in the sampling layer, i.e., in each 8×8 pixel section of the grid, only one landmark location with the maximum landmark probability is sampled.

Another challenge in the CT scan imaging data comes from a large number of pixels belonging to the background. Traditionally, the image is cropped to the center to avoid prediction of landmarks in the background or on the patient table. However, this strategy requires an additional pre-processing step during inference. To avoid this, we computed a *valid mask* for each image, which contained the value 1 at the location of body pixels and 0 elsewhere. The valid mask was generated by image binarization using intensity thresholding and removing small connected components in the binarized image. The network is trained to predict high landmark probabilities as well as feature descriptor matching probabilities only in the matching locations that correspond to a value of 1 in the valid mask. This allows the network to learn a content-based prior on the landmark locations and avoids the need for image pre-processing during inference.

2.3.3. END-TO-END

The conventional approach to establish landmark correspondences between an image pair utilizes the following steps:

- Landmark detection, in which landmarks are detected in both the images independently.
- Feature description, wherein a vector (often called “descriptor") is calculated to describe the image properties surrounding the landmark location. An example of a feature descriptor is Scale Invariant Feature Transform (SIFT [14]), which calculates the histograms of orientations from the image patches of different scales around the landmark.
- Landmark matching, wherein landmark descriptors in both the images are matched using a matching algorithm. A straightforward matching algorithm is brute force matching, which aims at finding the best match among all the landmark locations in the source image for each landmark location in the target image.

Our approach replaces each of the abovementioned components with a neural network module, and connects the neural network modules such that the gradients flow from the end to the inputs. The modules of landmark detection and description are represented by the CNN branches of the Siamese network. The task of landmark matching is performed by the descriptor matching module. It is important to mention that the key feature of our approach lies in the assembling of different modules to provide a simple end-to-end deep learning solution for simultaneous landmark detection, description, and matching automatically. Therefore, the proposed approach can be easily modified, e.g., it may be improved by the use of a different neural network in any of the modules.

2.3.4. INFERENCE

During inference, only the locations in I_1 and I_2 with landmark probabilities above a threshold ($thresh_{landmark}$) are considered. Further, landmark pairs from different images are only matched if their matching is inverse consistent.

Suppose, locations $i \in \{1,..,K_1\}$ in I_1 and locations $j \in \{1,..,K_2\}$ in I_2 have landmark probabilities above $thresh_{landmark}$. A pair (i^*, j^*) is considered matching if there is no other pair (i^*, j') where $j' \in \{1,..,K_2\}$ or (i', j^*) where $i' \in \{1,..,K_1\}$ with higher descriptor matching probabilities or lower L2-norms for their feature descriptor pairs $(f_{i^*}^{I_1}, f_{j'}^{I_2})$ or $(f_{i'}^{I_1}, f_{j^*}^{I_2})$.

2.3.5. IMPLEMENTATION DETAILS

We implemented our approach using PyTorch[16]. We trained the network for 50 epochs using the Adam[13] optimizer with learning rate 10^{-3} and a weight decay of 10^{-4}. The training was done with a batchsize of 4 and took 28 GPU (NVIDIA GeForce RTX 2080 Ti) hours. To allow for batching, a constant K (set to 400) landmarks were sampled from all the images. The threshold for Euclidean distance while generating the ground truth

($thresh_{pixels}$) was 2 pixels. The margin for the L2-norm of matching feature descriptors (m_{pos}) was set to 0.1 and the margin for the L2-norm of non-matching pairs (m_{neg}) was set to 1. During inference, $thresh_{landmark}$ = 0.5 was used.

The empirical values for the hyperparameters were decided based on experience in the preliminary experiments. For example, the number for landmarks to be sampled during training (K) was decided such that the entire image was covered with sufficient landmark density, which was inspected visually. Similarly, the decision for $thresh_{pixels}$ was motivated by the fact that a threshold less than 2 pixels did not yield any matching landmarks in the first few iterations of the training and hence the network could not be trained. We initially trained the network with default values of m_{pos}, and m_{neg} (m_{pos} = 0, and m_{neg} = 1). However, we noticed on the validation set that all the predicted landmark pairs were clustered in regions of no deformation. To avoid this behaviour, we trained the network with m_{pos} = 0.1 and m_{pos} = 0.2 so that the gradients were not affected by the hinge loss corresponding to easy landmark matches. The final results are reported corresponding to the run with m_{pos} = 0.1 as it had a better trade off between number of landmarks per image pair and difficulty of landmark locations. The value of $thresh_{landmark}$ was chosen to give the best trade off between the number of landmarks per image pair and the spatial matching error on the validation set.

2.4. EXPERIMENTS

2.4.1. BASELINE

Scale Invariant Feature Transform (SIFT[14]) based keypoint detectors and feature descriptors are prevalent approaches used in both natural image analysis as well as in medical image analysis [8]. Therefore, we used the OpenCV[3] implementation of SIFT as the baseline approach for comparison. We used two matching strategies for SIFT: a) brute-force matching with inverse consistency (similar to our approach, we refer to this approach as SIFT-InverseConsistency), b) brute-force matching with ratio test (as described in the original paper[14], we refer to this approach as SIFT-RatioTest). Default values provided in the OpenCV implementation were used for all other hyperparameters.

2.4.2. DATASETS

The performance is evaluated on two test sets. First, for quantitative evaluation, we transformed all 7,402 testing images from 54 CT scans with three different types of transformations corresponding to intensity (jitter in pixel intensities = ±20% maximum intensity), affine (pixel displacement: median = 29 mm, Inter Quartile Range (IQR) = 14 mm - 51 mm), and elastic transformations (pixel displacement: median = 12 mm, IQR = 9 mm - 15 mm), respectively. Elastic transformations were generated by deforming the original image according to a deformation vector field representing randomly-generated 2D Gaussian deformations. The extent of transformations was decided such that the intensity variations and the displacement of pixels represented the typical variations in thoracic and abdominal CT scan images [23, 17]. This resulted in three sets of 7,402 2D image pairs (total 22,206 pairs).

Second, to test the generalizability of our approach in a clinical setting, image pairs were taken from two CT scans of the same patient but acquired on different days. The two scans were aligned with each other using affine registration in the SimpleITK [15] package. This process was repeated for three patients.

2.4.3. EVALUATION

For quantitative evaluation, we projected the predicted landmarks in the target images to the reference images and calculated the Euclidean distance to their corresponding matches in the reference images. We report the cumulative distribution of landmark pairs with respect to the Euclidean distance between them.

The performance of our approach on clinical data was assessed visually. We show the predicted results on four transverse slices belonging to different anatomical regions. To visually trace the predicted correspondences of landmarks, the colors of the landmarks in both the images vary according to their location in the original CT slice. Similarly colored dots between slices from original and follow-up image represent matched landmarks.

2.5. RESULTS

Table 2.1: **Description of predicted landmark matches.** Median number of landmark matches per image pair with Inter Quartile Range (IQR) in parentheses are provided together with the spatial matching error. SIFT-IC: SIFT-InverseConsistency, SIFT-RT: SIFT-RatioTest. The entries in bold represent the best value among all approaches.

Transformations		**Intensity**	**Affine**	**Elastic**
No. of landmarks	Proposed	639 (547 - 729)	466 (391 - 555)	370 (293 - 452)
	SIFT-IC	**711 (594 - 862)**	**610 (509 - 749)**	**542 (450 - 670)**
	SIFT-RT	698 (578 - 849)	520 (426 - 663)	418 (330 - 541)
Spatial matching error (mm)	Proposed	**0.0 (0.0 - 0.0)**	**1.0 (0.0 - 1.4)**	**1.0 (1.0 - 1.4)**
	SIFT-IC	1.0 (1.0 - 1.4)	1.0 (1.0 - 1.4)	1.0 (1.0 - 2.0)
	SIFT-RT	1.0 (1.0 - 1.4)	1.0 (1.0 - 1.4)	**1.0 (1.0 - 1.4)**

The inference time of our approach per 2D image pair is within 10 seconds on a modern CPU without any parallelization. On the GPU the inference time is ~20 milliseconds. The model predicted on average 639 (IQR = 547 - 729), 466 (IQR = 391 - 555), and 370 (IQR = 293 - 452) landmark matches per image pair for intensity, affine, and elastic transformations, respectively.

2.5.1. SIMULATED TRANSFORMATIONS

Table 2.1 describes the number of landmark matches per image pair and the spatial matching error for both our approach and the two variants of SIFT. Though our approach finds less landmarks per image as compared to the two variants of SIFT, the predicted landmarks have smaller spatial matching error than the SIFT variants. Further, Figure 2.2 shows the cumulative distribution of landmark pairs with respect to the Euclidean distance between them. All the approaches are able to find more than

90% of landmark matches within 2 mm error for intensity transformations. Predicting landmark correspondences under affine and elastic transformations is considerably more difficult; this can also be seen in the worse performance of all approaches. However, our approach is still able to find more than 99% of landmark matches within a spatial matching error of 4 mm and 8 mm, respectively for affine and elastic transformations. However, a noticeable percentage (about 2% for affine transformations and 3% for elastic transformations) of landmarks detected by SIFT-RatioTest are wrongly matched with landmarks from far apart regions (more than 64 mm). It should be noted that if landmark matches with such high inaccuracies are used for providing guidance to a registration method, it may have a deteriorating effect on the registration if the optimizer is not sufficiently regularized.

For visual comparison, the landmark correspondences in pairs of original and elastic transformed images are shown in Figure 2.3 (rows a-b) for our approach as well as for SIFT. As can be seen, the cases of mismatch in predictions from our approach (i.e., the number of landmarks in transformed slices not following the color gradient in the original slice) are rather scarce in comparison to the baseline approaches. Another interesting point to note is the difference in the landmark locations from our approach and the two baseline approaches. Since SIFT is designed to predict landmarks at locations of local extrema, the landmark matches are concentrated on the edges in the images. Our approach, however, predicts matches in soft tissue regions as well. Further inspection reveals that our approach predicts a considerable number of landmark matches even in the deformed regions in contrast to the baseline approaches. The capability to establish landmark correspondences in the soft tissues and deformed regions is important because DIR methods can especially benefit from guidance information in these regions.

2.5.2. CLINICAL TRANSFORMATIONS

Rows c-f in Figure 2.3 show landmark correspondences in pairs of transverse slices corresponding to the lower abdominal region in the original and follow-up CT for our approach as well as for SIFT. As can be seen, the original and follow-up slices have large differences in local appearance of structures owing to contrast agent, bladder filling, presence or absence of gas pockets, which was not part of the training procedure. It is notable that the model is able to find considerable landmark matches in image pairs despite these changes in local appearance. Moreover, the spatial matching error of landmarks seems similar to that of images with simulated transformations, in contrast to the baseline approach SIFT-InverseConsistency. Further, SIFT-RatioTest predicts fewer mismatched landmarks compared to SIFT-InverseConsistency, but this is achieved at the cost of a large decrease in the number of landmark matches per image pair.

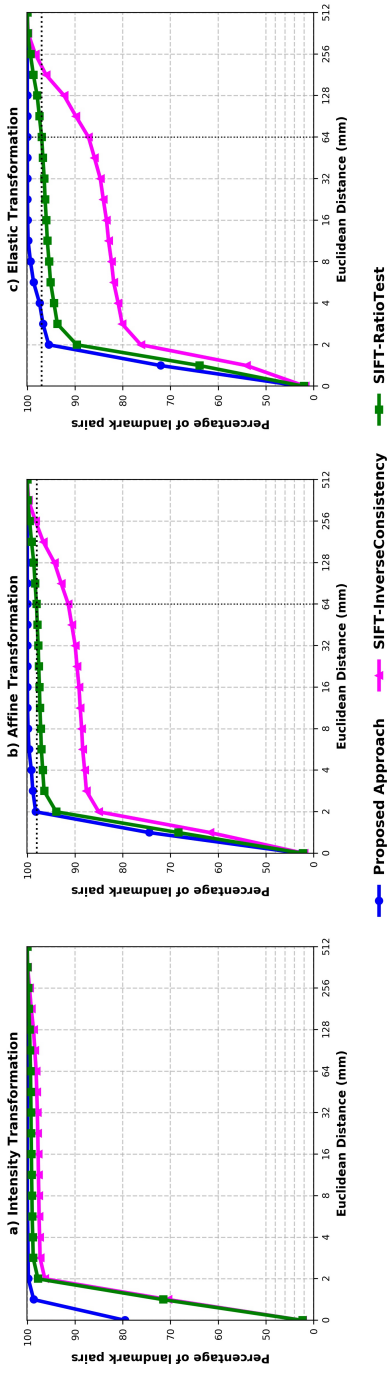


Figure 2.2: **Cumulative distribution of landmarks.** The cross-hairs in (b) and (c) correspond to the percentile of landmarks in SIFT-RatioTest at 64 mm.

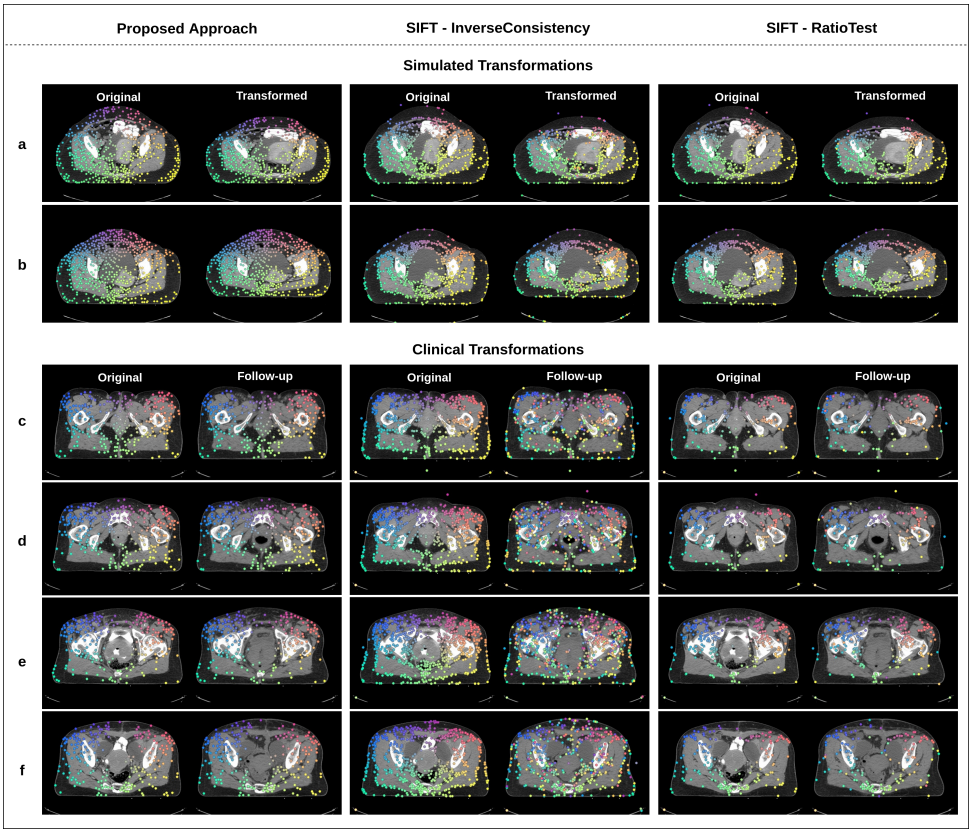


Figure 2.3: **Landmark correspondences for pairs of different transverse slices in abdominal CT scans.** The landmark correspondences predicted by our approach are shown in comparison with two variants of SIFT. **Rows (a-b)** show predictions on pairs of original (left) and elastic transformed (right) slices. **Rows (c-f)** show transverse slices taken from different anatomical regions. The slices in the original CT (left) are matched with a similar slice from a follow-up CT scan (right) by affine registration.

2.6. DISCUSSION

With a motivation to provide additional guidance information for DIR methods of medical images, we developed an end-to-end deep learning approach for the detection and matching of landmarks in an image pair. To the best of our knowledge, this is the first approach that simultaneously learns landmark locations as well as the feature descriptors for establishing landmarks correspondences in medical imaging. While the final version of this manuscript was being prepared, we came across one research on retinal images [21], whose approach for landmark detection using UNet architecture in a semi-supervised manner is partly similar to ours. However, our approach not only learns the landmark locations, but also the feature descriptors and the feature matching such that the entire pipeline for finding landmark correspondences can be replaced by a neural network. Therefore, our approach can be seen as an essential extension to the mentioned approach.

Our proposed approach does not require any expert annotation or prior knowledge regarding the appearance of landmarks in the learning process. Instead, it learns landmarks based on their distinctiveness in feature space despite local transformations. Such a definition of landmarks is generic so as to be applicable in any type of image and sufficient for the underlying application of establishing correspondences between image pairs. Further, in contrast to the traditional unsupervised approaches for landmark detection in medical imaging, the proposed approach does not require any pre- or post-processing steps, and has fewer hyperparameters.

The main challenge for intensity based DIR methods is to overcome local optima caused by multiple low contrast regions in the image, which result in image folding and unrealistic transformations in the registered image. It can be speculated that the availability of landmark correspondences in the low contrast image regions may prove to be beneficial for DIR methods. Moreover, a uniform coverage of entire image is desirable for improved performance. Upon visual inspection of the landmarks predicted by our approach, we observed that our approach not only finds landmark correspondences in bony anatomical regions but also in soft tissue regions lacking intensity gradients. Moreover, a considerable density of landmarks (approximately 400 landmarks per image pair) was observed despite the presence of intensity, affine, or elastic transformations. Based on these observations, we are optimistic about the potential added value of our approach to the DIR methods.

We validated our approach on images with simulated intensity, affine, and elastic transformations. The quantitative results show low spatial matching error of the landmarks predicted by our approach. Additionally, the results on clinical data demonstrate the generalization capability of our approach. We compared the performance of our approach with the two variants of widely used SIFT keypoint detection approach. Our approach not only outperforms the SIFT based approach in terms of matching error under simulated transformations, but also finds more accurate matches in the clinical data. As such the results look quite promising. However, the current approach is developed for 2D images i.e., it overlooks the possibility of the out-of-plane correspondences in two CT scans, which is quite likely especially in lower abdominal regions. The extension of the approach to 3D is, therefore, imperative so as to speculate into its benefits in providing additional guidance information to the DIR methods.

BIBLIOGRAPHY

[1] Tanja Alderliesten, Peter A. N. Bosman, and Arjan Bel. "Getting the most out of additional guidance information in deformable image registration by leveraging multi-objective optimization". In: *Medical Imaging 2015: Image Processing*. Vol. 9413. Proc. SPIE. International Society for Optics and Photonics. 2015, 94131R.

[2] Bastian Bier et al. "X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 55–63.

[3] G. Bradski. "The OpenCV library". In: *Dr. Dobb's Journal* 25 (2000), pp. 120–125. URL: https://ci.nii.ac.jp/naid/10028167478/en/.

[4] J. Chen et al. "A Partial Intensity Invariant Feature Descriptor for Multimodal Retinal Image Registration". In: *IEEE Transactions on Biomedical Engineering* 57.7 (2010), pp. 1707–1718. DOI: 10.1109/TBME.2010.2042169.

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. "Superpoint: Self-supervised interest point detection and description". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 224–236.

[6] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), p. 115.

[7] Georgios Georgakis et al. "End-to-end learning of keypoint detector and descriptor for pose invariant 3D matching". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1965–1973.

[8] Zeinab Ghassabi et al. "An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors". In: *EURASIP Journal on Image and Video Processing* 2013.1 (2013), p. 25. ISSN: 1687-5281. DOI: 10.1186/1687-5281-2013-25. URL: https://doi.org/10.1186/1687-5281-2013-25.

[9] Varun Gulshan et al. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". In: *JAMA Network Open* 316.22 (2016), pp. 2402–2410. ISSN: 0098-7484. DOI: 10.1001/jama.2016.17216. eprint: https://jamanetwork.com/journals/jama/articlepdf/2588763/joi160132.pdf. URL: https://doi.org/10.1001/jama.2016.17216.

[10] Yulan Guo et al. "A Comprehensive Performance Evaluation of 3D Local Feature Descriptors". In: *International Journal of Computer Vision* 116.1 (2016), pp. 66–89. ISSN: 1573-1405. DOI: 10.1007/s11263-015-0824-y. URL: https://doi.org/10.1007/s11263-015-0824-y.

[11] Dong Han et al. "Robust anatomical landmark detection with application to MR brain image registration". In: *Computerized Medical Imaging and Graphics* 46 (2015), pp. 277–290. ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.

compmedimag.2015.09.002. URL: http://www.sciencedirect.com/science/article/pii/S089561111500124X.

[12] Álvaro S. Hervella et al. "Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement". In: *Procedia Computer Science* 126 (2018). Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia, pp. 97–104. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.07.213. URL: https://www.sciencedirect.com/science/article/pii/S1877050918311876.

[13] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations*. 2015. URL: http://arxiv.org/abs/1412.6980.

[14] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110.

[15] Bradley C Lowekamp et al. "The design of SimpleITK". In: *Frontiers in Neuroinformatics* 7 (2013), p. 45.

[16] Adam Paszke et al. "Automatic differentiation in PyTorch". In: *Advances in Neural Information Processing Systems-W*. 2017. URL: https://github.com/pytorch/pytorch.

[17] Thomas Polzin et al. "Combining automatic landmark detection and variational methods for lung CT registration". In: *Fifth International Workshop on Pulmonary Image Analysis*. 2013, pp. 85–96.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[19] J. Rühaak et al. "Estimation of Large Motion in Lung CT by Integrating Regularized Keypoint Correspondences into Dense Deformable Registration". In: *IEEE Transactions of Medical Imaging* 36.8 (2017), pp. 1746–1757. ISSN: 0278-0062. DOI: 10.1109/TMI.2017.2691259.

[20] James Thewlis, Hakan Bilen, and Andrea Vedaldi. "Unsupervised Learning of Object Landmarks by Factorized Spatial Embeddings". In: *The IEEE International Conference on Computer Vision*. 2017, pp. 5916–5925.

[21] Prune Truong et al. "GLAMpoints: Greedily Learned Accurate Match points". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 10732–10741.

[22] Ahmet Tuysuzoglu et al. "Deep Adversarial Context-Aware Landmark Detection for Ultrasound Imaging". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 151–158.

[23] Eliana M. Vásquez Osorio et al. "Accurate CT/MR vessel-guided nonrigid registration of largely deformed livers". In: *Medical Physics* 39.5 (2012), pp. 2463–2477. DOI: 10.1118/1.3701779. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.3701779. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3701779.

[24] René Werner et al. "Assessing accuracy of non-linear registration in 4D image data using automatically detected landmark correspondences". In: *Medical Imaging*

2013: Image Processing. Vol. 8669. Proc. SPIE. International Society for Optics and Photonics. 2013, 86690Z.

[25] Deshan Yang et al. “A method to detect landmark pairs accurately between intra-patient volumetric medical images”. In: *Medical Physics* 44.11 (2017), pp. 5859–5872.

[26] Yuting Zhang et al. “Unsupervised discovery of object landmarks as structural representations”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2694–2703.

3

Automatic Landmarks Correspondence Detection in 3D and Application to Deformable Image Registration

Purpose: *Deformable Image Registration (DIR) can benefit from additional guidance using corresponding landmarks in the images. However, the benefits thereof are largely understudied, especially due to the lack of automatic landmark detection methods for three-dimensional (3D) medical images.*

Approach: *We present a Deep Convolutional Neural Network (DCNN), called DCNN-Match, that learns to predict landmark correspondences in 3D images in a self-supervised manner. We trained DCNN-Match on pairs of Computed Tomography (CT) scans containing simulated deformations. We explored five variants of DCNN-Match that use different loss functions and assessed their effect on the spatial density of predicted landmarks and the associated matching errors. We also tested DCNN-Match variants in combination with the open-source registration software Elastix to assess the impact of predicted landmarks in providing additional guidance to DIR.*

Results: *We tested our approach on lower-abdominal CT scans from cervical cancer patients: 121 pairs containing simulated deformations and 11 pairs demonstrating clinical deformations. The results showed significant improvement in DIR performance when landmark correspondences predicted by DCNN-Match were used in the case of simulated ($p = 0e^0$) as well as clinical deformations ($p = 0.030$). We also observed that the spatial density of the automatic landmarks with respect to the underlying deformation*

affect the extent of improvement in DIR. Finally, DCNN-Match was found to generalize to Magnetic Resonance Imaging (MRI) scans without requiring retraining, indicating easy applicability to other datasets.
***Conclusions**: DCNN-Match learns to predict landmark correspondences in 3D medical images in a self-supervised manner, which can improve DIR performance.*

The content of this chapter is based on the following publication: **Grewal, M.**, Wiersma, J., Westerveld, H., Bosman, P. A. N., & Alderliesten, T. (2023). *Automatic Landmark Correspondence Detection in Medical Images with an Application to Deformable Image Registration.* Journal of Medical Imaging, 10(1), 014007-014007.

3.1. INTRODUCTION

Deformable Image Registration (DIR) is a task of aligning a source (or moving) image to a target (or fixed) image by optimizing a Deformation Vector Field (DVF). The aligned source image can then be computed by resampling the source image at the spatial locations specified by the mapping. DIR has tremendous application possibilities in the radiation treatment workflow required for cancer treatment e.g., automatic contour propagation [5, 13], dose accumulation [33, 27, 6]. However, DIR in regions such as the pelvis is challenging due to large local deformations and appearance differences caused by physical processes such as bladder filling, and the presence of gas pockets and contrast agents [13]. In such DIR scenarios, the existing non-linear intensity-based registration approaches [22, 35, 36] often get stuck in a local minimum[27]. Many previous studies [1, 37, 26, 29, 18, 16] have shown that landmark correspondences between the images to be registered can provide additional guidance to the intensity-based DIR methods and help overcome local minima. However, to the best of our knowledge, such an approach has not been tested on pelvic scans.

Manual annotation of landmarks for DIR in the clinic is not practically tractable due to two main reasons. First, a high number of landmarks is desired, and it is difficult to unambiguously define such a high number of landmarks manually. Second, manual annotations require lots of time from clinicians, which is hardly available. Therefore, an automatic method for finding landmark correspondences is required. Although many endeavours have been made in the direction of automatic landmarks correspondence detection in medical images [39, 16, 3], there remain significant gaps to fill. The existing methods usually employ large pipelines consisting of multiple components, each component using multiple hyperparameters derived from image features specific to the underlying dataset. Consequently, the entire pipeline is sensitive to small variations in local image intensities and choices of hyperparameters, making application to a new dataset difficult. Moreover, in datasets such as pelvic scans with ill-defined boundaries between soft tissues, intensity gradient based landmark detection may not work at all.

Convolutional Neural Networks (CNNs) are known to learn deep features from images, which are robust to small variations in local image intensities. In recent years, deep CNNs have not only shown remarkable performance in difficult computer vision tasks in medical imaging [14, 10], but also good generalization to unseen data. Moreover, with the advances in the available computational resources, CNN-based solutions turn out to be faster than their traditional counterparts. Therefore, there is a strong motivation to replace the entire pipeline for automatically finding landmark correspondences by a deep CNN. Recently some deep CNN methods have been developed for automatic landmark detection in medical images [34, 12], but these are limited to either 2D datasets or supervised learning of a few manually annotated landmarks.

Other relevant works include methods for landmark propagation from a template image by learning pixel-wise anatomical embeddings [38] or through deformable image registration [7]. While such methods allow for single shot landmark detection in a new image, the requirement of manual annotation of landmarks on the template image still

exists. Another study uses unsupervised image registration as a proxy task to discover landmarks shape descriptors [2], but this method is limited to discovering a small number of landmarks (~ 100 landmarks per image pair).

In this study, we present a deep CNN (referred to as “DCNN-Match") for automatic landmarks correspondence detection (i.e., simultaneous landmark detection as well as matching) in 3D images. The presented method is an extension of our method for 2D images (described in Chapter 2). Briefly, the neural network is trained on pairs of 3D lower abdominal Computed Tomography (CT) scans such that the network learns to predict landmarks at salient locations in both the images along with the correspondence score of each landmark pair. One key feature of the presented method is that unlike supervised methods, the neural network in the presented method is trained in a self-supervised manner without using any manual annotations. This is important because manual annotations on medical images are not always readily available, mainly because it is time-consuming to create them.

It is essential to investigate the added value of automatic landmarks correspondence detection towards the improvement of the DIR solutions to estimate the potential deployability of landmarks-guided DIR approaches in the clinic. Existing studies have investigated the added value of automatic landmark correspondences towards DIR independently of the underlying automatic landmark detection method [37, 26, 16]. Since change in the automatic landmarks correspondence detection method changes the aspects of the automatic landmarks e.g., spatial distribution and matching accuracy, the effect of the automatic landmarks on the DIR performance is likely to be affected as well. Therefore, we believe that developing a method for automatic landmarks correspondence detection and at the same time integrating it with a DIR pipeline can provide numerous insights. To this end, we have integrated our method for automatic landmark detection and matching with an existing DIR software so that the added value of using landmark correspondences in solving DIR problems can be assessed. Further, we investigate five different variants of the developed method by use of different loss functions during training that each predict landmark correspondences with different spatial distributions and matching errors, to assess the effect of different types of automatic landmark correspondences towards the improvement of DIR. The present work has the following contributions:

- We extended our previously published end-to-end self-supervised deep learning method for automatically finding landmark correspondences in medical images from 2D to 3D. The key highlights of the method are:
  - the method does not set any prior on the definition of landmarks
  - the method does not require manual annotations for training
- We integrated our automatic landmark correspondence detection method in 3D (DCNN-Match) with an open-source registration software Elastix [31, 22] to develop a DIR pipeline that utilizes additional guidance information from automatic landmark correspondences. We used this DIR pipeline to investigate the added value of automatic landmark correspondences in providing additional guidance to the DIR method and finding better DIR solutions.

- We varied the landmarks correspondence detection method and investigated how it affected the added value to the DIR method. We explored five different variants of the proposed automatic landmarks correspondence method.
- We experimentally investigated the generalization capability of our proposed automatic landmarks correspondence detection method to a Magnetic Resonance Imaging (MRI) dataset.

3.2. MATERIALS AND METHODS

In the following sections, we describe DCNN-Match (Section 3.2.1), and the DIR pipeline which uses the information from automatic landmark correspondences predicted by DCNN-Match to guide the registration (Section 3.2.2). Sections 3.2.3 and 3.2.4 provide details of implementation and hyperparameters for reproducibility. Sections 3.2.5, 3.2.6, 3.2.7, and 3.2.8 describe the datasets, experiments, evaluation metrics, and statistical testing used in the experiments, respectively.

3.2.1. DCNN-MATCH

We extended our approach in Chapter 2 for finding landmark correspondences in 2D CT scan slices to work on 3D CT scans. The different components of the 3D approach are illustrated in Figure 3.1. Briefly, the approach proposed in Chapter 2 consists of a Siamese network with three modules: a) two CNN branches with shared weights, b) a sampling layer, c) a descriptor matching module. The CNN branches comprise an image-to-image translation network that maps an input image to a feature map. The architecture of the network is derived from the famous UNet architecture [28] proposed for image segmentation. For a given pair of target image (I_{target}) and source image (I_{source}), the CNN branches predict a landmark probability map describing the probability $\hat{p}_i^{I_x}$ ($x \in \{target, source\}$) of each spatial location i being a landmark. The sampling layer is a parameter-free module that samples K (hyperparameter) landmark locations with top landmark probabilities during training. During inference, the sampling layer samples all landmark locations with landmark probabilities above a threshold. We used the value 0.5, same as in Chapter 2.

Additionally, the sampling layer samples a feature vector from the feature maps of the last two downsampling levels in the CNN branch at the coordinates of each i^{th} landmark location and constructs the feature descriptor $f_i^{I_x}$ by concatenating the sampled feature vectors. This allows for efficient use of the network weights by simultaneous learning the landmark detection as well as feature description of each landmark without unnecessarily increasing the network size. Moreover, the concatenation of features from different downsampling levels emulates the behavior of multi-scale feature description, which otherwise, is achieved by calculating features from a Gaussian pyramid representation of the image. Following the calculation of feature descriptors for each landmark location, the sampling layer creates feature descriptor pairs $(f_i^{I_{target}}, f_j^{I_{source}}) \forall\, i = 1,2,...,K$ in I_{target} and $\forall\, j = 1,2,...,K$ in I_{source} to feed to the descriptor matching module. The descriptor matching module predicts the landmark matching probabilities corresponding to each feature descriptor pair.

SELF-SUPERVISED TRAINING

The network is trained in a self-supervised manner on pairs of target (I_{target}) and source (I_{source}) lower abdominal CT scans containing simulated deformations. The details on the generation of target and source image pairs are provided in Section 3.2.3.

Following the sampling of landmark locations $i = 1,2,...,K$ in I_{target} and $j = 1,2,...,K$ in I_{source} along with their corresponding feature descriptors $f_i^{I_{target}}$ and $f_j^{I_{source}}$, feature descriptor pairs $(f_i^{I_{target}}, f_j^{I_{source}})$ are constructed in the sampling layer. The feature descriptor pairs are considered corresponding to all i and j, allowing for feature descriptor matching between far-away locations in the images without requiring encoding of the underlying deformation field explicitly. Since the simulated deformations used to create source and target image pairs during training can not represent the complex large deformations in a clinical setup exactly, learning the feature descriptor matching not explicitly dependent on the underlying deformation field is likely to help the neural network generalize better to clinical scenario.

The ground truth $c_{i,j}$ of the correspondence of each feature descriptor pair is calculated on-the-fly based on the known simulated deformation. Each sampled landmark location in the target image is projected onto the source image based on the known simulated deformation and the nearest predicted landmark (within a distance of 2 voxels = 4 mm) in the source image is considered its match. We used a threshold of 4 mm (instead of image resolution = 2 mm) in order to find a reasonable number of landmark matches from random predictions in the beginning of the training to ensure sufficient supervision. The value of $c_{i,j}$ is set to 1 for matching and 0 for non-matching feature descriptor pairs. Subsequently, the ground truth $p_i^{I_x}$ for the landmark probability of landmark location i in image $I_x, x \in \{target, source\}$ is determined as follows:

$$p_i^{I_x} = \begin{cases} 1 & \text{if } \exists!\, j \in \{0,1,2,...,K\} \text{ in image } I_y, y \in \{target, source\},\ y! = x \wedge c_{i,j} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The ground truths $c_{i,j}$ are used directly as ground truths for the matching probability of the feature descriptor pairs $(f_i^{I_{target}}, f_j^{I_{source}})$. In other words, the ground truth is generated such that the landmark probability as well as the descriptor matching probability is high for the matching locations between the two images and low otherwise. The network is trained by minimizing a multi-task loss defined as follows:

$$Loss = LandmarkProbabilityLoss_{I_{target}} + LandmarkProbabilityLoss_{I_{source}} + DescriptorMatchingLoss \tag{3.2}$$

The $LandmarkProbabilityLoss_{I_x}$ for the probabilities of landmarks in image $I_x, x \in \{target, source\}$ is defined as:

$$LandmarkProbabilityLoss_{I_x} = \frac{1}{K}\sum_{i=1}^{K}\left((1-\hat{p}_i^{I_x}) + CrossEntropyLoss(\hat{p}_i^{I_x}, p_i^{I_x})\right) \tag{3.3}$$

where $CrossEntropyLoss$ is the cross entropy loss between predicted landmark probability $\hat{p}_i^{I_x}$ and ground truth $p_i^{I_x}$ of the i^{th} sampled location. K is the total number of sampled landmark locations in image I_x. Further details of the $LandmarkProbabilityLoss$ are omitted for brevity and can be found in Chapter 2.
The $DescriptorMatchingLoss$ allows the network to learn feature descriptor matching automatically and is defined as follows:

$$DescriptorMatchingLoss = DescriptorHingeLoss + DescriptorCELoss \tag{3.4}$$

$DescriptorHingeLoss$ is defined as follows:

$$DescriptorHingeLoss = \sum_{i=1,j=1}^{K,K}\left(\frac{c_{i,j}\,max(0, ||f_i^{I_{target}} - f_j^{I_{source}}||^2 - m_{pos})}{K_{pos}} + \frac{(1-c_{i,j})\,max(0, m_{neg} - ||f_i^{I_{target}} - f_j^{I_{source}}||^2)}{K_{neg}}\right) \tag{3.5}$$

where, $f_i^{I_{target}}$ and $f_j^{I_{source}}$ are the feature descriptors corresponding to the i^{th} and j^{th} landmark locations in the input images I_{target} and I_{source}, respectively; $c_{i,j}$ is the ground truth matching probability for the feature descriptor pair $(f_i^{I_{target}}, f_j^{I_{source}})$; m_{pos} and m_{neg} are the margins for the L2-norm of matching (positive class) and non-matching (negative class) feature descriptor pairs. The Hinge losses corresponding to positive and negative classes are normalized by K_{pos} (number of positive feature descriptor pairs) and K_{neg} (number of negative feature descriptor pairs), respectively to account for the class imbalance between positive and negative feature descriptor matches. $DescriptorCELoss$ is defined as follows:

$$DescriptorCELoss = \sum_{i=1,j=1}^{K,K}\left(\frac{WeightedCrossEntropy(\hat{c}_{i,j}, c_{i,j})}{(K_{pos}+K_{neg})}\right) \tag{3.6}$$

where $\hat{c}_{i,j}$ is the predicted matching probability; $WeightedCrossEntropy$ represents the binary cross entropy loss where the loss corresponding to the positive class is weighted by the frequency of negative examples and vice versa.

In the beginning of the training, the predicted landmark probability maps by the CNN branches are random and by chance only a few landmark locations have correct correspondence (i.e., $c_{i,j} = 1$) between images. The loss defined in (3.2) encourages high landmark probability at these locations as well as high feature descriptor matching

probability for the feature descriptor pairs of these locations and low landmark probability and feature descriptor matching probability otherwise. Additionally, the term $(1 - \hat{p}_i^{I_x})$ in (3.3) encourages high landmark probability at all locations i.e., encourages more landmark locations to have correct correspondence in the other image. Consequently as the training progresses, the network learns to identify salient locations in the images that have correct correspondence in the other image as well and predicts high landmark probabilities at the these locations.

EXTENSION TO 3D IMAGES

We have extended our original approach proposed in the previous chapter (Chapter 2) to work on 3D images by performing three modifications. The first obvious modification was to use 3D convolutional kernels (kernel size = $3 \times 3 \times 3$) instead of 2D convolutional kernels in the CNN branches. The sampling layer and the feature descriptor matching module were also adapted for 5D tensors arising from training on 3D images. The generation of a *valid mask* during training as described in Chapter 2 section 2.4 was also adapted for 3D images. The *valid mask* makes the network learn a content-based prior to predict landmarks only in the regions that include patient anatomy and not in the background or the CT couch.

Second, since we had a considerably large training dataset (details in Section 3.2.5) as opposed to Chapter 2, we kept the same number of kernels in each layer as the original UNet architecture [28]. Third, we trained the network on 3D patches of the entire CT due to GPU memory constraints. During inference, we evaluated the network on the patches belonging to the same spatial locations in the target and source images. The patches were cut with 50% overlap and the final output combined the predicted landmark pairs in all patches. All the corresponding landmarks predicted in all the overlapping patches were considered landmarks. Using a small patch size restricts the network from learning landmark matches in locations that are far apart in the two images. Therefore, the patch size has to be decided while keeping in mind the spatial extent of deformations we want the network to learn. This is further described in the hyperparameters section (Section 3.2.4).

3.2.2. DIR WITH ADDITIONAL GUIDANCE FROM AUTOMATIC LANDMARK CORRESPONDENCES

We integrated DCNN-Match with the open-source registration software Elastix [22, 31, 23] to create a pipeline for DIR that utilizes the additional guidance information from automatic landmark correspondences. A schematic of the DIR pipeline is provided in Figure 3.2.

DIR requires calculation of a DVF that maps each spatial location in the target image to a spatial location in the source image. In Elastix, the DVF is parameterized by B-splines and the coefficients of B-splines are optimized by non-linear optimization. We align the source CT scans with the target CT scans using affine registration before performing DIR. The parameters of the 3D affine transformation matrix (i.e., translation, rotation, scale, and shear) are optimized by maximizing the normalized mutual information between the target and source scans. The target and the affine

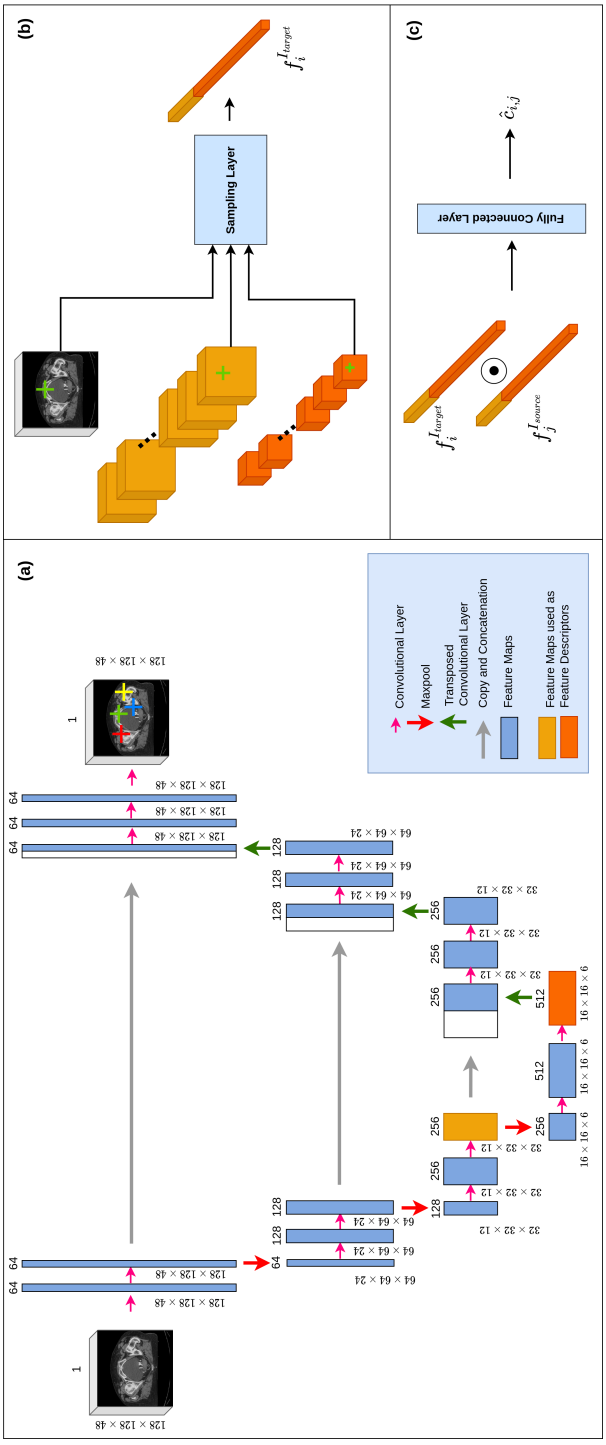


Figure 3.1: **Illustration of the components of DCNN-Match.** (a) Illustration of different layers in the shared **CNN branch** used for landmark detection and feature description. (b) The **sampling layer** samples the feature maps of the last two downsampling levels in the CNN branch at the locations described by the landmark probability map. (c) The **descriptor matching module** realized by a fully connected layer predicts the matching probability of a feature descriptor pair.

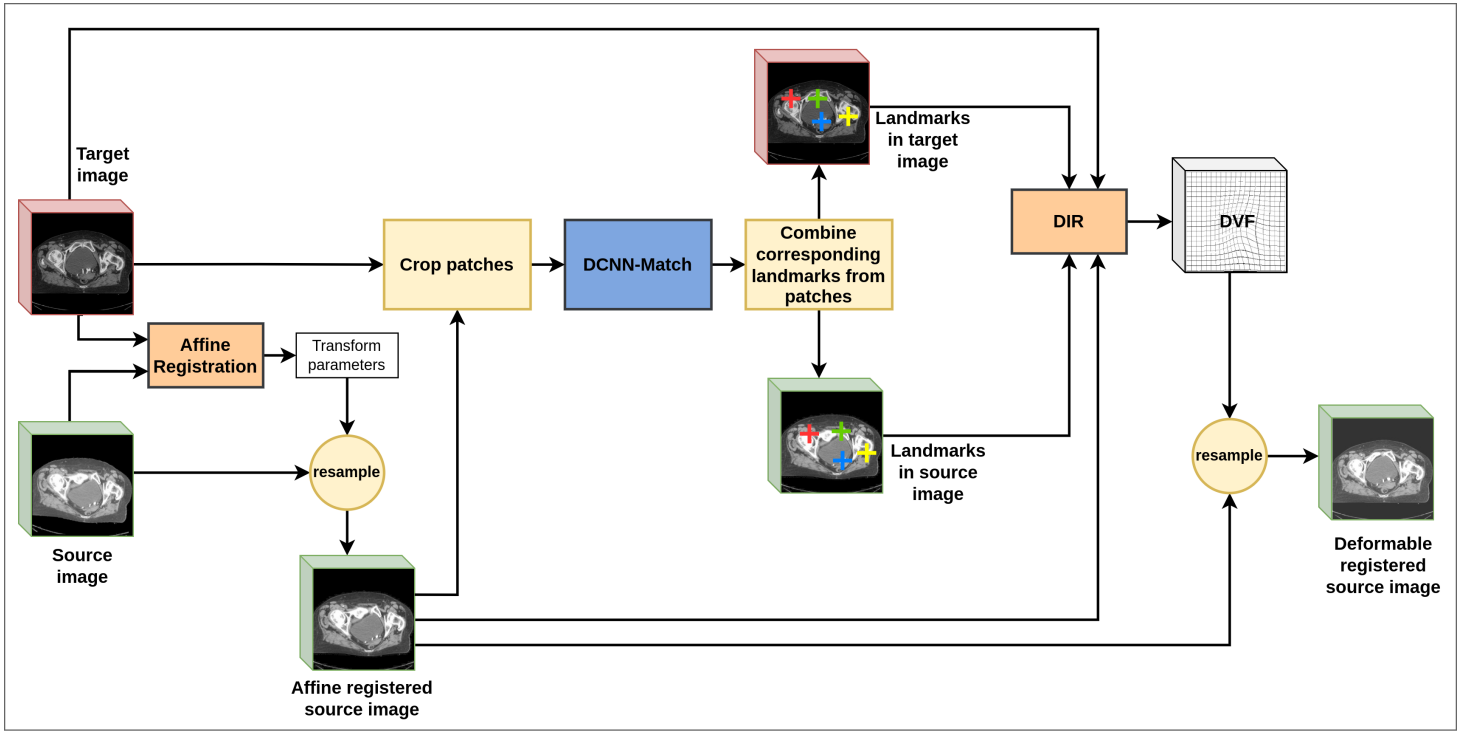


Figure 3.2: **DIR pipeline with automatic landmarks correspondence detection using DCNN-Match.** The source image is affine registered with the target image followed by automatic landmarks correspondence detection using DCNN-Match. DCNN-Match provides the locations of corresponding landmarks (shown with similar colored cross-hairs) in both the target and affine registered source image. The DIR module finds a DVF by utilizing the additional guidance information from automatic landmark correspondences. The final transformed (deformable registered) source image is obtained by resampling the affine registered source image according to the obtained DVF.

registered source CT scan are input to the DCNN-Match, which provides the locations of corresponding landmarks in both the scans. The DIR module in Elastix takes the target image, affine registered source image, and the pairs of corresponding landmarks in both the images as input. The DIR is performed by optimizing the following objective function:

$$\begin{aligned} f_{Guidance} = & \, weight_0 \, AdvancedMattesMutualInformation \\ & + weight_1 \, TransformBendingEnergyPenalty \\ & + weight_2 \, CorrespondingPointsEuclideanDistanceMetric \end{aligned} \tag{3.7}$$

where $AdvancedMattesMutualInformation$ represents the maximization of mutual information between two scans (for details refer to [32]).
$TransformBendingEnergyPenalty$ is a regularization term that penalizes large transformations, and $CorrespondingPointsEuclideanDistanceMetric$ is used for minimizing the Euclidean distance between the landmarks in the target CT and the landmarks in the source CT. $weight_0$, $weight_1$, and $weight_2$ control the relative contribution of each term towards the objective function.

3.2.3. IMPLEMENTATION

The DIR pipeline was developed in Python. We used the PyTorch framework [24] for developing DCNN-Match. The training was done on an RTX 2080 Ti GPU and took approximately 21 hours. The weights of DCNN-Match were initialized using the He norm method [17]. The training was done using the Adam optimizer [21] with a learning rate of $1e^{-4}$. The neural network weights were regularized by using a weight decay of $1e^{-4}$.

We randomly cropped 3D patches of dimension 128 × 128 × 48 from the entire CT scan volume and used them as target images. The source images were generated on-the-fly by applying one of the following random transformations on the target images: translation, rotation, scale, or elastic transformations. The magnitudes of the affine transformations along all axes were sampled from the following uniform distributions: $U(-12mm, 12mm)$, $U(-20°, 20°)$, and $U(0.9, 1.1)$ for translation, rotation, and scale respectively. The elastic transformations were applied so as to simulate the two types of soft tissue deformations present in the lower abdominal scans: a) large local deformations e.g., bladder filling, b) small tissue deformations everywhere in the image. The large local deformations were simulated by a 3D Gaussian DVF (DVF_{large}) of magnitude at center = $U(2mm, 24mm)$ and $\sigma = U(64mm, 128mm)$ at a random location in the image. The small deformations everywhere in the image were simulated by Gaussian smoothing of a random DVF ($DVF_{small} = U(1mm, 12mm)$) at each location. DVF_{large} and DVF_{small} were additively applied to the target image to generate the source image with elastic transformation.

3.2.4. HYPERPARAMETERS

Apart from the conventional hyperparameters involved in designing and training a DCNN e.g., network depth and width, optimizer, and learning rate, there are two hyperparameters specific to DCNN-Match: patch dimensions and the number of sampling points during training (K). As indicated in the previous section, we used a patch size of 128 × 128 × 48 (256 mm × 256 mm × 96 mm). This way the neural network's Field-Of-View (FOV) was maximum given the network depth and GPU memory constraints, which ensured that the landmark correspondences could be learned for deformations as large as 128 mm in-plane and 48 mm along the transverse axis. Similar to Chapter 2, $K = 512$ was used based on the visual inspection that the predicted landmarks in the validation set (details in Section 3.2.5) covered the image sufficiently.

In Elastix, we used the advanced mattes mutual information as a similarity metric because it has been found successful in earlier studies on DIR [13]. For deciding other hyperparameters such as the number of iterations, step size, step decay, $weight_0$, $weight_1$, and $weight_2$, we used the development set (details in Section 3.2.5). For this purpose, the pairs of target and source images were generated in a manner similar to the training set. 100 locations were sampled randomly on the target image and their corresponding location in the source image was established by transforming the coordinates with the inverse DVF used for generating the source image. The hyperparameters were tuned based on the following observations on the development set: the transformed source image after registration was not distorted and showed no visible folding, the image alignment at the 100 randomly sampled locations improved after registration. The exact configuration of Elastix used for affine registration and DIR is provided in the Appendix 3.6.1.

3.2.5. DATA

An overview of the data is provided in Figure 3.3. We retrospectively included the CT and MRI scans from female patients (age range 22 - 95 years), who received radiation treatment in the lower abdominal region between the year 2009 and 2019 at the Amsterdam University Medical Centers, location AMC, the Netherlands. The data was transferred in anonymized form through a data transfer agreement. A subset of these scans was the same as used in Chapter 2.

TRAINING AND VALIDATION SET

A total of 1671 CT scans of 831 patients were used for developing the approach: 1335 CT scans for training and 336 CT scans for validation. A subset containing 10 CT scans from the validation set (referred to as the development set) was used to tune the hyperparameters of the DIR pipeline. All the CT scans were resampled to have 2 mm × 2 mm × 2 mm voxel spacing and the image intensities were converted from the Hounsfield units to a range of 0 to 1 after windowing.

SIMULATED DEFORMATIONS TEST SET - CT

We tested the performance of DCNN-Match and the DIR pipeline on a curated dataset of 121 CT scans belonging to 121 patients, who received radiation treatment for cervical

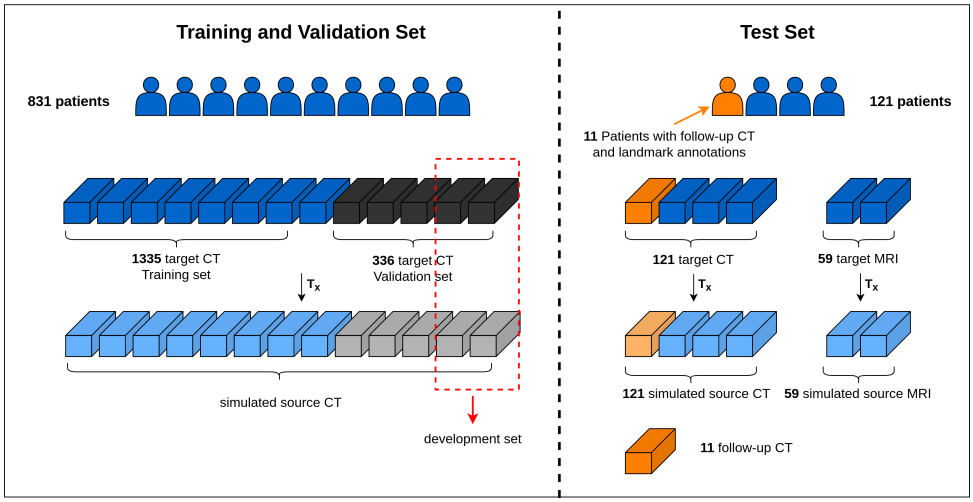


Figure 3.3: **Data Overview.** The vertical dashed gray line depicts the patient-level split between the training and validation set, and test set.

cancer. The mean FOV of acquisition of the CT scans was 546 mm × 546 mm × 368 mm and the scans were resampled to 2 mm × 2 mm × 2 mm voxel spacing. The available CT scans were used as target images and corresponding source images were simulated by applying random elastic transformations to the target CT scans according to the method described in the Section 3.2.3 above. Further, an example of the simulated deformation and the obtained source CT is shown in Figure 3.4 (a).

In each pair of target and source image, 100 corresponding locations were sampled with uniform random distribution. These sampled locations were used as validation landmarks for assessing the performance of DCNN-Match and the DIR pipeline.

CLINICAL DEFORMATIONS TEST SET - CT

The CT scans in a clinical setup exhibit complex bio-mechanical deformations including discontinuities in the deformation field around sliding tissues and large deformations that may not be Gaussian. The random Gaussian DVF used for deforming the images to obtain a simulated test set is an oversimplification of the underlying situation. Therefore, it is essential to investigate if the observations on the simulated deformations test set hold in the clinical setting as well. To this end, additional CT scans (referred to as follow-up scans) were searched in the clinical database for a subset of patients in the test set (11 patients). The first CT scans from these patients were used as target images and the corresponding follow-up CT scans were used as source images.

Corresponding landmarks at 29 locations were manually identified in each target and source CT scan by a clinical expert. These landmarks included six fiducial markers in the vaginal wall, and anatomical landmarks e.g., aortic bifurcation, cervical os, and os coccygis. Since clinically available scans were used, the number of fiducial markers were different in each patient in accordance with the treatments given to the patients.

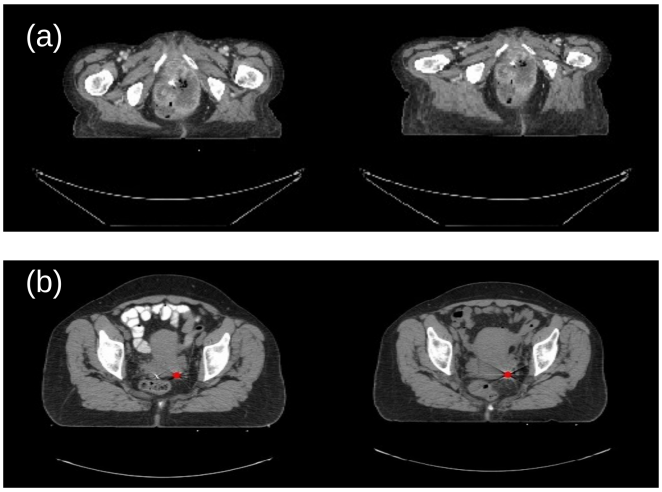


Figure 3.4: **Transverse slices from representative examples.** (a) simulated deformations test set: the source CT (right) is obtained by applying an elastic transformation to the target CT (left). (b) clinical deformations test set: the landmark at the location of a fiducial marker (shown with red dot) in the target (left) and source (right) CT is shown. Note the appearance difference in the bowel due to contrast.

The majority of the patients' scans had three fiducial markers, while some had less or more. If a patient's scan had less than three fiducial markers, calcification (if present) in corresponding anatomical locations were used as landmarks. If a patient's scan had more than three fiducial markers, only three of them were used. An example landmark location is shown in Figure 3.4 (b) and the complete list of landmark locations is provided in Appendix 3.6.2.

SIMULATED DEFORMATIONS TEST SET - MRI

MRI scans of 59 cervical cancer patients (subset of the 121 cervical cancer patients mentioned in Section 3.2.5, who received brachytherapy treatment) acquired during brachytherapy treatment delivery were used to investigate the generalization capability of DCNN-Match. The mean FOV of acquisition of the MRI scans was 199 mm × 199 mm × 152 mm and the scans were resampled to 2 mm × 2 mm × 2 mm voxel spacing. The pairs of source and target scans were generated in a similar way to the CT scans (Section 3.2.5).

3.2.6. EXPERIMENTS

We conducted three types of experiments. The first type of experiments were aimed to gain insights in the working of DCNN-Match by changing the $DescriptorMatchingLoss$ (sections 3.2.6 and 3.2.6). The second type of experiments were done to investigate the effect of automatically predicted landmark correspondences on the performance of DIR (Section 3.2.6). We also investigated how the changes in $DescriptorMatchingLoss$ affected the added value of the automatic landmark correspondences toward the performance of DIR. Third, we investigated the generalization capability of DCNN-Match on a different modality (Section 3.2.6).

DESCRIPTOR LOSS

We trained three versions of DCNN-Match, each with a different descriptor loss. The first version was trained with only the $DescriptorHingeLoss$ defined in Equation (3.5). This version is referred to as DCNN-Match Hinge. DCNN-Match Hinge was trained with $m_{pos} = 0$ and $m_{neg} = 1$. In the second version, only $DescriptorCELoss$ Equation (3.6) was employed. We refer to this version as DCNN-Match CE. Next, we trained the network with a linear combination of $DescriptorHingeLoss$ and $DescriptorCELoss$ Equation (3.4), which is referred to as DCNN-Match Hinge+CE.

POSITIVE MARGIN IN THE HINGE LOSS

We considered that the L2-norm of the descriptor pairs of highly deformed regions would be high and these pairs would be difficult to match. Further, it is intuitive to think that the landmark matches in regions of high deformation would provide more added value to the DIR approach. To allow the network to focus more on matching these pairs, we trained DCNN-Match Hinge+CE with two values for m_{pos}: 0.1 and 0.2. These versions are referred to as DCNN-Match Hinge0.1+CE and DCNN-Match Hinge0.2+CE, respectively. The value of $m_{pos} > 0$ in the $DescriptorHingeLoss$ makes the loss term 0 for descriptor pairs whose L2-norm is less than m_{pos} i.e., the network already identifies the descriptor pairs as matching. Thus, the gradients are influenced only by the descriptor pairs which are difficult to match. Consequently, the network should be able to predict difficult landmark correspondences in the highly deformed regions accurately.

EFFECT OF ADDITIONAL GUIDANCE FROM AUTOMATIC LANDMARK CORRESPONDENCES

To assess the effect of additional guidance from automatic landmark correspondences on the DIR, we compared the results from the DIR pipeline with ($weight_2 = 0.01$ in Equation (3.7) as obtained from hyperparameter tuning on the development set) and without ($weight_2 = 0$ in Equation (3.7)) automatic landmarks correspondence detection.

GENERALIZATION TO MRI DATASET

Given the capability of deep neural networks to learn robust features, and the self supervised nature of our training approach, optimistically one would expect that the developed approach would generalize to different datasets. To this end, we tested DCNN-Match on pairs of MRI scans containing simulated deformations (described in Section 3.2.5) without retraining. Compared to the training set, the MRI scans were not only different in imaging modality, but also in the FOV of acquisition.

3.2.7. EVALUATION

SPATIAL MATCHING ERRORS OF LANDMARK CORRESPONDENCES

In the simulated deformations test set, the landmarks on the source CT scans were projected on the target CT scans using the known transformation between them. The Euclidean distances between the landmarks on the target CT scans and the projection of their corresponding landmarks predicted by the network were calculated. The Euclidean distance gives a measure of the spatial matching error of the predicted

landmark correspondences. The spatial matching errors were compared between all versions of DCNN-Match.

Quantitative analysis of the spatial matching errors of the predicted landmark correspondences is not feasible in the clinical deformations test set due to the absence of the ground truth DVF. To provide some insights into the performance on the clinical deformations test set, we conducted a validation study on a subset of the data. For this purpose, we randomly sampled 75 predicted landmarks from DCNN-Match CE in two patients (total 150 landmark correspondences). A radiation oncologist (henceforth, referred to as clinician) ranked these landmark correspondences on a 3 point Likert scale: 1 = good match (roughly within 5 mm distance), 2 = moderate match (roughly within 10 mm distance), 3 = poor or wrong match (roughly more than 15 mm distance) in a 3D (axial, sagittal, and coronal) image viewer. The (approximate) spatial matching errors were calculated based on the ranking provided by the clinician. The clinician also labelled the anatomical location of the landmarks in target CT scans according to the following categories: a) bony anatomy, b) soft tissue (i.e., muscles, fatty tissue, and fascia), c) bowel i.e., rectum, large and small bowel including gas pockets, d) other (including organs and blood vessels i.e., veins and arteries). We also analyzed the spatial matching errors of the predicted landmark correspondences separately for each anatomical category.

TARGET REGISTRATION ERROR

In the clinical deformations test set, we transformed the manually annotated landmarks in the target images according to the estimated DVF after DIR using the transformix module in SimpleElastix [23] (documentation on using transformix in SimpleElastix can be found at SimpleElastix documentation and Elastix manual). We calculated their Euclidean distance with the corresponding landmarks in the source image. This measure is often referred to as “Target Registration Error" or TRE. We calculated the TRE values after initial affine registration and before the DIR (TRE_{before}) and after DIR (TRE_{after}) for all experiments. In the simulated deformations test set, TRE calculations were done using the randomly sampled validation landmarks described in Section 3.2.5.

It should be noted that the TRE calculations were done in image space i.e., the landmarks were represented by the center of a voxel. We chose this setup because the automatic landmarks are predicted in image space. However, this setup may give rise to discretized TRE values.

LANDMARK CORRESPONDENCES VS. UNDERLYING DEFORMATION

It is intuitive to think that the DIR performance in a specific region is dependent on the underlying deformation in that region. Concordantly, the distribution of landmarks with respect to the underlying deformation would impact the additional guidance provided by the landmarks overall. Therefore, it is important to investigate the choice of landmark locations by the network with respect to the extent of deformation at those locations. To this end, we partitioned the voxels in the source images in the simulated deformations test set into bins of different deformations. For each DCNN-Match variant, the spatial density of predicted landmark correspondences was calculated in

each bin of the underlying deformation by dividing the number of landmarks with the number of voxels in each bin.

Similarly, we calculated the percentage of automatic landmarks below 4 mm spatial matching errors (as a surrogate for landmarks correspondence accuracy) and TRE values of validation landmarks (as a measure of DIR performance) in each deformation region. The threshold of 4 mm was chosen because the same threshold was used during training. In each deformation region, we analyzed the TRE values in light of the spatial density and landmarks correspondence accuracy to gain insights about what aspects of automatic landmarks affect the DIR performance.

DETERMINANT OF SPATIAL JACOBIAN

Evaluating the performance of DIR is a difficult task and TRE can only give an estimate of performance on sparse image locations. Moreover, TRE can give a biased perspective of the DIR performance because of the observer subjectivity in the manual annotation of landmark locations. In order to assess whether the obtained DVF is anatomically plausible or not, the determinant of the spatial Jacobians of the DVF is a good measure. The negative values in the determinant of the spatial Jacobian represent singularities in the DVF and indicate image folding in those regions. Therefore, we also investigated the determinant of the spatial Jacobians of the obtained DVFs after DIR.

3.2.8. STATISTICAL TESTING

The statistical testing was done using IBM SPSS Statistics for Ubuntu (Version 27.0, IBM Corp. Released 2020. Armonk, NY: IBM Corp)[19]. We tested the null hypothesis that the TRE_{after} values in the test sets were the same in the following experimental scenarios: DIR without additional guidance from corresponding landmarks, and DIR with additional guidance from five different variants of DCNN-Match.

Kolmogorov-Smirnov tests for normality revealed that the TRE_{after} values were not normally distributed in any of the experimental scenarios. Therefore, we used the related samples Friedman's two way Analysis of Variance by Ranks test followed by post-hoc pairwise comparisons using Dunn-Bonferroni test [8]. An alpha of 0.05 with Bonferroni correction for multiple comparisons was considered significant.

3.3. RESULTS

The average inference time of DCNN-Match variants for predicting landmark correspondences in one CT scan pair was ~20s. A representative example of predicted landmark correspondences is shown in Figure 3.5. The images in the figure are shown with the couch table cropped for better visualization, but the automatic landmark correspondence detection as well as DIR were performed on full CT scans without any cropping.

3.3.1. NUMBER OF LANDMARK CORRESPONDENCES

The number of landmark correspondences predicted per image on the simulated test set and clinical test set is described in Table 3.1 and Figure 3.6. As can be seen in Table 3.1 and Figure 3.6, DCNN-Match Hinge and DCNN-Match CE approaches

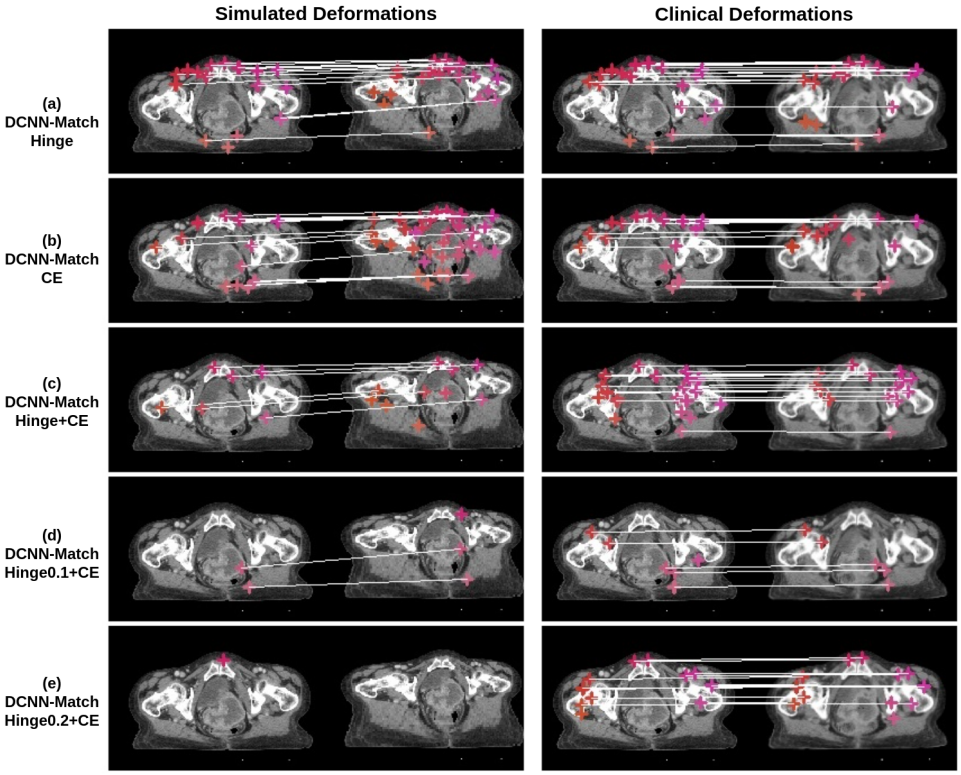


Figure 3.5: **Visualization of predicted landmark correspondences** by (a) DCNN-Match Hinge, (b) DCNN-Match CE, (c) DCNN-Match Hinge+CE, (d) DCNN-Match Hinge0.1+CE, and (e) DCNN-Match Hinge0.2+CE. A transverse slice from target and source CTs in the simulated deformations test set (left) and the clinical deformations test set (right) is shown. The corresponding landmarks are shown with the same colored cross-hairs in target and source image and a white line is drawn for in-slice corresponding landmarks. Please note that some corresponding landmarks may lie on a different slice and are therefore not visible in the figure.

Table 3.1: Number of predicted landmark correspondences per CT scan pair for each variant of DCNN-Match. Mean (M) ± standard deviation (SD), and range (5^{th} percentile – 95^{th} percentile) are provided.

	Hinge	CE	Hinge+CE	Hinge0.1+CE	Hinge0.2+CE
Simulated Deformations					
M ± SD	5488 ± 2258	7761 ± 2540	1698 ± 888	1735 ± 959	1220 ± 871
Range	2160 – 9580	2999 – 11400	595 – 3462	563 – 3563	244 – 3028
Clinical Deformations					
M ± SD	3708 ± 1052	7427 ± 1682	946 ± 391	1062 ± 479	511 ± 307
Range	2563 – 5344	5394 – 10340	491 – 1569	455 – 1819	193 – 1000

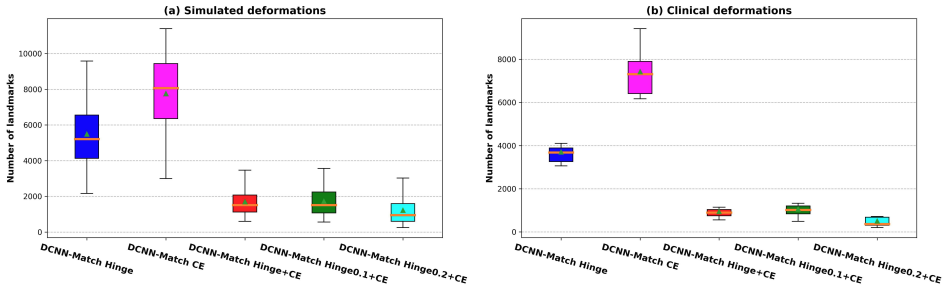


Figure 3.6: **Distribution of predicted automatic landmark correspondences across patients.** The boxes extend from the lower to upper quartile values of the data, with a line at the median. Mean is shown with a triangular marker and whiskers represent the range from 5^{th} to 95^{th} percentile.

predicted a large number of landmarks per CT scan pair. In DCNN-Match Hinge+CE, the use of an auxiliary loss allows for applying an additional constraint on the landmark correspondences. Consequently, the number of predicted landmark correspondences per image was fewer than with using either of the loss separately. Further, the DCNN-Match Hinge0.1+CE and DCNN-Match Hinge0.2+CE predicted even fewer landmarks per CT scan pair, possibly due to the additional constraint posed by the positive margin m_{pos} used in the Hinge loss. It should be noted that irrespective of the differences within different DCNN-Match variants, a considerable number of landmark correspondences were predicted by all of them in both the simulated as well as the clinical deformations test set.

The cumulative distribution of the predicted landmark correspondences in the simulated test set is plotted against their spatial matching errors in Figure 3.7. Both DCNN-Match Hinge and DCNN-Match CE predicted more than 70% landmarks with less than 2 voxels (equivalent to 4 mm) spatial matching error. But, DCNN-Match CE predicted a higher percentage of landmarks within a specific spatial matching error as compared to DCNN-Match Hinge. The decrease in spatial matching errors could be attributed to the added parameters used in the dedicated descriptor matching module in DCNN-Match CE as opposed to the parameter-free module in DCNN-Match Hinge. Further, DCNN-Match Hinge+CE takes advantage of the auxiliary loss and therefore, the landmark correspondences are predicted with lower spatial matching errors. About 90% of the predicted landmarks had a spatial matching error of less than 4 mm.

As expected, training with $m_{pos} > 0$ yielded landmarks with lower spatial matching errors as compared to DCNN-Match Hinge+CE (Figure 3.7). Specifically for DCNN-Match Hinge0.2+CE, more than 90% of the predicted landmark correspondences had spatial matching errors of less than 1 voxel, which is equivalent to 2 mm (image resolution). This finding indicates the potential of the automatic landmark correspondences predicted by the DCNN-Match variant for use in clinical applications.

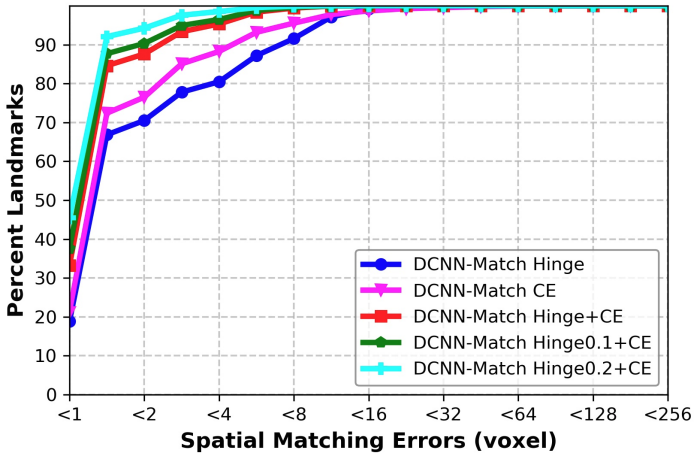


Figure 3.7: Cumulative distribution of the landmarks with respect to the spatial matching errors for different versions of DCNN-Match on the simulated deformations test set - CT.

3.3.2. SPATIAL MATCHING ERRORS IN CLINICAL DATA

In Figure 3.5, the predicted landmark correspondences from DCNN-Match variants on a representative transverse slice from the clinical deformations test set are shown for the reader's perusal. More examples are shown in Figure 3.8 (a)-(d). The border colors indicate the ranking given by the clinician: green = good, yellow = moderate, red = wrong match. Figure 3.8 (a) demonstrates a good match in the small bowel despite the difference of the underlying contrast and Figure 3.8 (b) demonstrates a good match in the muscle despite a change in the muscle deformation. In Figure 3.8 (c), both the landmarks are present in the rectum, but in different locations, although it was challenging to review because of the presence of the gas pocket and change in the muscle deformation. Figure 3.8 (d) shows an example of a wrong match in the bowel. It is important to note the underlying challenges visible between the two scans in Figure 3.8 (d) e.g., difference in contrast, and content mismatch due to presence of gas pockets.

3.3.3. SPATIAL MATCHING ERRORS OF LANDMARK CORRESPONDENCES

Figure 3.8 (e) shows that more than 72% landmark correspondences were ranked as good match i.e., approximately within 5 mm distance and about 90% landmark correspondences were ranked to be within 10 mm distance. These results indicate only a small performance difference in comparison to the simulated deformations test set (magenta curve in Figure 3.7 and Figure 3.8 (e)), which is expected due to the presence of additional challenges in the clinical data.

Further, in Figure 3.8 (f), the percentage of landmarks in bony anatomy, soft tissue, bowel, and other regions is plotted. The bars in the plot are shaded in green, yellow,

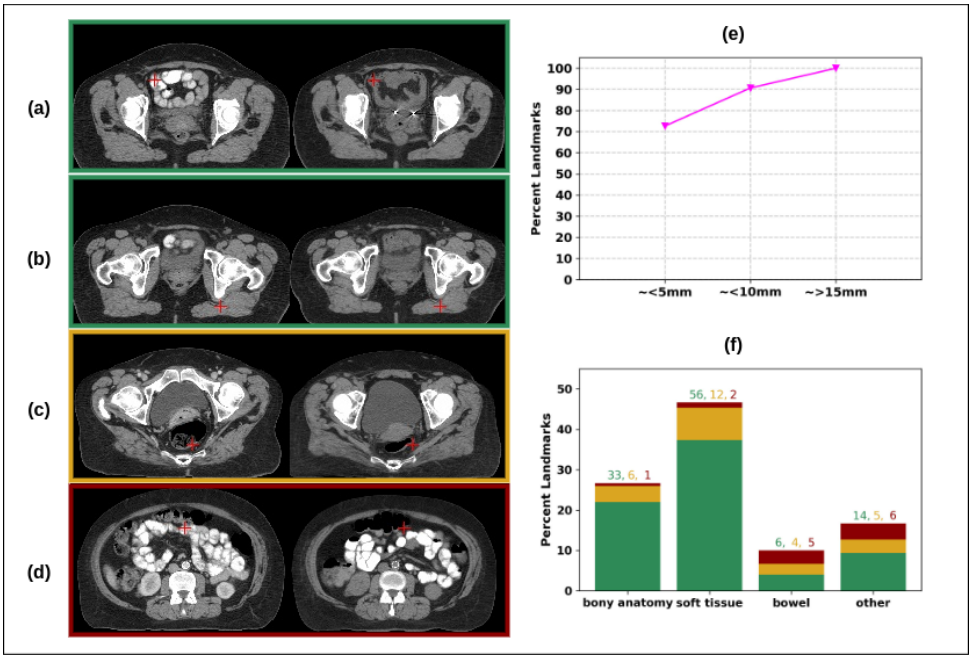


Figure 3.8: **Validation of landmark correspondences in clinical deformations test set.** Representative examples of (a)-(b): good match despite contrast variation and difference in muscle deformation, (c): moderate match, and (d): wrong match. (e): Cumulative distribution of landmarks with respect to (approximate) spatial matching errors. (f): Distribution of landmarks in different anatomical categories. The bars are shaded in proportion to the number of landmarks corresponding to a rank: green = good, yellow = moderate, red = wrong. In each anatomical category, the total number of landmarks representing a rank is provided in the text above bars in the corresponding color.

and red colors in proportion to the ranking of the landmarks (green = good, yellow = moderate, red = wrong) in that anatomical category. It is worth noting that the wrong matches are mainly in the bowel region, where content mismatch may happen along with large deformations and intensity variations.

3.3.4. EFFECT OF LANDMARK CORRESPONDENCES ON DIR

In Table 3.2, the TRE values in the simulated and clinical deformations test sets are provided. In Figure 3.9, boxplots of TRE values are provided. In both test sets, there was a significant main effect of the experimental scenario (i.e., DIR without landmarks and with landmarks predicted by either one of the DCNN-Match variants) on the observed TRE_{after} values, ($\chi(5) = 6620.117$, $p = 0e^{0}$) in the simulated test set and ($\chi(5) = 34.051$, $p = 0.000002$) in the clinical test set. Note that in the simulated test set, the sample size was quite large (100 landmarks per scan × 121 scans = 12100) giving rise to near zero p values in the statistical testing.

In the simulated deformations test set, the post-hoc comparisons revealed that TRE_{after} values from registration using additional guidance from landmark

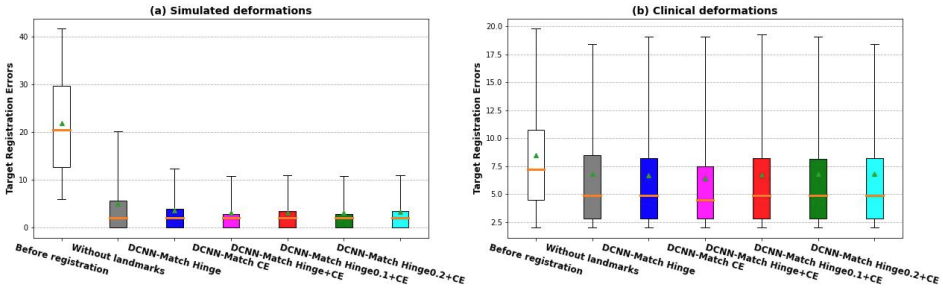


Figure 3.9: **Distribution of Target Registration Errors (TRE).** The boxes extend from the lower to upper quartile values of the data, with a line at the median. Mean is shown with a triangular marker and whiskers represent the range from 5^{th} to 95^{th} percentile.

Table 3.2: Target Registration Errors (TREs) in mm of pre-specified landmarks (for details refer to 3.2.7) before DIR but after affine registration (TRE_{before}) and after DIR with different approaches (TRE_{after}). Mean (M) ± standard deviation (SD), and range (5^{th} percentile – 95^{th} percentile) are provided. Best TRE values are highlighted in bold. * represents significance in post-hoc comparison against TRE_{after} without landmarks.

		Simulated Deformations		**Clinical Deformations**	
		M ± SD	Range	M ± SD	Range
TRE_{before}		21.99 ± 12.67	6.00 – 41.76	8.50 ± 5.81	2.00 – 19.96
TRE_{after}	Without landmarks	5.07 ± 9.98	0.00 – 20.20	6.85 ± 5.79	2.00 – 19.12
	Hinge	3.58 ± 8.80 *	0.00 – 12.33	6.69 ± 5.84	2.00 – 19.53
	CE	**3.14 ± 8.61***	0.00 – 10.77	**6.42 ± 5.79***	2.00 – 19.94
	Hinge+CE	3.21 ± 8.63*	0.00 – 10.95	6.74 ± 5.77	2.00 – 19.47
	Hinge0.1+CE	3.18 ± 8.62*	0.00 – 10.77	6.79 ± 5.83	2.00 – 19.31
	Hinge0.2+CE	3.27 ± 8.65*	0.00 – 10.95	6.82 ± 5.86	2.00 – 19.53

correspondences predicted by any of the DCNN-Match variants were significantly lower than TRE_{after} values from registration without using additional guidance from landmark correspondences. However, the strongest effect was observed with landmark correspondences from DCNN-Match CE ($p = 0e^0$).

On the clinical deformations test set, although the TRE_{after} values from registration with the use of additional guidance by automatic landmark correspondences were smaller than the TRE_{after} values from registration without using additional guidance from landmark correspondences, the differences were small. Only the post-hoc comparison between TRE_{after} values from registration by using landmark correspondences predicted by DCNN-Match CE and TRE_{after} values from registration without using landmark correspondences yielded statistical significance after correction for multiple comparisons ($p = 0.030$).

3.3.5. DIFFERENTIAL EFFECT OF DCNN-MATCH VARIANTS ON DIR

The post-hoc analysis indicated that the landmarks predicted by DCNN-Match CE had significantly more added value (as reflected by the TRE_{after} values) as compared to the landmarks predicted by DCNN-Match Hinge on the simulated test set (p = $0e^0$). However, a similar finding could not be corroborated on the clinical deformations test set – pairwise comparison of TRE_{after} values obtained by DCNN-Match CE and DCNN-Match Hinge did not yield significance after correcting for multiple comparisons (p = 0.406).

Based on the observed spatial matching errors, it is intuitive to expect that DCNN-Match Hinge+CE would yield lower TRE values after registration as compared to DCNN-Match CE. However, surprisingly this is not the case (Table 3.2). TRE_{after} values using DCNN-Match CE were significantly lower than TRE_{after} values using DCNN-Match Hinge+CE in the simulated deformations test set (p = 0.013). In the clinical deformations test set also, the TRE_{after} values using DCNN-Match CE were significantly lower than TRE_{after} values using DCNN-Match Hinge+CE (p = 0.046).

Furthermore, the TRE values after registration were not affected by increasing m_{pos} in the simulated test set. The post-hoc pairwise comparisons of TRE_{after} values by using DCNN-Match Hinge+CE vs DCNN-Match Hinge0.1+CE were not significant (p = 0.783) on the simulated deformations test set. In fact, the TRE_{after} values by using DCNN-Match Hinge0.2+CE values were significantly higher than TRE_{after} values by using DCNN-Match Hinge0.1+CE (p = 0.000244). This indicates that even though an increase in m_{pos} predicts landmark correspondences with lower spatial matching errors, there is no additional benefit toward DIR performance. The observations on clinical deformations also corroborated the findings on simulated deformations. None of the post-hoc comparisons between experimental scenarios with different m_{pos} values were significantly different in the clinical deformations test set.

Overall, the results from pairwise comparisons between the TRE_{after} indicate that the added value of the automatic landmark correspondences towards the improvement of DIR performance is dependent on the underlying approach for identifying automatic landmark correspondences.

3.3.6. RELATION BETWEEN ASPECTS OF AUTOMATIC LANDMARKS AND DIR PERFORMANCE

In Figure 3.10 (a), the landmarks correspondence accuracy (averaged over 121 patients) as described in Section 3.2.7 in the regions of different underlying deformation is plotted for each DCNN-Match variant. As can be seen, the correspondence accuracy of the automatic landmarks predicted by DCNN-Match Hinge deteriorated as the underlying deformation increased. A similar trend was observed for DCNN-Match CE, but to a lesser extent. As expected, the correspondence accuracy of the landmarks predicted by DCNN-Match Hinge+CE was higher than both DCNN-Match Hinge as well as DCNN-Match CE in all regions of the underlying deformation. Further, the purpose of experimenting with m_{pos} = 0.1 and m_{pos} = 0.2 to encourage high landmarks correspondence accuracy in the regions of high deformation seems to be

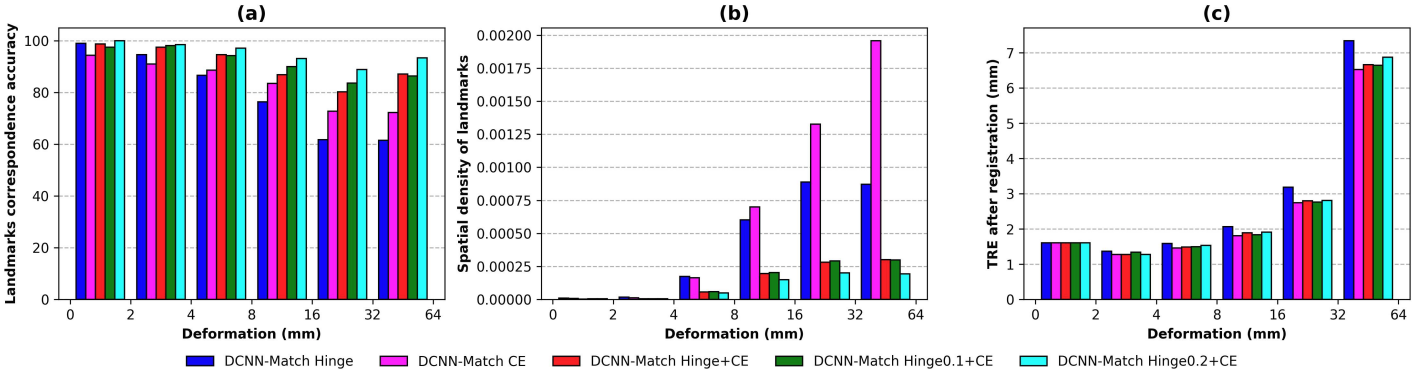


Figure 3.10: **Analysis of relation between aspects of landmark correspondences and DIR performance.** (a) Landmarks correspondence accuracy in different regions of underlying deformation represented by the percentage of landmarks predicted within 4 mm spatial matching errors. (b) Spatial density of landmarks (number of landmarks per voxel) predicted in different regions of underlying deformation. (c) TRE of validation landmarks after DIR by using automatic landmarks.

fulfilled. The landmarks correspondence accuracy was high irrespective of the extent of the underlying deformation for DCNN-Match Hinge0.1+CE and even higher for DCNN-Match Hinge0.2+CE.

In Figure 3.10 (b), the spatial density of predicted landmark correspondences (averaged over 121 patients) in different regions of the underlying deformation is plotted for each DCNN-Match variant. The plot shows that DCNN-Match CE predicted more landmarks in regions with high deformations as compared to other DCNN-Match variants, which is purely empirical.

In Figure 3.10 (c), the TRE_{after} values of the validation landmarks (averaged over 121 patients) in different region of the underlying deformation are plotted for each DCNN-Match variant. As is apparent from the figure, the TRE_{after} values were lowest in all deformation regions when the automatic landmarks predicted by DCNN-Match CE were used as compared to the other DCNN-Match variants.

If we analyze the plots in the Figure 3.10 collectively, we observe that in high deformation regions, DCNN-Match CE predicted landmarks with lower landmarks correspondence accuracy but higher spatial density as compared to DCNN-Match Hinge+CE, DCNN-Match Hinge0.1+CE, and DCNN-Match Hinge0.2+CE. Further, the DIR performance in the highly deformed regions was higher (reflected by lower TRE_after values) with the use of the automatic landmarks predicted by DCNN-Match CE as compared to DCNN-Match Hinge+CE, DCNN-Match Hinge0.1+CE, and DCNN-Match Hinge0.2+CE. This implies that a larger number of slightly less accurate landmarks in highly deformed regions may be more favorable for guiding the DIR approach as compared to a smaller number of highly accurate landmarks.

3.3.7. DETERMINANT OF SPATIAL JACOBIAN AND QUALITATIVE EVALUATION

The determinant of the spatial Jacobian of the obtained DVFs was observed to be non-negative in all the registrations obtained in all the experimental scenarios. This indicates that all the obtained registrations were anatomically plausible.

Figure 3.11 shows a representative example of registration without using landmarks and registration with the DCNN-Match CE approach. The source image has a large local deformation in the center along with small random deformations globally. The transformed source images obtained after DIR have been overlaid onto the target image (columns (b) and (d)) using complementary colors such that the aligned structures look grey and misalignment is highlighted in colors. As can be seen in column (b), many regions are not aligned properly after the registration, but, with the additional guidance information (column (d)), the anatomical structures look perfectly aligned. The corresponding landmark pairs are shown with cross-hairs of the same color in the target and source images. It is worth noting that DCNN-Match CE can find landmark correspondences in highly deformed regions as well. As a result, DIR with landmark correspondences can find a better estimation of the underlying deformation field as compared to the baseline DIR approach. Columns (c) and (e) represent the Root Mean Square Errors (RMSE) of the ground truth DVF and the DVF obtained after DIR without and with landmark correspondences. Further, Figure 3.12 shows an example of

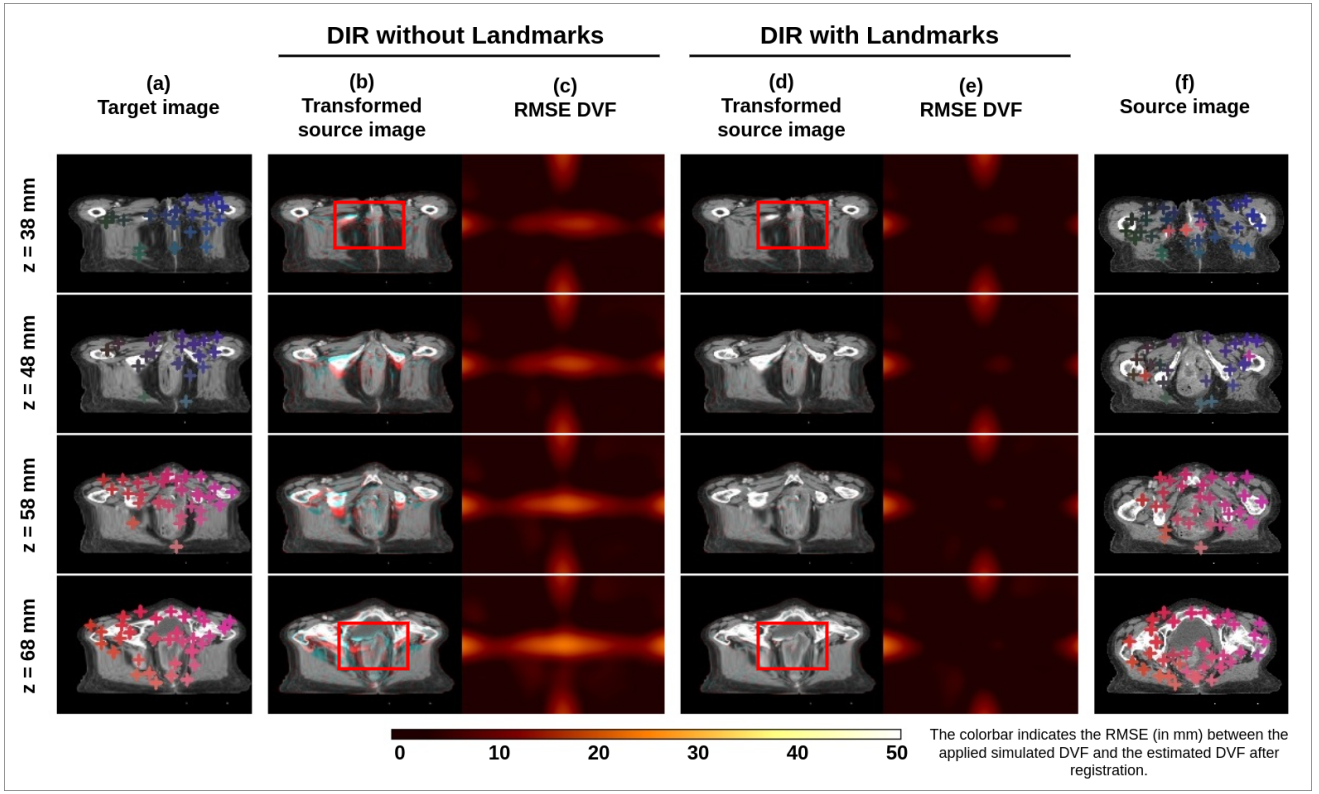


Figure 3.11: **Qualitative results on simulated deformations test set.** Transverse slices from 10 mm apart from a representative example are shown in different rows. Column (a): target image, columns (b) and (c): transformed source image and Root Mean Square Error (RMSE) plot between the ground truth and estimated DVF after registration without automatic landmarks, respectively, columns (d) and (e): transformed source image and RMSE plot between the ground truth and estimated DVF after registration with using automatic landmarks predicted by DCNN-Match CE, respectively, column (f): source image. Landmark correspondences between the target and source images are shown in similar colored cross-hairs in columns (a) and (f). Note: some of the landmarks may have correspondences in the transverse slices not shown in the figure. The red rectangles highlight the effect of using landmark correspondences in a highly deformed region.

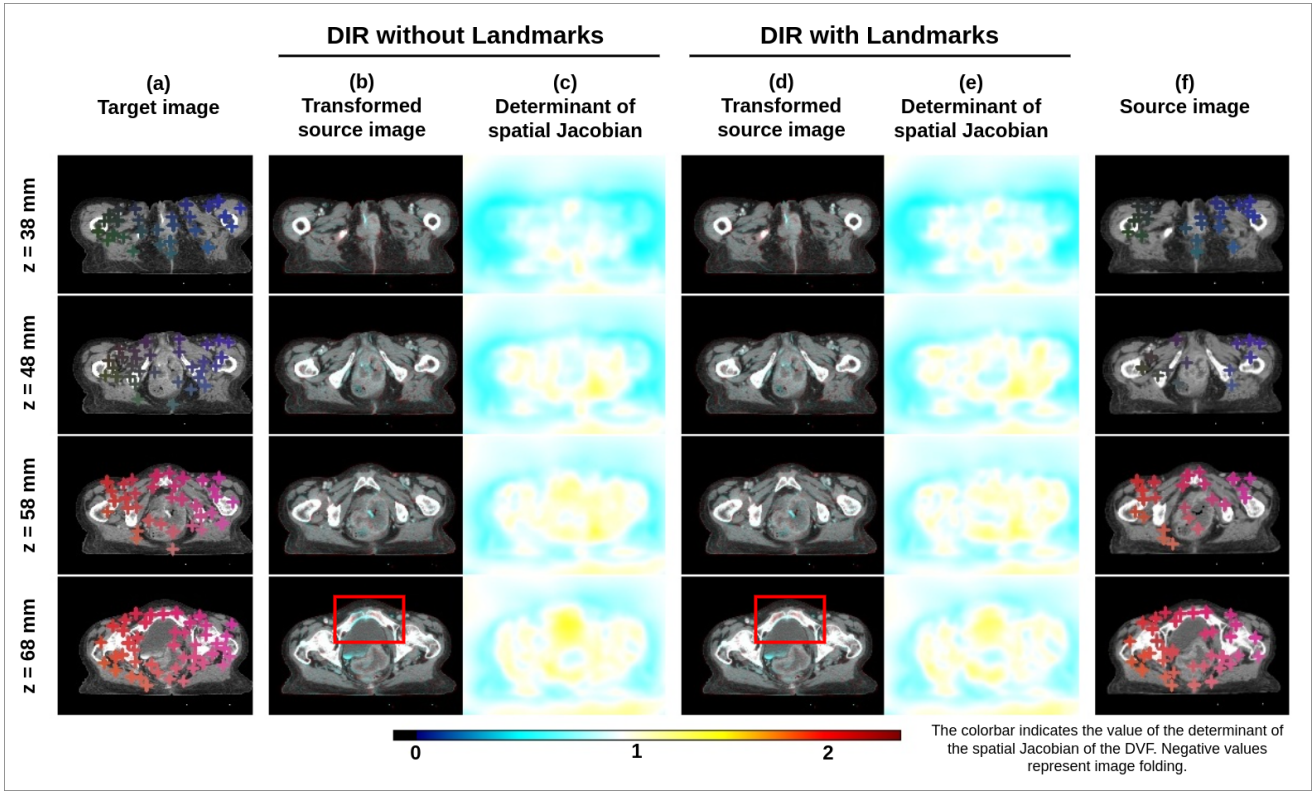


Figure 3.12: **Qualitative results on clinical deformations test set.** Transverse slices from 10 mm apart from a representative example are shown in different rows. Column (a): target image, columns (b) and (c): transformed source image and determinant of the spatial Jacobian after registration without automatic landmarks, respectively, columns (d) and (e): transformed source image and determinant of the spatial Jacobian after registration with using automatic landmarks predicted by DCNN-Match CE, respectively, column (f): source image. Landmark correspondences between the target and source images are shown in similar colored cross-hairs in columns (a) and (f). Note: some of the landmarks may have correspondences in the transverse slices not shown in the figure. The red rectangle highlights a region where improvement by adding landmarks correspondences in the DIR is visible.

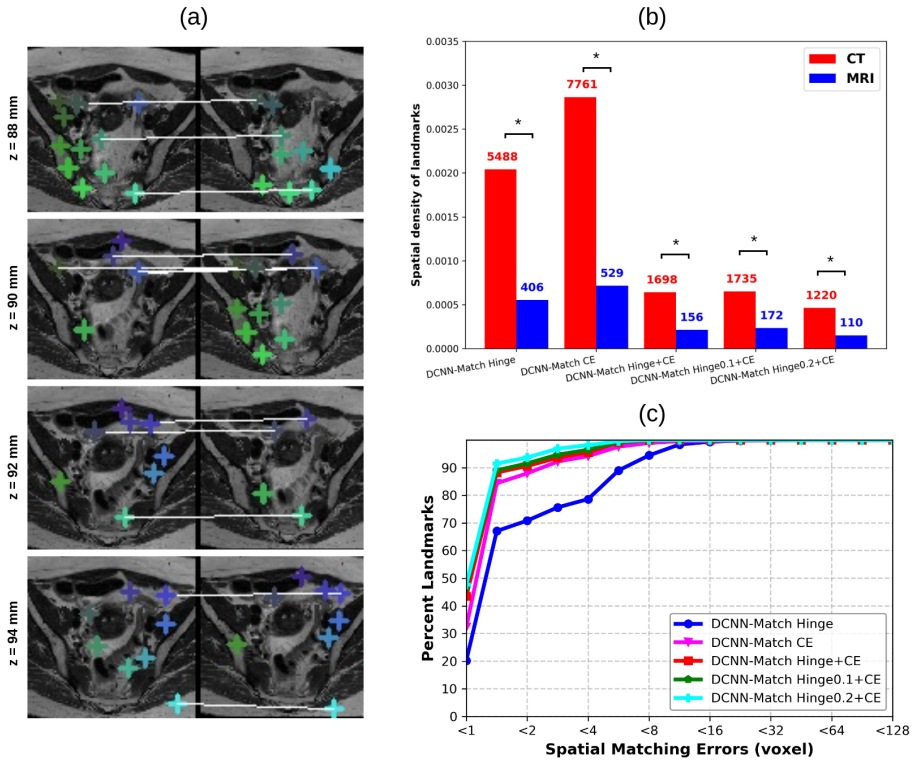


Figure 3.13: **Generalization results on the simulated deformations test set - MRI.** (a) Predicted corresponding landmarks in the target and source MRI. Corresponding landmarks are shown with similar colored cross-hairs in the target and source images. Note that some of the landmarks match across slices following the underlying deformation in 3D. (b) Comparison of the spatial density of predicted landmarks (averaged over all patients) between simulated deformations test set - CT and simulated deformations test set - MRI for each DCNN-Match variant. The average number of predicted landmarks is shown in the text above bars. * indicates significant difference after Mann-Whitney U test. (c) Spatial matching errors of predicted landmark correspondences.

DIR without and with using landmarks for clinical deformations. While the output of registration without and with using landmark correspondences looks similar in most cases, a subtle improvement in alignment can still be spotted in some regions of the images (also highlighted with a red rectangle in the figure) with the use of landmark correspondences in the DIR. The determinant of the spatial Jacobian shown in Figure 3.12 (c) and (e) shows no visible image folding in the DIR solutions obtained by either of the approaches.

3.3.8. GENERALIZATION TO MRI DATASET

A representative example of predicted landmark correspondences by DCNN-Match CE on MRI scans without retraining is shown in Figure 3.13 (a). Upon visual inspection, the predicted landmark correspondences seem to be accurate despite the different modality of the test scans. Further, the FOV of the acquisition of MRI scans was approximately 16

times smaller than the FOV of the acquisition of CT scans in the test set. To make a direct comparison between the number of predicted landmark correspondences in CT and MRI datasets, we calculated the spatial density of predicted landmarks by dividing the number of landmarks by the total number of voxels in each image. In CT scan images, a large portion of the image consists of background voxels where the DCNN-Match variants do not predict landmark correspondences. Therefore, we considered only the voxels in the patient's anatomy by counting the number of voxels in the largest connected component after binarizing the image through intensity thresholding.

The spatial density of predicted landmarks in both CT and MRI test sets is shown in Figure 3.13 (b). Since the networks were not trained on MRI scans, the spatial density of the predicted landmarks was reduced in MRI scans. Still, a considerable number of landmarks (on average for all patients) were predicted in the MRI test set by each approach (shown as the text above bars). Further, the spatial matching errors (shown in Figure 3.13 (c)) of the predicted landmark correspondences on MRI scans were comparable to the spatial matching errors observed for CT scans. Overall, the above results demonstrate the generalization potential of DCNN-Match on cross-modality data without retraining.

3.4. DISCUSSION

We developed a self-supervised deep learning method (DCNN-Match) for automatic landmarks correspondence detection in 3D medical images. We have also presented quantitative and qualitative evidence that a high number of landmark correspondences with good spatial matching accuracy can be predicted within seconds with the help of our proposed approach. Furthermore, we integrated DCNN-Match with a DIR pipeline and assessed the added value of automatic landmark correspondences toward the improvement of intra-patient DIR performance. To the best of our knowledge, this is the first study to develop a self-supervised deep learning approach for predicting automatic landmark correspondences in 3D medical images and investigating their applicability in improving DIR.

We developed five variants of the proposed approach, which differed in the way feature descriptor matching is learned. We observed that a separate module for learning feature descriptor matching (DCNN-Match CE) yields landmark correspondences with not only reduced spatial matching errors but also an increased number of matches in regions of high deformation. The results also showed that the added value to the performance of DIR was most prominent by the use of automatic landmark correspondences predicted by DCNN-Match CE. While three other variants predicted automatic landmark correspondences with better spatial matching accuracy than DCNN-Match CE, the numbers of predicted landmarks by these variants were fewer than the number of landmarks predicted by DCNN-Match, especially in regions of high deformation. This implies that the spatial density of predicted landmarks with respect to the underlying deformation plays a role in the extent of the added value provided by the automated landmark correspondences.

The results also showed that the additional guidance by automatic landmark correspondences improved the performance of DIR irrespective of the variance in the number, spatial matching errors, and spatial distribution of the automatic landmarks in both simulated as well as clinical deformations test sets. These findings are in line with the existing literature on the use of automatic landmarks for the improvement of DIR in chest CT [9, 26, 29], head and neck CT [20], retinal images [18], and brain MRI images [15, 16]. A study on DIR of thoracic CT scans [30] reported that automatic landmarks-based optimization of the regularization parameter reduced the TRE of expert landmarks on average by 0.07 mm. Another study on registration of CT scans corresponding to end-inspiration and end-expiration phases reported a reduction of TRE of expert landmarks from 1.34 ± 2.00 mm to 0.82 ± 0.97 mm by the use of automatic landmarks in DIR [29]. Our experiments showed that the TRE of validation landmarks in the simulated deformations test set reduced from 5.07 ± 9.98 to 3.14 ± 8.61, and the TRE of expert landmarks in the clinical deformations test set reduced from 6.85 ± 5.79 to 6.42 ± 5.79 on average by the use of automatic landmark correspondences predicted by DCNN-Match CE in DIR. Since the improvement in DIR performance reported in terms of TRE values of the expert landmarks is affected by several factors e.g., the number and location of the expert landmarks, image resolution, and TRE values before registration, a comparison in absolute values of TRE improvement cannot be made. Nevertheless, the current study adds to the existing evidence on the added value of automatic landmark correspondences in improving DIR by providing experimental results from pelvic CT scan registrations, which otherwise did not exist.

Two other studies have looked into intra-patient DIR in cervical cancer patients [27, 4]. The authors in one of the studies [27] have focused on dose mapping and do not report TRE values. The average TRE values after registration reported in the other study [4] are the following: 3.5 ± 2.4 mm for bladder top, 8.5 ± 5.2 mm for cervix tip, 5.7 ± 2.1 mm for markers, and 4.6 ± 2.2 mm for the midline. As such, a direct correspondence between the landmarks used in our study and landmarks in the earlier study cannot be ascertained. Moreover, the underlying dataset and methods used are also different. Still, the mean TRE value obtained after registration with additional guidance information from landmark correspondences predicted by DCNN-MatchCE (6.42 ± 5.79 mm) seems to be within the range of reported TRE values, which gives some confidence that the obtained DIR results are satisfactory.

The extent of the added value provided by the use of automatic landmark correspondences in DIR was lower in the clinical deformations test set as compared to the simulated deformations test set. Our retrospective analysis (provided in Appendix 3.6.3) revealed no obvious patterns regarding the spatial distribution of the automatic landmarks in relation to manual landmarks used for TRE calculations that could explain the lower added value of using automatic landmarks in the clinical deformations test set. The DIR performance in case of clinical deformations as reflected by TRE of manually annotated landmarks is affected by several factors e.g., choice of manual landmarks, inter- and intra-observer variation in the placement of manual landmarks, hyperparameters in the parameter map used for Elastix, limitations of Elastix in modeling large deformations, sliding tissue, and singularities in DVF.

Therefore, establishing a direct relationship between the quality of automatic landmark correspondences and the DIR performance is difficult. However, we can speculate on a few factors that impacted the quality of automatic landmark correspondences and hence could have impacted the added value to DIR. In the clinical test set, the CT scans were acquired with contrast administered via a rectal tube or intravenously. Consequently, one or multiple regions (e.g., vagina, bladder, bowel bag, or vascular regions) were contrast-enhanced giving rise to large differences in appearance between the CT scan pairs, which was not a part of the training for DCNN-Match. An example of appearance variation due to contrast is shown in Figure 3.4 (b). This appearance variance between the source and target CT scans often overlapped with the large and complex deformations in the bladder and bowel bag. This posed an additional challenge for finding landmark correspondences between scans. Although all DCNN-Match variants were still able to find landmark correspondences in these scans despite the aforementioned challenges, they failed to find correspondences in regions where appearance was strongly different due to a combination of contrast administration and underlying deformation. We expect that incorporating a model for simulating contrast differences between scans and a better (probably a bio-mechanical based) model for simulating deformations due to physical phenomena such as bladder filling would lead to the prediction of automatic landmarks in the aforementioned challenging scenario as well and yield a larger added value of using automatic landmark correspondences in DIR. We are considering pursuing this direction for a future study.

Another factor affecting the DIR performance in the clinical deformations test set is that we tuned the hyperparameters used in Elastix (weights of the objectives used in DIR, $weight_1$, and $weight_2$) based on the DIR of CT scan pairs in the validation set consisting of simulated deformations. We used these hyperparameters for all the registrations in both simulated as well as clinical deformation test sets. This does not acknowledge the fact that each DIR problem is unique and therefore, a single setting for all source and target pairs is sub-optimal. Earlier research has also pointed out the importance of tuning the weights of different objectives in the DIR separately for each image pair to achieve the best DIR performance [30, 25]. We conducted retrospective experiments by changing the weights of the objectives in DIR, which revealed that $weight_1$, and $weight_2$ values corresponding to best DIR performance (quantified in terms of minimum TRE values) were indeed different for each CT scan pair in the clinical deformations test set. Unfortunately, the tuning of $weight_1$, and $weight_2$ separately for each CT scan pair in the clinical deformations test set could not be done objectively and automatically due to the unavailability of the underlying ground truth. Note that the manually annotated landmarks were used to evaluate the DIR performance and therefore using them for tuning $weight_1$, and $weight_2$ would have produced biased results. However, the purpose of this research was not to obtain the best DIR performance for each CT scan pair but to quantify the effect of additional guidance provided by the automatic landmark correspondences.

Further, the added value of the additional guidance provided by the automatic landmark correspondences may be limited by erroneous matches. While the results on the simulated data indicated the benefits of more landmarks toward DIR performance,

the adverse effect of erroneous matches remains unclear. It would be interesting to investigate in a future study how much value can be gained by removing the erroneous landmark matches either using RANSAC [11] or a deep learning approach [40]. Another interesting direction for future research can be to simultaneously learn a deep learning model for landmark matching as well as performing DIR. Such a model can be used to investigate how many landmarks are optimal for improving the DIR. However, care needs to be taken to avoid degeneracy because landmark matching essentially is performing DIR on a sparse grid and the optimal number of landmark matches to improve DIR could quite likely be the total number of voxels in the image.

Remarkably the proposed approach for finding landmark correspondences could find landmark correspondences on cross-modality data without retraining. Based on this observation, we expect that with retraining (which requires minimal effort because manual annotations are not needed), the proposed approach should be able to find automatic landmark correspondences on any type of medical imaging data. Furthermore, since a considerable number of landmarks were predicted in the MRI scans with spatial matching errors comparable to the CT scans, we expect that the use of automatic landmarks should lead to performance gain in DIR on MRI scans also. With retraining on MRI scans, we expect that the added value to the DIR performance will be similar to as observed in the CT scans.

3.5. CONCLUSIONS

We developed a self-supervised method for automatic landmarks correspondence detection in abdominal CT scans and investigated the effect of different variants of our automatic landmarks correspondence detection approach on the performance of DIR. The obtained results provide strong evidence for the added value of using automatic landmark correspondences in providing additional guidance information to DIR. The added value of automatic landmarks in DIR is consistent across different variants of our approach and for both simulated as well as clinical deformations. Additionally, we observed that the spatial distribution of automatic landmark correspondences with respect to the underlying deformation has a considerable effect on the extent of the added value provided by landmark correspondences. A higher number of automatic landmark correspondences in highly deformed regions has more added value than more accurate but fewer landmark correspondences. Therefore, further research in the direction of developing landmark detection approaches that are aware of the underlying deformation is recommended.

In conclusion, the current study affirms the added value of using automatic landmark correspondences for solving challenging DIR problems and provides insights into what type of landmark correspondences (in terms of spatial distribution and matching errors) may be more beneficial to DIR than others.

3.6. APPENDIX

3.6.1. ELASTIX PARAMETER MAPS

AFFINE REGISTRATION

```
(AutomaticParameterEstimation "true")
(AutomaticTransformInitialization "true")
(AutomaticTransformInitializationMethod "Origins")
(CheckNumberOfSamples "true")
(DefaultPixelValue 0)
(FinalBSplineInterpolationOrder 1)
(FixedImagePyramid "FixedSmoothingImagePyramid")
(ImageSampler "RandomCoordinate")
(Interpolator "LinearInterpolator")
(MaximumNumberOfIterations 1024)
(MaximumNumberOfSamplingAttempts 8)
(Metric "AdvancedMattesMutualInformation")
(MovingImagePyramid "MovingSmoothingImagePyramid")
(NewSamplesEveryIteration "true")
(NumberOfResolutions 4)
(NumberOfSamplesForExactGradient 4096)
(NumberOfSpatialSamples 4096)
(Optimizer "AdaptiveStochasticGradientDescent")
(Registration "MultiResolutionRegistration")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")
(Transform "AffineTransform")
```

DEFORMABLE IMAGE REGISTRATION

```
(AutomaticParameterEstimation "true")
(BSplineInterpolationOrder 1)
(CheckNumberOfSamples "true")
(DefaultPixelValue 0)
(FinalBSplineInterpolationOrder 1)
(FinalGridSpacingInPhysicalUnits 8)
(FixedImageDimension 3)
(FixedImagePixelType "float")
(FixedImagePyramid "FixedRecursiveImagePyramid")
(HowToCombineTransforms "Compose")
(ImageSampler "RandomCoordinate")
(Interpolator "BSplineInterpolator")
(MaximumNumberOfIterations 300 600 900 1200)
(Metric "AdvancedMattesMutualInformation" "TransformBendingEnergyPenalty"
        "CorrespondingPointsEuclideanDistanceMetric")
(Metric0Weight 1)
(Metric1Weight 1)
(Metric2Weight 0.01)
(MovingImageDimension 3)
(MovingImagePixelType "float")
(MovingImagePyramid "MovingRecursiveImagePyramid")
(NewSamplesEveryIteration "true" "true" "true" "true")
(NumberOfHistogramBins 32 32 32 32)
(NumberOfResolutions 4)
(NumberOfSpatialSamples 5000 5000 5000 5000)
(Optimizer "StandardGradientDescent")
(Registration "MultiMetricMultiResolutionRegistration")
(ResampleInterpolator "FinalBSplineInterpolator")
```

```
(Resampler "DefaultResampler")
(SP_A 100 200 300 400)
(SP_a 35000 30000 25000 20000)
(SP_alpha 0.602 0.602 0.602 0.602)
(ShowExactMetricValue "false" "false" "false" "false")
(Transform "BSplineTransform")
(UpsampleGridOption "true")
```


3.6.2. LIST OF MANUALLY ANNOTATED LANDMARKS

- fiducial markers in the vaginal wall near the cervix at the locations: posterior left, anterior mid, posterior right, posterior mid, anterior left, and anterior right
- bifurcation aorta
- os coccygis
- medial tip of right and left trochanter minor
- most caudal, dorsal, and ventral part of the corpus of lumbar vertebrae 3
- most caudal, dorsal, and ventral part of the corpus of lumbar vertebrae 5
- right and left bifurcation vena iliaca communis
- right and left bifurcation of artery iliaca communis
- umbilicus
- caudal tip of right and left kidney
- external and internal anal sphincter
- cervical ostium
- external and internal urethral ostium
- right and left ureteral ostium
- uterus top

3.6.3. RETROSPECTIVE ANALYSIS

The extent of the added value provided by the use of automatically-identified landmark correspondences in DIR was lower in the clinical deformations test set as compared to the simulated deformations test set. Therefore, we analyzed the TRE values of each manual landmarks in the clinical deformations test set to understand the possible causes for the lack of performance gain by using automatic landmarks in DIR. Specifically, we calculated the number of automatic landmarks in proximity (16 mm) to each of the manual landmark. We plotted this value against the TRE_{before} value (representative of the underlying deformation in that region) of that manual landmark (shown in Figure 3.14 (a)). The plot shows that automatic landmarks were predicted in the regions of high deformation as well, especially by DCNN-Match CE. Therefore, a lack of the presence of automatic landmarks in highly deformed regions could not be the sole cause for the lack of performance gain in DIR.

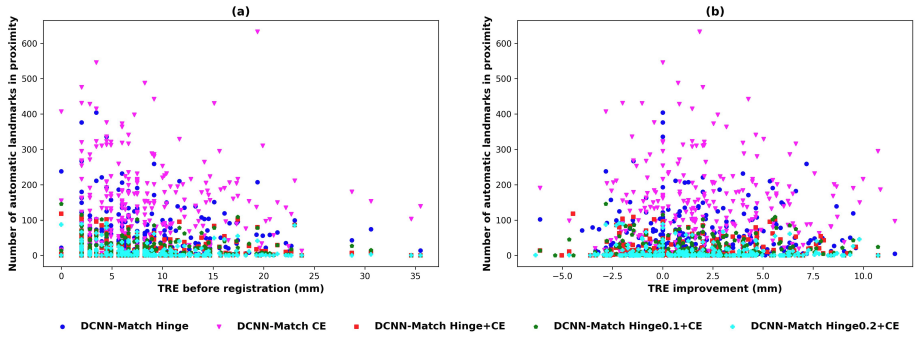


Figure 3.14: **Results on clinical deformations test set.** For each manual landmark, the number of automatic landmark correspondences predicted in 16 mm proximity to that manual landmark has been plotted against (a) the corresponding TRE_{before} value and (b) TRE improvement value as obtained by subtracting TRE_{after} value from TRE_{before} value.

Further, we calculated the TRE improvement for each manual landmark by subtracting TRE_{after} from TRE_{before} values. A positive high number indicates higher improvement in TRE values (or DIR performance). In Figure 3.14 (b), the TRE improvement values have been plotted against the number of automatic landmarks in proximity for each manual landmark. We observed that the TRE value of some of the manual landmarks in some of the patients did not improve despite the presence of automatic landmarks in their proximity.

In conclusion, the above analysis shows that a straightforward pattern regarding the spatial distribution of automatic landmarks relative to the manual landmarks cannot be established in case of clinical deformations. Consequently, a direct relationship between the quality of automatic landmark correspondences and the DIR performance cannot be established.

BIBLIOGRAPHY

[1] Tanja Alderliesten, Peter A. N. Bosman, and Arjan Bel. "Getting the most out of additional guidance information in deformable image registration by leveraging multi-objective optimization". In: *Medical Imaging 2015: Image Processing*. Vol. 9413. Proc. SPIE. International Society for Optics and Photonics. 2015, 94131R.

[2] Riddhish Bhalodia et al. "Leveraging unsupervised image registration for discovery of landmark shape descriptor". In: *Medical Image Analysis* 73 (2021), p. 102157.

[3] Bastian Bier et al. "X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 55–63.

[4] Luiza Bondar et al. "A symmetric nonrigid registration method to handle large organ deformations in cervical cancer patients". In: *Medical Physics* 37.7, Part 1 (2010), pp. 3760–3772.

[5] Ming Chao, Yaoqin Xie, and Lei Xing. "Auto-propagation of contours for adaptive prostate radiation therapy". In: *Physics in Medicine & Biology* 53.17 (2008), p. 4533.

[6] Indrin J Chetty and Mihaela Rosu-Bubulac. "Deformable registration for dose accumulation". In: *Seminars in Radiation Oncology*. Vol. 29. 3. Elsevier. 2019, pp. 198–208.

[7] Jay Devine et al. "A registration and deep learning approach to automated landmark detection for geometric morphometrics". In: *Evolutionary Biology* 47.3 (2020), pp. 246–259.

[8] Olive Jean Dunn. "Multiple comparisons using rank sums". In: *Technometrics* 6.3 (1964), pp. 241–252.

[9] Jan Ehrhardt et al. "Automatic landmark detection and non-linear landmark-and surface-based registration of lung CT images". In: *Medical Image Analysis for the Clinic-A Grand Challenge, MICCAI* 2010 (2010), pp. 165–174.

[10] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), p. 115.

[11] Martin A Fischler and Robert C Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6 (1981), pp. 381–395.

[12] Florin C. Ghesu et al. "An Artificial Agent for Anatomical Landmark Detection in Medical Images". In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*. Ed. by Sebastien Ourselin et al. Cham: Springer International Publishing, 2016, pp. 229–237. ISBN: 978-3-319-46726-9.

[13] Soumya Ghose et al. “A review of segmentation and deformable registration methods applied to adaptive cervical cancer radiation therapy treatment planning”. In: *Artificial Intelligence in Medicine* 64.2 (2015), pp. 75–87.

[14] Varun Gulshan et al. “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs”. In: *JAMA* 316.22 (2016), pp. 2402–2410. ISSN: 0098-7484. DOI: 10.1001/jama.2016.17216.

[15] Dong Han et al. “Robust anatomical landmark detection for MR brain image registration”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2014, pp. 186–193.

[16] Dong Han et al. “Robust anatomical landmark detection with application to MR brain image registration”. In: *Computerized Medical Imaging and Graphics* 46 (2015), pp. 277–290. ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.compmedimag.2015.09.002. URL: http://www.sciencedirect.com/science/article/pii/S089561111500124X.

[17] Kaiming He et al. “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1026–1034.

[18] Álvaro S. Hervella et al. “Multimodal registration of retinal images using domain-specific landmarks and vessel enhancement”. In: *Procedia Computer Science* 126 (2018). Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia, pp. 97–104. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2018.07.213. URL: https://www.sciencedirect.com/science/article/pii/S1877050918311876.

[19] IBM Corp. *IBM SPSS Statistics*. Version 27. 2020. URL: https://www.ibm.com/nl-en/analytics/spss-statistics-software.

[20] Vasant Kearney et al. “Automated landmark-guided deformable image registration”. In: *Physics in Medicine & Biology* 60.1 (2014), p. 101.

[21] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations*. 2015. URL: http://arxiv.org/abs/1412.6980.

[22] Stefan Klein et al. “Elastix: A toolbox for intensity-based medical image registration”. In: *IEEE Transactions on Medical Imaging* 29.1 (Jan. 2010), pp. 196–205.

[23] K. Marstal et al. “SimpleElastix: A user-friendly, multi-lingual library for medical image registration”. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 574–582.

[24] Adam Paszke et al. “Automatic differentiation in PyTorch”. In: *Advances in Neural Information Processing Systems-W*. 2017. URL: https://github.com/pytorch/pytorch.

[25] Kleopatra Pirpinia et al. “The feasibility of manual parameter tuning for deformable breast MR image registration from a multi-objective optimization perspective”. In: *Physics in Medicine & Biology* 62.14 (2017), p. 5723.

[26] Thomas Polzin et al. “Combining automatic landmark detection and variational methods for lung CT registration”. In: *Fifth International Workshop on Pulmonary Image Analysis*. 2013, pp. 85–96.

[27] Bastien Rigaud et al. “Deformable image registration for dose mapping between external beam radiotherapy and brachytherapy images of cervical cancer”. In: *Physics in Medicine & Biology* 64.11 (2019), p. 115023.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[29] J. Rühaak et al. “Estimation of Large Motion in Lung CT by Integrating Regularized Keypoint Correspondences into Dense Deformable Registration”. In: *IEEE Transactions of Medical Imaging* 36.8 (2017), pp. 1746–1757. ISSN: 0278-0062. DOI: 10.1109/TMI.2017.2691259.

[30] Alexander Schmidt-Richberg et al. “Landmark-driven parameter optimization for non-linear image registration”. In: *Medical Imaging 2011: Image Processing*. Ed. by Benoit M. Dawant and David R. Haynor. Vol. 7962. International Society for Optics and Photonics. SPIE, 2011, 79620T. DOI: 10.1117/12.877059. URL: https://doi.org/10.1117/12.877059.

[31] Denis P Shamonin et al. “Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer’s disease”. In: *Frontiers in Neuroinformatics* 7.50 (2014), pp. 1–15.

[32] P. Thevenaz and M. Unser. “Optimization of mutual information for multiresolution image registration”. In: *IEEE Transactions on Image Processing* 9.12 (2000), pp. 2083–2099. DOI: 10.1109/83.887976.

[33] Maria Thor et al. “Evaluation of an application for intensity-based deformable image registration and dose accumulation in radiotherapy”. In: *Acta Oncologica* 53.10 (2014), pp. 1329–1336.

[34] Ahmet Tuysuzoglu et al. “Deep Adversarial Context-Aware Landmark Detection for Ultrasound Imaging”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 151–158.

[35] Tom Vercauteren et al. “Diffeomorphic demons: Efficient non-parametric image registration”. In: *NeuroImage* 45.1 (2009), S61–S72.

[36] Ola Weistrand and Stina Svensson. “The ANACONDA algorithm for deformable image registration in radiotherapy”. In: *Medical Physics* 42.1 (2015), pp. 40–53.

[37] René Werner et al. “Assessing accuracy of non-linear registration in 4D image data using automatically detected landmark correspondences”. In: *Medical Imaging 2013: Image Processing*. Vol. 8669. Proc. SPIE. International Society for Optics and Photonics. 2013, 86690Z.

[38] Ke Yan et al. “SAM: Self-Supervised Learning of Pixel-Wise Anatomical Embeddings in Radiological Images”. In: *IEEE Transactions on Medical Imaging* 41.10 (2022), pp. 2658–2669. DOI: 10.1109/TMI.2022.3169003.

[39] Deshan Yang et al. “A method to detect landmark pairs accurately between intra-patient volumetric medical images”. In: *Medical Physics* 44.11 (2017), pp. 5859–5872.

[40] Kwang Moo Yi et al. “Learning to find good correspondences”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2666–2674.

4

ORGANS AT RISK SEGMENTATION

Deep learning models benefit from training with a large dataset (labeled or unlabeled). Following this motivation, we present an approach to learn a deep learning model for the automatic segmentation of Organs at Risk (OARs) in cervical cancer radiation treatment from a large clinically available dataset of Computed Tomography (CT) scans containing data inhomogeneity, label noise, and missing annotations. We employ simple heuristics for automatic data cleaning to minimize data inhomogeneity and label noise. Further, we develop a semi-supervised learning approach utilizing a teacher-student setup, annotation imputation, and uncertainty-guided training to learn in presence of missing annotations. Our experimental results show that learning from a large dataset with our approach yields a significant improvement in the test performance despite missing annotations in the data. Further, the contours generated from the segmentation masks predicted by our model are found to be equally clinically acceptable as manually generated contours.

The content of this chapter is based on the following publication: **Grewal, M.**, van Weersel, D., Westerveld, H., Bosman, P. A. N., & Alderliesten, T. (2024, January). *Learning Clinically Acceptable Segmentation of Organs at Risk in Cervical Cancer Radiation Treatment from Clinically Available Annotations.* In Medical Imaging with Deep Learning (pp. 260-273). PMLR.

4.1. INTRODUCTION

The planning for cervical cancer radiation treatment[1] requires manual contouring of the Organs at Risk (OARs) where the adverse effects of radiation must be minimized. Automatic segmentation of these OARs can save hours of manual work. In this chapter, we focus on the automatic segmentation of four OARs in cervical cancer radiation treatment: bowel bag, bladder, hips, and rectum. A few studies have focused on developing deep learning based automatic OARs segmentation methods for cervical cancer radiation treatment [9, 8, 16, 10, 13]. All of these studies use a traditional setup for developing a deep learning model, which involves: (a) obtaining a fully annotated clinically available dataset, (b) splitting the data into training, validation, and testing, and (c) training a model and evaluating it on the test dataset. A major drawback in this setup is the limited size of the datasets used for training and testing. A small training dataset limits the possibility of a deep learning model capturing large variance in real-world data. Further, evaluation results from a small test dataset do not inform sufficiently in regard to the true test performance of a deep learning model. Although in the medical imaging domain, such a setup is understandable because of the underlying requirement of clinical expertise for annotating the data, it would be of interest to investigate if clinically available data can be leveraged to increase the size of the training and testing datasets.

The size of the training dataset for automatic OARs segmentation for cervical cancer radiation treatment can be increased if the abdominal scans acquired for tumors other than cervical cancer are also included. However, all the OARs in cervical cancer radiation treatment may not be annotated in those scans. Furthermore, since the clinically available abdominal scans and annotations are retrospectively included, the acquisition protocols, contouring guidelines, and observers may be different giving rise to data inhomogeneity and label noise. In this chapter, we follow the motivation of harnessing the benefits of training on a large dataset. Therefore, we use the Computed Tomography (CT) scans and OARs contours delineated for clinical use during radiation treatment for tumors in the abdominal region to develop a deep learning model for segmentation of OARs in cervical cancer radiation treatment. We develop a semi-supervised learning approach to tackle the issue of missing annotations in data. Briefly, the key contributions of our work are the following:

1. We propose a teacher-student setup, wherein, the predictions from a teacher model are used to impute the missing annotations, and a student model is trained using the dataset containing imputed annotations. Additionally, we train the student with an uncertainty-guided loss to avoid the adverse effect of imperfect predictions from the teacher, and with additional augmentations to increase performance.

2. We perform an ablation study to investigate the effect of different components of the proposed approach. Furthermore, we perform a clinical validation study to assess the clinical acceptability of contours generated from automatic segmentation masks predicted by our deep learning model.

[1] Radiation treatment for cancer involves giving high doses of radiation to the tumor to kill cancer cells.

4.1.1. RELATED WORK

Our approach is closely related to previous works in the direction of semi-supervised learning by generation of pseudo-labels and self-training for medical image segmentation tasks [1, 6, 7, 19]. Different from these works, we use self-training with pseudo-labels in a teacher-student setup similar to [15, 18]. Further, we utilize uncertainty maps to reduce the adverse effect of imperfect pseudo-labels, which have been previously used in [15, 18, 19]. In contrast to [15, 18], we train a noisy student with the use of additional augmentations in the data because it has been shown to provide performance gain [17]. In the domain of learning an OARs segmentation model for cervical cancer radiation therapy by utilizing a large dataset, our work is similar to [12]. However, instead of learning a separate model for each OAR as in [12], we learn a single model for the segmentation of all OARs, which increases the potential for real-world deployment of our model.

4.2. DATA

We retrospectively selected the CT scans of female patients who were treated in an academic hospital for a tumor in the abdominal region from 2009 to 2019. A total of 1170 CT scans with associated clinically available contours from 1108 patients were received in anonymized form through a data transfer agreement. These scans were used for training and validation. For testing, we used 105 CT scans with associated clinically available contours from 95 cervical cancer patients who received radiation treatment in the same hospital.

4.2.1. PREPROCESSING

In all the CT scans (1170 from the training and validation dataset, and 105 from the test dataset), the clinically available annotations of four OARs in cervical cancer radiation treatment (bowel bag, bladder, hips, and rectum) were extracted by using the following steps: (1) standardize different variations of organ labels (e.g., bowel, bowel bag, Bowel bag, bowel_bag, Bowel_bag were all considered bowel bag), (2) combine left and right hip annotations as a single organ, (3) remove voxels annotated as bladder or rectum from the bowel bag annotation to avoid ambiguous labeling in those voxels. Next, the scans were resampled to 2.5mm×2.5mm×2.5mm voxel spacing. The Hounsfield units were converted to intensity values between 0 and 1 by windowing (window level=40, window width=400). In the training and validation dataset, the preprocessing resulted in a total of 186 scans that contained annotations for all the four OARs considered in this work (referred to as the fully annotated dataset, $\mathscr{D}_f$). The remaining scans had missing annotations for at least one of the OARs (referred to as the partially annotated dataset, $\mathscr{D}_p$). In total 383, 1103, 504, and 865 scans had annotations for bowel bag, bladder, hips, and rectum, respectively.

4.2.2. AUTOMATIC DATA CLEANING

Since the data was accumulated over 10 years and the scans belonging to patients who were treated for a tumor anywhere in the abdominal region were included, the data exhibited inhomogeneity in the cranial extent of the scan (causing an increase in the

number of background voxels and potentially less efficient training), and the cranial border of the bowel bag annotations (attributing to label noise).

To make the data more homogeneous so that the adverse effects of inefficient training and label noise could be reduced, we analyzed the histograms of $\mathscr{D}_f$ and decided on thresholds such that the histograms represented a unimodal distribution corresponding to the most frequently used scanning protocol and annotation style (details are provided in Appendix 4.7.2). Based on these thresholds, the scans were cropped in the cranial direction to remove the chest region. The bowel bag annotations in the abdominal region roughly above the level of the lumbar (L4) spinal segment were deleted. The scans that did not contain bowel bag annotations in the entire pelvic region were discarded. These steps resulted in a decrease in the size of $\mathscr{D}_f$ from 186 to 134. The resulting dataset of 134 scans is referred to as $\mathscr{D}_f^{clean}$ in the rest of the chapter.

4.3. APPROACH

We developed a semi-supervised learning approach utilizing a teacher-student setup (Figure 4.1). We train a teacher model using the small, fully annotated dataset ($\mathscr{D}_f^{clean}$). The predictions from the trained teacher model are used to impute the remaining large dataset with missing annotations ($\mathscr{D}_p$). Then, a student is trained with the entire dataset ($\mathscr{D}_f^{clean} + \mathscr{D}_p$) containing the clinically available and imputed annotations.

4.3.1. UNCERTAINTY-GUIDED TRAINING

Epistemic uncertainty refers to the lack of knowledge in a model about the underlying data. Estimating epistemic uncertainty enables the estimation of the reliability of a model's prediction for a specific sample. We train the teacher model to also estimate the epistemic uncertainty maps for each sample. For this purpose, we use a K-head neural network, similar to [19]. At each iteration of training, a single head is selected randomly for backpropagation. During inference, we use the mean prediction from K-heads as confidence and the entropy of the mean prediction as an estimate of epistemic uncertainty. We selected the K-head approach because it allows independence between predictions from different heads with faster inference times as compared to the Monte-Carlo (MC) dropout approach [2]. Moreover, the memory overhead is not much compared to fully independent deep ensembles [5].

We train the student model with an uncertainty-guided cross-entropy loss $\mathscr{L}_{uCE} = e^{-u} y \cdot log(\hat{y})$, where u is uncertainty in the teacher's predictions at each voxel, e^{-u} is the uncertainty-guided weight, y is the reference label, and $\hat{y}$ is the predicted probability. The weight e^{-u} ensures a large weight on voxels where the uncertainty in the teacher's predictions is small and vice-versa. We set $u = 0$ at the voxels where annotations are clinically available. In this way, the student model can benefit from training with a large dataset while avoiding deterioration in performance due to uncertain label predictions from the teacher model.

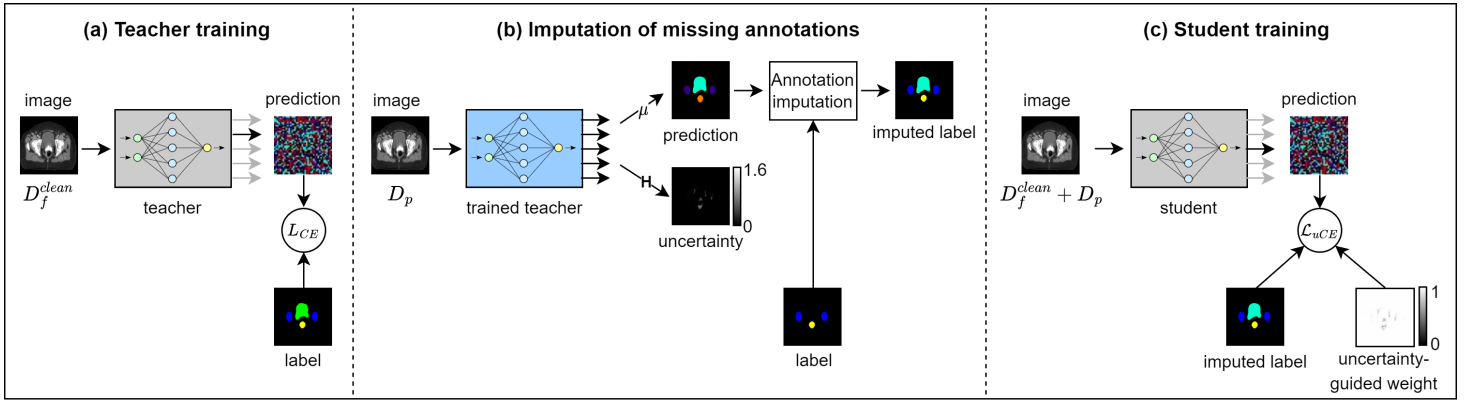


Figure 4.1: Schematic of the proposed approach. (a) A K-head (depicted by output arrows) teacher model is trained by randomly selecting a single head (highlighted in black) for backpropagation. (b) The clinically available 'label' contains annotation for hips (blue) and rectum (yellow) only. The annotation for bladder is missing. The mean prediction (of K-heads) from the trained teacher is used to impute the bladder annotation. (c) A K-head student model is trained with imputed label and uncertainty-guided loss. μ: mean, H: entropy, L_{CE}: cross-entropy loss, $\mathscr{L}_{uCE}$: uncertainty-guided loss.

4.3.2. IMPLEMENTATION DETAILS

As a baseline, we used the original U-Net architecture [14] after replacing the 2D convolutional layers with 3D convolutional layers and adding a batch normalization layer after each convolutional layer. The training was done using randomly cropped 3D patches (of depth 32 along the transverse direction) with a batchsize of 1 because of the GPU memory constraints. The implementation[2] was done in Python by using the PyTorch library [11] and the training was done on NVIDIA RTX2080 GPUs. Other hyperparameters were: optimizer=Adam [4]; network initialization=Kaiming He [3]; learning rate (LR)=$1e^{-3}$; weight decay=$1e^{-4}$; the number of training epochs=500 for teacher models, 250 for student models; learning schedule=step LR with step size=$\frac{1}{3}\times$total training steps; data augmentations=global brightness and contrast variations (±20%), random rotations (-10° to 10° along all axes); the number of heads (K) in teacher and student=5.

Method	Dice (%)	Surface Dice (%)	HD
3D U-Net + $\mathscr{D}_f$	83.47 (6.16)	80.23 (6.82)	16.06 (9.07)
3D U-Net + $\mathscr{D}_f^{clean}$	85.02 (5.92)*	82.00 (6.55)*	12.44 (10.58)*
basic teacher	85.36 (5.54)*	82.33 (6.18)*	11.61 (7.94)*
basic student	87.01 (4.62)*†	84.64 (5.18)*†	10.64 (8.00)*†
robust teacher	85.31 (5.25)*	82.30 (5.72)*	11.57 (7.73)*
basic teacher + robust student	87.11 (4.28)*†	84.76 (4.85)*†	10.39 (6.68)*†
robust teacher + robust student	87.16 (4.19)*†	84.82 (4.68)*†	9.92 (4.72)*†
robust teacher + robust student - iter. 2	87.40 (4.13)*†	85.30 (4.60)*†	9.85 (4.86)*†
robust teacher + robust student - iter. 3	87.35 (4.10)*†	85.24 (4.63)*†	9.96 (4.84)*†

Table 4.1: Mean (standard deviation) of mean test performance per scan of the best models obtained from 5-fold cross-validation. Aug.: additional augmentations, HD: Hausdorff distance in mm at 95 percentile. Surface Dice were computed at a tolerance of 2.5mm (voxel spacing). *significant differences compared to 3D U-Net + $\mathscr{D}_f$, †significant differences compared to 3D U-Net + $\mathscr{D}_f^{clean}$.

4.4. ABLATION EXPERIMENT

We conducted an ablation experiment to look into the individual effect of the components of our approach. As a baseline, we used two models: 3D U-Net trained with $\mathscr{D}_f$, and 3D U-Net trained with $\mathscr{D}_f^{clean}$. Note that the 3D U-Net trained with $\mathscr{D}_f^{clean}$ is similar to the traditional setup of deep learning model development. In the first stage of ablation, we trained a K-head 3D U-Net teacher model with $\mathscr{D}_f^{clean}$ (referred to as '*basic teacher*') followed by K-head 3D U-Net student model with the large dataset ($\mathscr{D}_f^{clean} + \mathscr{D}_p$) and uncertainty-guided loss (referred to as '*basic student*'). In the next

[2]The source code is available at https://github.com/monikagrewal/OrganSegmentation.

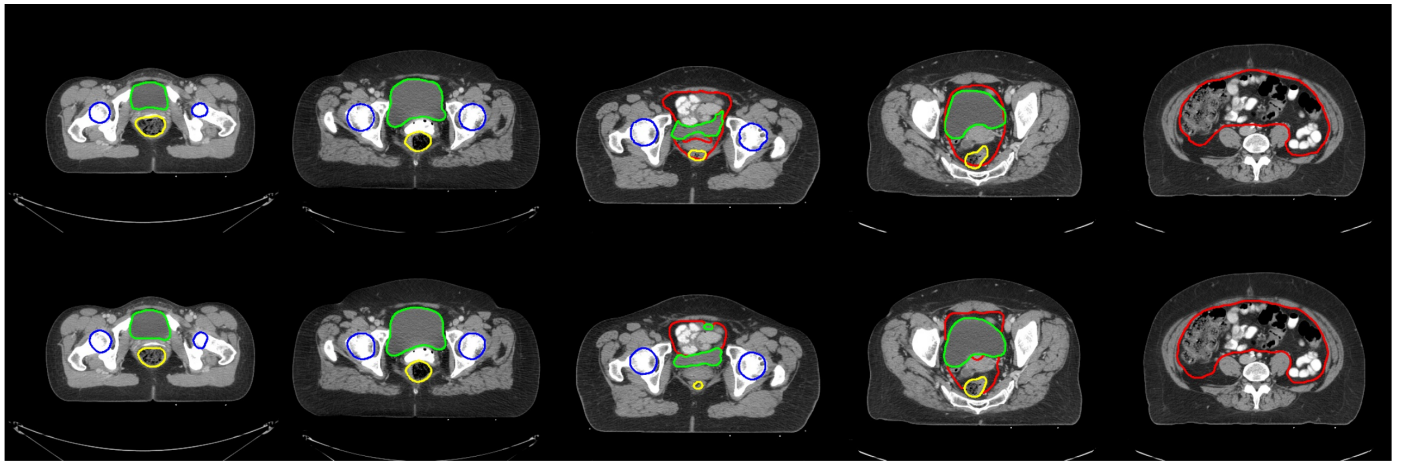
stage, we employed the following additional data augmentations to introduce noise in the data: left-right flipping, masking an organ with a random intensity to simulate contrast, global elastic deformations, and elastic deformations centered in either bowel bag or bladder as additional augmentations. We compared the performance of three models: a teacher model trained with $\mathscr{D}_f^{clean}$ and additional augmentations (referred to as '*robust teacher*'), a student model trained with $\mathscr{D}_f^{clean} + \mathscr{D}_p$ and additional augmentation, and using the imputed annotations from *basic teacher* (referred to as '*basic teacher* + *robust student*'), and a student model trained with $\mathscr{D}_f^{clean} + \mathscr{D}_p$ and additional augmentation, and using the imputed annotations from *robust teacher* (referred to as '*robust teacher* + *robust student*'). Further, we performed 3 iterations of teacher-student training for *robust teacher* + *robust student*, wherein in each subsequent iteration, the student model became the teacher and a new student model was trained.

The mean and standard deviations of the performance metrics on test data from the best models obtained after 5-fold cross-validation are reported in Table 4.1. The distributions of performance metrics for each method (N = 105 test scans × 5 models) were tested for normality using the Kolmogorov-Smirnov test. This was followed by a Friedman test for the main effect and Wilcoxon signed-rank test for post-hoc comparisons. A p-value less than 0.05 with adjustment for multiple comparisons was considered significant.

The automatic data cleaning had a significant impact on the test performance ($p = 5.96e^{-}18$, $p = 6.76e^{-}17$, $p = 2.18e^{-}29$ for Dice, Surface Dice (SD), and Hausdorff distance (HD), respectively), which was mainly due to better bowel bag segmentation. The automatic data cleaning increased the mean Dice coefficient of the bowel bag from 0.7947 to 0.8477 (performance metrics for all the OARs separately are provided in Appendix 4.7.1). Furthermore, learning from a large dataset with the proposed teacher-student setup, annotation imputation, and uncertainty-guided training (*basic student*) provided a significant gain of 2.34% in mean Dice coefficient ($p = 4.51e^{-}38$), 3.22% in mean SD ($p = 1.21e^{-}35$), and 14.47% in mean HD ($p = 1.51e^{-}15$) as compared to learning from a small, fully annotated dataset (U-Net + D_f^{clean}). Adding noise to the data through additional augmentations provided only a marginal gain in the mean performance of the student model, but a considerable decrease in the standard deviations of HD indicating increased robustness towards variations in the test data. Further, iterating the teacher-student training yielded some performance gains, but only till the second iteration. A few representative examples from the results obtained by *basic teacher* + *robust student* are shown in Figure 4.2.

4.4.1. COMPARISON WITH THE STATE-OF-THE-ART (SOTA)

In comparison to SOTA approaches for CT image segmentation for OARs in cervical cancer radiation treatment (shown in Table 4.2), the performance of our approach seems better for the bowel bag, similar for the bladder and hips, but slightly worse for the rectum. Note that the results in [16, 9, 8, 13] correspond to a small test dataset resulting from a single random split, which is susceptible to bias introduced during the splitting of the data. In terms of test dataset size, a comparison with [12] is more



	manual	automatic	manual	automatic	manual	automatic	manual	automatic	manual	automatic
bowel bag	-	-	-	-	3	2	3	3	2	1
bladder	1	1	1	1	2	2	1	1	-	-
hips	2	2	2	2	1	1	-	-	-	-
rectum	2	1	1	1	1	3	1	1	-	-

Figure 4.2: Representative examples of OARs contours. *Top row*: clinically available contours (manual), *Bottom row*: contours generated from OARs segmentation masks predicted by our approach (automatic). Further, the clinical acceptability grades (smaller value indicates better quality) are reported for each OAR.

	A	B	C	D	E1	E2	Ours
Bowel bag	-	-	0.85	-	0.78	0.78	0.86
Bladder	0.91	0.92	0.91	0.89	0.90	0.91	0.92
Hips	0.88	0.905	0.90	0.935	0.89	0.92	0.93
Rectum	0.81	0.79	0.82	0.81	0.77	0.77	0.78
Number of test samples	25	14	27	140	30	30	105

Table 4.2: Mean Dice coefficients reported in A:[16], B:[9], C:[8], D:[12], E1:[13] model 1, E2:[13] model 2, and Ours: *robust teacher + robust student*.

suitable. However, [12] had a comparatively larger training dataset also and trained separate models for each OAR. We believe that using our approach in combination with the data from [12] may result in a better performance with a single model.

4.5. CLINICAL ACCEPTABILITY TEST

We conducted a validation study to assess the clinical acceptability of the automatically generated OARs segmentations. We used the *basic teacher + robust student* model from the first data-split, to predict OARs segmentation masks in the first 4 scans in the test dataset, which were used to generate automatic contours. We showed[3] both the clinically available contours and the automatically generated contours to a radiation oncologist (henceforth referred to as 'clinical expert'), without informing them about the method used to generate the contours. The clinical expert graded each contour for its clinical acceptability according to a 4-point Likert scale: 1=acceptable as it is, 2=acceptable but marginally deviating from exact anatomical definition (subjective to an observer), 3=acceptable with minor corrections because either a part of the organ was not delineated or a peripheral tissue was included in the contour, 4=not acceptable because a correction involving both deletion, as well as delineation of an additional contour, was required.

The clinical acceptability grades for the automatically and manually generated contours for all the graded 2D transverse slices and OARs are shown in Figure 4.3. None of the contours were given grade 4 implying that all the contours were of clinically acceptable quality either as it is or with adaptations. Further, not all of the clinically available contours were graded as 1, representing inter-observer variation. A Chi-squared test of goodness of fit indicated that the histograms of clinical acceptability grades of the automatically generated contours were significantly different from the manually generated contours for the bowel bag ($\chi^2(1, N = 58) = 11.402, p = 0.003$). However, as shown in the Figure 4.3, it was unclear which contours (automatically or manually generated) were better. The clinical acceptability grades for automatically and manually generated contours were not significantly different for the bladder ($\chi^2(1, N = 27) = 2.667, p = 0.102$), and hips ($\chi^2(1, N = 18) = 2.250, p = 0.134$). For the rectum, the Chi-squared test statistics could not be obtained because the frequency counts corresponding to grade 3 were less than 5, however, it is apparent from the Figure

[3]The contours were presented on 2D transverse slices spaced at a 10mm distance to make it similar to the clinical scenario where the contours are delineated on 2D transverse slices. The clinical expert optionally inspected the contours and scans in coronal and sagittal view also to ensure comprehensiveness.

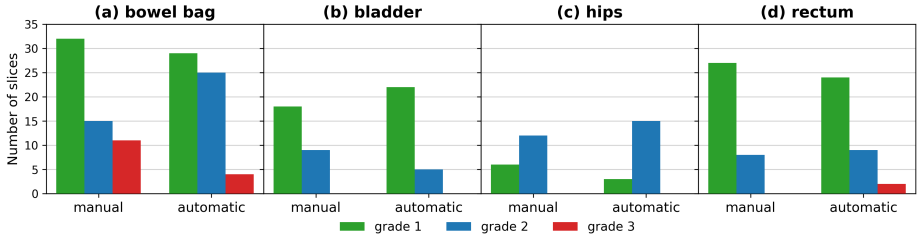


Figure 4.3: Comparison of clinical acceptability grades (smaller value indicates better quality) for clinically available contours (manual) and the contours generated from OARs segmentation masks predicted by our approach (automatic) for (a): bowel bag, (b): bladder, (c): hips, and (d) rectum.

4.3 that the frequency counts in each category were similar for both the automatically and manually generated contours.

Qualitatively, the differences in grade 1 and grade 2 in all the organs were mainly attributed to inter-observer variance. In the case of hips, the window width and window level settings used to visualize the CT scans also influenced the difference between grade 1 and grade 2. Grade 3 corresponded to contours including mesorectum as a part of the bowel bag, and difference in cranial-caudal extent in the rectum.

4.6. DISCUSSION AND CONCLUSIONS

We investigated the possibility of using a large clinically available dataset of the abdominal region to learn a deep learning model for the automatic segmentation of OARs in cervical cancer radiation treatment. To the best of our knowledge, this is one of the few works in the direction of utilizing a large clinically available dataset containing missing annotations for learning a deep learning model. Our experimental results show that learning from a large dataset using our proposed approach yields significant performance gain despite missing annotations in the data. The obtained segmentations from our deep learning model were of clinically acceptable quality, which is encouraging.

Limitations of our work include an ablation study involving only a single run (i.e., network initialization), and a lack of experiments with different semantic segmentation architectures. Both decisions were consciously taken to find sensible results despite the expensive nature of training deep neural networks. Interesting future directions are 1) extending the current work to automatic segmentation of more OARs in cervical cancer radiation treatment e.g., sigmoid and anal canal, and 2) evaluating and learning from datasets of multiple hospitals and demographics to investigate and reduce possible bias in the predictions.

In conclusion, we demonstrated that training a deep learning model without using curated and specifically annotated medical imaging data, but with the capability of predicting clinically acceptable segmentation is possible. Apart from saving clinicians' time, our proposed approach leads to faster development time because of using the readily available data and increased test performance because of the increased dataset.

4.7. APPENDIX

4.7.1. PERFORMANCE METRICS FOR ALL OARS

Table 4.3: Mean (standard deviation) of Dice coefficient of the best models obtained from 5-fold cross-validation. Aug.: additional augmentations.

Method	Bowel bag	Bladder	Hips	Rectum
3D U-Net $\mathscr{D}_f$	79.47 (11.24)	89.25 (16.05)	91.57 (3.73)	73.58 (13.93)
3D U-Net $\mathscr{D}_f^{clean}$	84.77 (6.71)	90.24 (15.62)	91.91 (2.23)	73.15 (15.32)
basic teacher	84.88 (6.21)	90.61 (14.59)	91.81 (2.50)	74.15 (14.05)
basic student	86.31 (5.59)	92.13 (11.43)	92.65 (2.14)	76.95 (12.63)
robust teacher	84.69 (7.01)	90.23 (15.43)	91.73 (2.40)	74.58 (12.70)
basic teacher + robust student	85.86 (5.57)	92.08 (10.03)	92.62 (2.19)	77.86 (11.96)
robust teacher + robust student	86.25 (5.54)	91.93 (10.58)	92.34 (2.22)	78.10 (10.99)
robust teacher + robust student - iter. 2	86.12 (5.50)	92.39 (8.88)	92.69 (2.16)	78.39 (11.92)
robust teacher + robust student - iter. 3	86.40 (5.54)	92.31 (7.60)	92.76 (2.23)	77.92 (12.68)

Table 4.4: Mean (standard deviation) of Surface Dice computed at a tolerance of 2.5mm (voxel spacing) of the best models obtained from 5-fold cross-validation. Aug.: additional augmentations.

Method	Bowel bag	Bladder	Hips	Rectum
3D U-Net $\mathscr{D}_f$	61.55 (10.77)	88.24 (16.99)	96.62 (4.71)	74.51 (14.99)
3D U-Net $\mathscr{D}_f^{clean}$	66.37 (9.27)	90.07 (16.36)	97.03 (3.27)	74.52 (15.51)
basic teacher	66.45 (8.68)	90.55 (15.61)	96.86 (3.84)	75.46 (14.10)
basic student	68.91 (8.57)	93.13 (11.97)	97.55 (3.18)	78.96 (12.92)
robust teacher	66.09 (8.84)	90.46 (16.25)	96.80 (3.47)	75.86 (12.59)
basic teacher + robust student	68.33 (8.61)	92.95 (10.43)	97.53 (3.23)	80.24 (12.10)
robust teacher + robust student	68.97 (8.40)	92.74 (10.78)	97.29 (3.34)	80.30 (11.37)
robust teacher + robust student - iter. 2	69.10 (8.26)	93.57 (9.44)	97.50 (3.21)	81.01 (11.95)
robust teacher + robust student - iter. 3	69.58 (8.32)	93.26 (8.55)	97.52 (3.30)	80.59 (12.61)

4.7.2. DESCRIPTION OF THRESHOLDS FOR AUTOMATIC DATA CLEANING

The histograms of the cranial border of the scans, and the cranial border of the bowel bag annotation with respect to the most cranial point of the hip annotations in D_f are shown in Figure 4.4. The thresholds to crop the scans and delete the bowel bag annotations in the cranial direction are marked in the Figure 4.4.

Table 4.5: Mean (standard deviation) of Hausdorff distance at 95 percentile of the best models obtained from 5-fold cross-validation. Aug.: additional augmentations.

Method	Bowel bag	Bladder	Hips	Rectum
3D U-Net $\mathscr{D}_f$	35.27 (24.77)	7.84 (10.51)	4.16 (20.14)	16.98 (11.86)
3D U-Net $\mathscr{D}_f^{clean}$	19.34 (11.70)	9.68 (37.38)	2.93 (1.00)	17.80 (13.26)
basic teacher	18.43 (9.97)	7.96 (26.75)	2.95 (1.03)	17.10 (12.61)
basic student	17.26 (10.47)	6.31 (22.70)	2.87 (1.04)	16.11 (18.35)
robust teacher	18.43 (10.83)	7.56 (22.30)	2.95 (1.05)	17.34 (17.86)
basic teacher + robust student	17.55 (10.23)	5.57 (15.12)	2.87 (1.05)	15.58 (18.00)
robust teacher + robust student	17.10 (11.23)	5.12 (7.22)	2.90 (1.08)	14.57 (10.63)
robust teacher + robust student - iter. 2	17.23 (10.74)	4.71 (6.55)	2.88 (1.08)	14.58 (11.29)
robust teacher + robust student - iter. 3	17.41 (11.76)	4.49 (5.22)	2.88 (1.14)	15.05 (11.94)

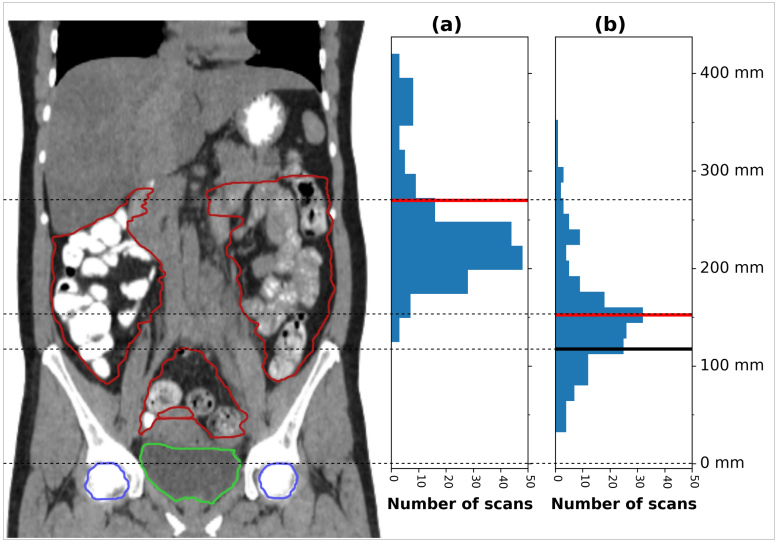


Figure 4.4: The histograms of the number of scans with respect to the distance from the most cranial point of the hip annotations to (a) the cranial border of the scan, and (b) the cranial border of the bowel bag annotation. On the left, a representative CT scan and reconstructed contours in the coronal view are shown (red: bowel bag, green: bladder, blue: hips). Red lines in (a) and (b): thresholds to crop the FOV and delete the bowel bag annotations in the cranial direction. The black line in (b): threshold for discarding the scans, where the bowel bag annotations did not cover the pelvic region. Dashed lines: corresponding anatomy for each threshold.

BIBLIOGRAPHY

[1] Wenjia Bai et al. “Semi-supervised learning for network-based cardiac MR image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*. Springer. 2017, pp. 253–260.

[2] Yarin Gal and Zoubin Ghahramani. “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning”. In: *International Conference on Machine Learning*. 2016, pp. 1050–1059.

[3] Kaiming He et al. “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1026–1034.

[4] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations*. 2015. URL: http://arxiv.org/abs/1412.6980.

[5] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[6] Jiahui Li et al. “Signet Ring Cell Detection with a Semi-supervised Learning Framework”. In: *Information Processing in Medical Imaging*. Ed. by Albert C. S. Chung et al. Cham: Springer International Publishing, 2019, pp. 842–854. ISBN: 978-3-030-20351-1.

[7] Yuexiang Li et al. “Self-Loop Uncertainty: A Novel Pseudo-Label for Semi-supervised Medical Image Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel et al. Cham: Springer International Publishing, 2020, pp. 614–623. ISBN: 978-3-030-59710-8.

[8] Zhikai Liu et al. “Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy”. In: *Radiotherapy and Oncology* 153 (2020). Physics Special Issue: ESTRO Physics Research Workshops on Science in Development, pp. 172–179. ISSN: 0167-8140. DOI: https://doi.org/10.1016/j.radonc.2020.09.060. URL: https://www.sciencedirect.com/science/article/pii/S0167814020308355.

[9] Zhikai Liu et al. “Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network”. In: *Physica Medica* 69 (2020), pp. 184–191. ISSN: 1120-1797. DOI: https://doi.org/10.1016/j.ejmp.2019.12.008. URL: https://www.sciencedirect.com/science/article/pii/S1120179719305290.

[10] Reza Mohammadi et al. “Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer”. In: *Radiotherapy and Oncology* 159 (2021), pp. 231–240. ISSN: 0167-8140. DOI: https://doi.org/10.1016/

j.radonc.2021.03.030. URL: https://www.sciencedirect.com/science/article/pii/S0167814021061703.

[11] Adam Paszke et al. "Automatic differentiation in PyTorch". In: *Advances in Neural Information Processing Systems-W*. 2017. URL: https://github.com/pytorch/pytorch.

[12] Dong Joo Rhee et al. "Automatic contouring system for cervical cancer using convolutional neural networks". In: *Medical Physics* 47.11 (2020), pp. 5648–5658. DOI: https://doi.org/10.1002/mp.14467. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14467. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.14467.

[13] Bastien Rigaud et al. "Automatic Segmentation Using Deep Learning to Enable Online Dose Optimization During Adaptive Radiation Therapy of Cervical Cancer". In: *International Journal of Radiation Oncology*Biology*Physics* 109.4 (2021), pp. 1096–1110. ISSN: 0360-3016. DOI: https://doi.org/10.1016/j.ijrobp.2020.10.038. URL: https://www.sciencedirect.com/science/article/pii/S0360301620344849.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 234–241. ISBN: 978-3-319-24574-4.

[15] Suman Sedai et al. "Uncertainty Guided Semi-supervised Segmentation of Retinal Layers in OCT Images". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Cham: Springer International Publishing, 2019, pp. 282–290. ISBN: 978-3-030-32239-7.

[16] Zhi Wang et al. "Evaluation of deep learning-based auto-segmentation algorithms for delineating clinical target volume and organs at risk involving data for 125 cervical cancer patients". In: *Journal of Applied Clinical Medical Physics* 21.12 (2020), pp. 272–279. DOI: https://doi.org/10.1002/acm2.13097. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/acm2.13097. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.13097.

[17] Qizhe Xie et al. "Self-training with noisy student improves ImageNet classification". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10687–10698.

[18] Lequan Yu et al. "Uncertainty-Aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Cham: Springer International Publishing, 2019, pp. 605–613. ISBN: 978-3-030-32245-8.

[19] Hao Zheng et al. "Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 802–812.

5

MULTI-OBJECTIVE LEARNING

Real-world problems are often multi-objective, with decision-makers unable to specify a priori which trade-off between the conflicting objectives is preferable. Intuitively, building machine learning solutions in such cases would entail providing multiple predictions that span and uniformly cover the Pareto front of all optimal trade-off solutions. We propose a novel approach for multi-objective training of neural networks to approximate the Pareto front during inference. In our approach, we train the neural networks multi-objectively using a dynamic loss function, wherein each network's losses (corresponding to multiple objectives) are weighted by their hypervolume maximizing gradients. Experiments on different multi-objective problems show that our approach returns well-spread outputs across different trade-offs on the approximated Pareto front without requiring the trade-off vectors to be specified a priori. Further, results of comparisons with the state-of-the-art approaches highlight the added value of our proposed approach, especially in cases where the Pareto front is asymmetric.

The content of this chapter is based on the following publication: Deist, T. M.*, **Grewal, M.***, Dankers, F. J., Alderliesten, T., & Bosman, P. A. N. (2023, March). *Multi-Objective Learning using HV Maximization.* In International Conference on Evolutionary Multi-Criterion Optimization (pp. 103-117). Cham: Springer Nature Switzerland. **Authors contributed equally.*

5.1. INTRODUCTION

Multi-objective (MO) optimization refers to finding Pareto optimal solutions for multiple, often conflicting, objectives. In MO optimization, a solution is Pareto optimal if none of the objectives can be improved without a simultaneous detriment in performance on at least one of the other objectives [35]. MO optimization is used for MO decision-making in many real-world applications [32] e.g., e-commerce recommendation [21], treatment plan optimization [25, 27], and aerospace engineering [29]. In this chapter, we focus on learning-based MO decision-making i.e., MO training of machine learning (ML) models so that MO decision-making is possible during inference. Specifically, we focus on training neural networks to generate a finite number of Pareto optimal solutions for each sample[1], so that they together provide a discrete approximation of the Pareto front[2].

The most straightforward approach for MO optimization is linear scalarization, i.e., optimizing a linear combination of different objectives according to scalarization weights. The scalarization weights are based on the desired trade-off between multiple objectives which is often referred to as 'user preference'. A major issue with linear scalarization is that user preferences cannot always be straightforwardly translated to linear scalarization weights. Recently proposed approaches have tackled this issue and find solutions on the average Pareto front for conflicting objectives according to a pre-specified user preference vector [20, 23]. However, in many real-world problems, the user preference vector cannot be known a priori and decision-making is only possible *a posteriori*, i.e., after multiple solutions are generated that are (near) Pareto optimal for a specific sample [3]. For example, in neural style transfer [11] where photos are manipulated to imitate an art style from a selected painting, the user preference between the amount of semantic information (the photo's content) and artistic style can only be decided by looking at multiple different resultant images on the Pareto front (Figure 5.5). Moreover, defining multiple trade-offs, typically by defining multiple scalarizations, to evenly cover the Pareto front is far from trivial, e.g., if the Pareto front is asymmetric. Here, we define asymmetry in Pareto fronts as asymmetry in the distribution of Pareto optimal solutions in the objective space on either side of the 45°-line, the line which represents the trade-off of equal marginal benefit along all objectives (see Pareto fronts in Figure 5.1). We demonstrate and discuss this further in Section 5.4. To enable a posteriori decision-making per sample, multiple solutions spanning the Pareto front need to be generated without requiring the user preference vectors beforehand.

Despite many developments in the direction of MO training of neural networks with pre-specified user preferences, research on MO learning allowing for a posteriori decision-making is still scarce. Here, we present a novel method to multi-objectively train a set of neural networks to this end, leveraging the concept of hypervolume. Although we present our approach for training neural networks, the proposed

[1] Note that, during inference, only *near* Pareto optimal solutions can be generated due to the generalization gap between training and inference.
[2] The Pareto front is the set of all Pareto optimal solutions in objective space.
[3] For more information on a posteriori decision-making, please refer to [14].

formulation can be used for a wide range of ML models.

The hypervolume (HV) – the objective space dominated by a given set of solutions – is a popular metric to compare the quality of different sets of solutions approximating the Pareto front. It has its origins in the field of evolutionary algorithms [39], which are commonly accepted to be state of the art for multi-objective optimization. Theoretically, if the HV is maximal for a set of solutions, these solutions are on the Pareto front [9]. Additionally, HV not only encodes the proximity of a set of solutions to the Pareto front but also their diversity, which means that HV maximization provides a straightforward way for finding diverse solutions on the Pareto front. Therefore, we use hypervolume maximization for MO training of neural networks. We train the set of neural networks with a dynamically weighted combination of loss functions corresponding to multiple objectives, wherein the weight of each loss is based on the HV-maximizing gradients. In summary, this chapter has the following main contributions:

- An MO approach for training neural networks
  - using gradient-based HV maximization
  - predicting Pareto optimal and diverse solutions on the Pareto front per sample without requiring specification of user preferences
  - enabling learning-based a posteriori decision-making.
- Experiments to demonstrate the added value of the proposed approach, specifically in asymmetric Pareto fronts.

5.2. RELATED WORK

MO optimization has been used in machine learning for hyperparameter tuning of machine learning models [18, 2], multi-objective classification of imbalanced data [33], and discovering the complete Pareto set starting from a single Pareto optimal solution [22]. [15] used MO optimization for finding configurations of deep neural networks for conflicting objectives. [13] proposed optimizing the weights of an autoencoder multi-objectively for finding the Pareto front of sparsity and reconstruction error. [24] used the Tchebycheff procedure for multi-objective optimization of a single neural network with multiple heads for multi-task text classification. Although we do not focus on these directions, our proposed approach can be used in similar applications.

MO training of a set of neural networks such that their predictions approximate the Pareto front of multiple objectives is closely related to the work presented in this chapter. Similar to our work, [20, 23] describe approaches with dynamic loss formulations to train multiple networks such that the predictions from these multiple networks together approximate the Pareto front. However, in these approaches, the trade-offs between conflicting objectives are required to be known in advance whereas our proposed approach does not require knowing the set of trade-offs beforehand. Other approaches [28, 19] involve training a "hypernetwork" to predict the weights of another neural network based on a user-specified trade-off. Recently, it has been proposed to condition a neural network for an input user preference vector to allow for predicting multiple points near the Pareto front during inference [31]. While these

approaches can approximate the Pareto front by iteratively predicting neural network weights or outputs based on multiple user preference vectors, the process of sampling the user preference vectors may still be intensive for an unknown Pareto front shape. Another approach proposes growing dense Pareto fronts from sparse Pareto optimal solutions [22], for which our approach can provide baseline solutions.

Gradient-based HV maximization is a key component of our approach. [26] have described gradient-based HV maximization for single networks and formulated a dynamic loss function treating each sample's error as a separate loss. [1] applied this concept for training in generative adversarial networks. HV maximization is also applied in reinforcement learning [34, 38]. While these approaches use HV maximizing gradients to optimize the weights of a single neural network, our proposed approach formulates a dynamic loss based on HV maximizing gradients for a *set* of neural networks. Different from our approach, other concurrent approaches for HV maximization are based on transformation to $(m-1)D$ (where m is the number of objectives) integrals by use of polar coordinates [7], random scalarization [12], and a q-Expected hypervolume improvement function [3].

5

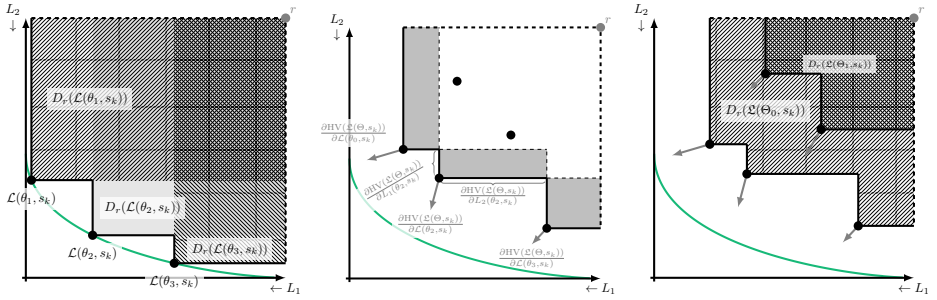
5.3. APPROACH

MO learning of a network parameterized by a vector θ can be formulated as minimizing a vector of n losses $\mathscr{L}(\theta, s_k) = [L_1(\theta, s_k), \ldots, L_n(\theta, s_k)]$ for a given set of samples $S = \{s_1, \ldots, s_k, \ldots, s_{|S|}\}$. These loss functions form the loss space, wherein the subspace attainable by a sample's losses is bounded by its Pareto front. To learn multiple networks with loss vectors on each sample's Pareto front, we replace θ by a set of parameters $\Theta = \{\theta_1, \ldots, \theta_p\}$, where each parameter vector θ_i represents a network. The corresponding set of loss vectors is $\{\mathscr{L}(\theta_1, s_k), \ldots, \mathscr{L}(\theta_p, s_k)\}$ and is represented by a stacked loss vector $\mathfrak{L}(\Theta, s_k) = \big[\mathscr{L}(\theta_1, s_k), \ldots, \mathscr{L}(\theta_p, s_k)\big]$. **Our goal is to learn a set of p networks such that loss vectors in $\mathfrak{L}(\Theta, s_k)$ corresponding to the networks' predictions for sample s_k lie on and span the Pareto front of loss functions for sample s_k.** In other words, each network's loss vector is Pareto optimal and lies in a distinct subsection of the Pareto front for each sample. To achieve this goal, we train networks so that the loss subspace Pareto dominated by the networks' predictions (i.e., the HV) is maximal.

The HV of a loss vector $\mathscr{L}(\theta_i, s_k)$ for a sample s_k is the volume of the subspace $D_r(\mathscr{L}(\theta_i, s_k))$ in loss space dominated by $\mathscr{L}(\theta_i, s_k)$. This is illustrated in Figure 5.1a. To keep this volume finite, the HV is computed with respect to a reference point r which bounds the space to the region of interest[4]. Subsequently, the HV of multiple loss vectors $\mathfrak{L}(\Theta, s_k)$ is the HV of the union of dominated subspaces $D_r(\mathscr{L}(\theta_i, s_k)), \forall i \in \{1, 2, ..., p\}$. The MO learning problem to maximize the mean HV of all $|S|$ samples is as follows:

$$\text{maximize} \frac{1}{|S|} \sum_{k=1}^{|S|} \text{HV}(\mathfrak{L}(\Theta, s_k)) \tag{5.1}$$

[4] The reference point is generally set to large coordinates in loss space to ensure that it is always dominated by all loss vectors.



(a) Dominated subspaces (b) HV gradients (c) Domination-ranked fronts

Figure 5.1: **(a)** Three Pareto optimal loss vectors $\mathscr{L}(\theta_i, s)$ on the Pareto front (green) with dominated subspaces $D_r(\mathscr{L}(\theta_i, s_k))$ with respect to reference point r. The union of dominated subspaces is the dominated hypervolume (HV) of $\mathfrak{L}(\Theta, s_k)$. **(b)** Gray markings illustrate the computation of the HV gradients $\frac{\partial \text{HV}(\mathfrak{L}(\Theta,s))}{\partial \mathscr{L}(\theta_i,s)}$ (gray arrows) in the three non-dominated solutions. **(c)** The same five solutions grouped into two domination-ranked fronts Θ_0 and Θ_1 with corresponding HV, equal to their dominated subspaces $D_r(\mathscr{L}(\theta_i, s_k))$, and HV gradients.

The update direction of gradient ascent for parameter vector θ_i of network i is:

$$\frac{\partial \frac{1}{|S|} \sum_{k=1}^{|S|} \text{HV}(\mathfrak{L}(\Theta, s_k))}{\partial \theta_i} \tag{5.2}$$

By exploiting the chain rule decomposition of HV gradients as described in [8], the update direction in Equation 5.2 for parameter vector θ_i of network i can be written as follows:

$$\frac{1}{|S|} \sum_{k=1}^{|S|} \frac{\partial \text{HV}(\mathfrak{L}(\Theta, s_k))}{\partial \mathscr{L}(\theta_i, s_k)} \cdot \frac{\partial \mathscr{L}(\theta_i, s_k)}{\partial \theta_i} \quad \forall i \in \{1, \ldots, p\} \tag{5.3}$$

The dot product of $\frac{\partial \text{HV}(\mathfrak{L}(\Theta, s_k))}{\partial \mathscr{L}(\theta_i, s_k)}$ (the HV gradients with respect to loss vector $\mathscr{L}(\theta_i, s_k)$) in loss space, and $\frac{\partial \mathscr{L}(\theta_i, s_k)}{\partial \theta_i}$ (the matrix of loss vector gradients in the network i's parameters θ_i) in parameter space, can be decomposed to

$$\frac{1}{|S|} \sum_{k=1}^{|S|} \sum_{j=1}^{n} \frac{\partial \text{HV}(\mathfrak{L}(\Theta, s_k))}{\partial L_j(\theta_i, s_k)} \frac{\partial L_j(\theta_i, s_k)}{\partial \theta_i} \quad \forall i \in \{1, \ldots, p\} \tag{5.4}$$

where $\frac{\partial \text{HV}(\mathfrak{L}(\Theta, s_k))}{\partial L_j(\theta_i, s_k)}$ is the scalar HV gradient in the single loss function $L_j(\theta_i, s_k)$, and $\frac{\partial L_j(\theta_i, s_k)}{\partial \theta_i}$ are the gradients used in gradient descent for single-objective training of network i for loss $L_j(\theta_i, s_k)$. Based on Equation 5.4, one can observe that mean HV maximization of loss vectors from a set of p networks for $|S|$ samples can be achieved by weighting their gradient descent directions for loss functions $L_j(\theta_i, s_k)$ with their corresponding HV gradients $\frac{\partial \text{HV}(\mathfrak{L}(\Theta, s_k))}{\partial L_j(\theta_i, s_k)}$ for all i, j. In other terms, the MO learning of a set of p networks can be achieved by minimizing[5] the following dynamic loss function

[5] Minimizing the dynamic loss function maximizes the HV because the reference point r is in the positive quadrant ("to the right and above 0").

for each network i:

$$\frac{1}{|S|}\sum_{k=1}^{|S|}\sum_{j=1}^{n}\frac{\partial \mathrm{HV}(\mathfrak{L}(\Theta,s_k))}{\partial L_j(\theta_i,s_k)}L_j(\theta_i,s_k) \quad \forall i \in \{1,\dots,p\} \tag{5.5}$$

The computation of the HV gradients $\frac{\partial \mathrm{HV}(\mathfrak{L}(\Theta,s_k))}{\partial L_j(\theta_i,s_k)}$ is illustrated in Figure 5.1b. These HV gradients are equal to the marginal decrease in the subspace dominated only by $\mathscr{L}(\theta_i,s_k)$ when increasing $L_j(\theta_i,s_k)$.

Note that Equation 5.5 maximizes the HV for each sample's losses instead of first averaging losses on the set of samples as commonly done in learning tasks. Consequently, the neural networks are trained on each sample's Pareto front separately, instead of on the front of averages losses. In [5], we experimentally illustrate that learning an average front may lead to undesired results for non-convex fronts.

5.3.1. HV MAXIMIZATION OF DOMINATION-RANKED FRONTS

5

A relevant caveat of gradient-based HV maximization is that HV gradients $\frac{\partial \mathrm{HV}(\mathfrak{L}(\Theta,s_k))}{\partial L_j(\theta_i,s_k)}$ in strongly dominated solutions are zero (because no movement in any direction will affect the HV, Figure 5.1b) and in weakly dominated solutions are undefined [8]. To resolve this issue, we follow [37]'s approach, which avoids the problem of dominated solutions by sorting all loss vectors into separate fronts Θ_l of mutually non-dominated loss vectors and optimizing each front separately (Figure 5.1c). l is the domination rank, and $q(i)$ is the mapping of network i to domination rank l. By maximizing the HV of each front, trailing fronts with domination rank >0 eventually merge with the non-dominated front Θ_0 and a single front is maximized by determining optimal locations for each loss vector on the Pareto front.

Furthermore, we normalize the HV gradients $\frac{\partial \mathrm{HV}(\mathfrak{L}(\Theta_{q(i)},s_k))}{\partial \mathscr{L}(\theta_i,s_k)}$ as in [6] such that their length in loss space is 1. The dynamic loss function with domination-ranking of fronts and HV gradient normalization is:

$$\frac{1}{|S|}\sum_{k=1}^{|S|}\sum_{j=1}^{n}\frac{1}{w_i}\frac{\partial \mathrm{HV}\left(\mathfrak{L}(\Theta_{q(i)},s_k)\right)}{\partial L_j(\theta_i,s_k)}L_j(\theta_i,s_k) \quad \forall i \in \{1,\dots,p\} \tag{5.6}$$

where $w_i = \left\| \frac{\partial \mathrm{HV}(\mathfrak{L}(\Theta_{q(i)},s_k))}{\partial \mathscr{L}(\theta_i,s_k)} \right\|$.

5.3.2. IMPLEMENTATION

We implemented the HV maximization of losses from multiple networks, as defined in Equation 5.6, in Python[6]. The neural networks were implemented using the PyTorch framework [30]. We used [10]'s HV computation reimplemented by Simon Wessing, available from [36]. The HV gradients $\frac{\partial \mathrm{HV}(\mathfrak{L}(\Theta_{q(i)},s_k))}{\partial L_j(\theta_i,s_k)}$ were computed following the algorithm by [8]. Networks with identical losses were assigned the same HV gradients. For non-dominated networks with one or more identical losses (which can occur in

[6] Code is available at https://github.com/timodeist/multi_objective_learning

training with three or more losses), the left- and right-sided limits of the HV function derivatives are not the same [8], and they were set to zero. Non-dominated sorting was implemented based on [4].

5.3.3. A TOY EXAMPLE

Consider an example of MO regression with two conflicting objectives: given a sample $x_k \in S$, from input variable $X \in [0, 2\pi]$, predict the corresponding output z_k that matches y_k^1 from target variable Y_1 and y_k^2 from target variable Y_2, simultaneously. The relation between X, Y_1, and Y_2 is as follows:

$$Y_1 = \cos(X), \quad Y_2 = \sin(X)$$

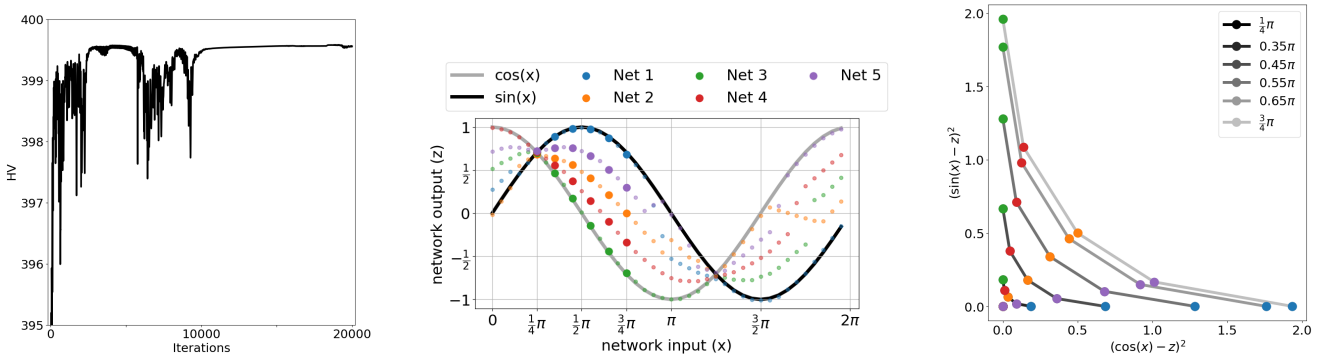
The corresponding mean square error formulations for loss functions are $L_j = \frac{1}{|S|}\sum_{k=1}^{|S|}(y_k^j - z_k)^2; j \in \{1,2\}$. We generated 200 samples of input and target variables for training and validation each. We trained five neural networks for 20000 iterations each with two fully connected linear layers of 100 neurons followed by ReLU nonlinearities. Figure 5.2a shows the HV over training iterations for the set of networks, which stabilizes visibly. Figure 5.2b shows predictions (y-axis) for validation samples evenly sampled from $[0, 2\pi]$ (x-axis). These predictions by five neural networks constitute Pareto front approximations for each sampled x_k, and correspond to precise predictions for $\cos(X)$ and $\sin(X)$, and trade-offs between both target functions. A network may generate predictions with changing trade-offs for different samples, as demonstrated Networks 2-5 in Figure 5.2b for $x \in [\frac{3/2}{\pi}, 2\pi]$. Figure 5.2c shows the predictions for the highlighted samples in Figure 5.2b in loss space, wherein they seem to be evenly distributed on the approximated Pareto front. It becomes clear from Figures 5.2b and 5.2c that each x_k has a differently sized Pareto front which the networks are able to predict. Figure 5.2c also demonstrates an a posteriori decision-making scenario. Upon visualizing the different Pareto fronts per sample, a user might decide to select predictions corresponding to different trade-offs for different samples.

5.4. EXPERIMENTS

We performed experiments with two MO problems: MO regression with differently shaped Pareto fronts and neural style transfer.[7] We compared the performance of our approach with **linear scalarization** and two state-of-the-art approaches: **Pareto MTL** [20] and **EPO** [23]. Pareto MTL and EPO try to find Pareto optimal solutions on the average Pareto front for a given trade-off vector using dynamic loss functions. For a consistent comparison, we used the trade-offs used in the original experiments of EPO for Pareto MTL, EPO, and as fixed weights in linear scalarization.

Experiments were run on systems using Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz with NVIDIA GeForce RTX 2080Ti, or Intel(R) Core(R) i5-3570K @ 3.40Ghz with NVIDIA GeForce GTX 1060 6GB. For training, the Adam optimizer [17] was used. The learning rate and β_1 of Adam were tuned for each approach separately based on the maximal HV of validation loss vectors.

[7] Additional experiments are provided in [5]: multi-observer medical image segmentation, MO regression with three losses, multi-style transfer, and a counter-example for initial loss normalization.



(a) (b) (c)

Figure 5.2: MO regression on two losses. (a) HV values for a set of networks over training iterations. (b) Network outputs for $X \in [0, 2\pi]$. (c) Generated Pareto front predictions for a selection of six samples from $[\frac{1}{4}\pi, \frac{3}{4}\pi]$ in loss space.

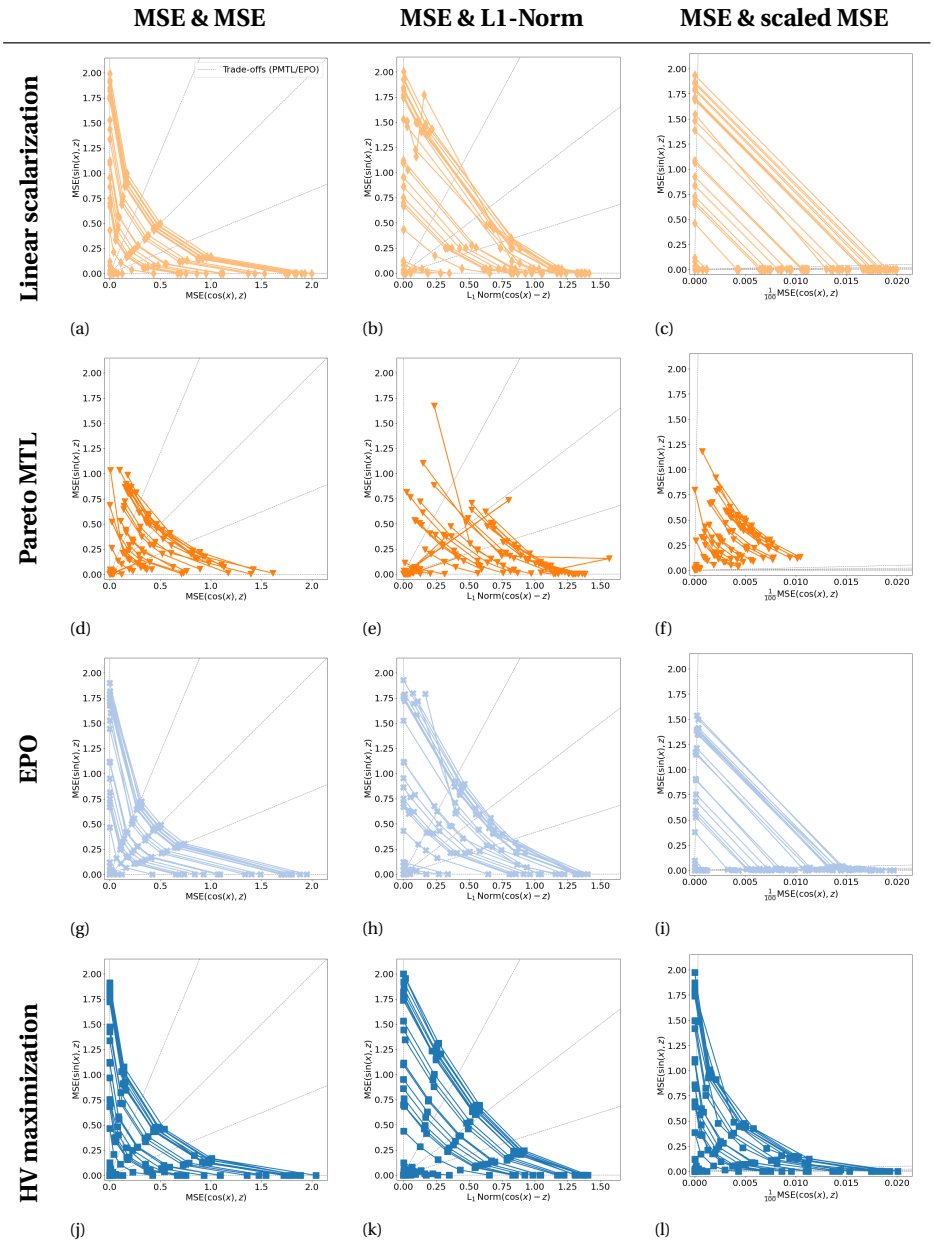


Figure 5.3: Pareto front approximations on a random subset of validation samples by sets of five neural networks trained using four approaches. Three different pairs of loss functions are used: (a)-(d) MSE and MSE, (e)-(h) MSE and L1-Norm, and (i)-(l) MSE and scaled MSE.

5.4.1. MO REGRESSION

We considered three cases for the MO regression toy problem described in Section 5.3.3 each demonstrating a different Pareto front shape: the symmetric case with two MSE losses as in Figure 5.2, and two asymmetric cases each with MSE as one loss and L1-norm or MSE scaled by $\frac{1}{100}$ as the second loss. The reference point for our proposed approach was set to $(20, 20)$.

Figure 5.3 shows Pareto front approximations for all three cases. Figures 5.3a & 5.3g show that fixed linear scalarizations and EPO produce networks generating well-distributed outputs with low losses that predict a sample's symmetric Pareto front for two conflicting MSE losses. The positions on the front approximated by linear scalarization seem to be far from the pre-specified trade-offs (gray lines). This is expected because, by definition of linear scalarization, the solutions should lie on the approximated Pareto front where the tangent is perpendicular to the search direction specified by the trade-offs. For Pareto MTL, networks are clustered closer to the center of the approximated Pareto front. Optimizing MSE and L1-Norm (Figures 5.3b-5.3k) results in an asymmetric Pareto front approximation. The predictions by our HV maximization-based approach remain well distributed across the fronts. EPO also still provides a decent spread albeit less uniform across samples whereas linear scalarization and Pareto MTL tend to both or mostly the lower extrema, respectively.

The difficulty of manually pre-specifying the trade-offs without knowledge of the Pareto front becomes more evident when optimizing losses with highly different scales (Figures 5.3c-5.3l). The pre-specified trade-offs do not evenly cover the Pareto fronts. Consequently, the networks trained by EPO do not cover the Pareto front evenly despite following the pre-specified trade-offs. Further, the networks optimized by Pareto MTL cover only the upper part of the fronts. Networks trained with fixed linear scalarizations tend towards both extrema. On the other hand, our approach trains networks that follow well-distributed trade-offs on the Pareto front. Normalizing losses from differing scales as in Figures 5.3c-5.3l might not sufficiently improve methods based on pre-specified trade-offs (Pareto MTL, EPO) or fixed linear scalarizations [5].

The mean HV over 200 validation samples is computed for all approaches and Table 5.1 displays the median and inter-quartile ranges (IQR) over 25 runs. The magnitude of the HV is largely determined by the position of the reference point. For $r = (20, 20)$ the maximal HV equals 400 minus the area bounded by the utopian point $(0, 0)$ and a sample's Pareto front. Even poor approximations of a sample's Pareto front can yield a HV ≥ 390. For these reasons, HVs in Table 5.1 appear large and minuscule differences between HVs are relevant. As expected, our approach finds higher HV values for the case of asymmetric front shapes (Table 5.1 columns 2 and 3, and Figures 5.3b-5.3l). In case of the symmetric front shape (Figure 5.3a), since the pre-specified trade-offs appear to span the Pareto front shape well, linear scalarization's training based on fixed loss weights is more efficient than training on a dynamic loss with varying weights as used by HV maximization. This increased efficiency of training using fixed weights that are suitable for symmetric MSE losses presumably results in a slightly higher HV for linear scalarization (Table 5.1 column 1).

Table 5.1: Comparison of HV across different approaches. The maximal median HV in each column is **highlighted**. Small increases in HV close to the maximum (10^6 or 400) matter: see Section 5.4.1. A statistically significant one-sided Wilcoxon signed rank test with correction for multiple comparison is indicated by: LS vs HV max. (*), PMTL vs HV max. (†), and EPO vs HV max. (‡). **Columns 1-3:** Median (inter-quartile range) values of the mean HV of the approximated Pareto fronts for 200 validation samples from 25 runs of MO regression problem are reported. **Column 4:** Median (inter-quartile range) HV of the approximated Pareto fronts for the 25 image sets used in neural style transfer are reported.

	MSE & MSE	**MSE & L1-Norm**	**MSE & scaled MSE**	**Style & content**
Linear scalarization (LS)	**399.5929*** **(399.5776, 399.6018)**	399.2909 (399.2738, 399.3045)	399.9859 (399.9857, 399.9864)	999990.7699 (999988.6580, 999992.5850)
Pareto MTL (PMTL)	397.1356 (396.3212, 397.6288)	392.2956 (392.0377, 393.4942)	398.3159 (397.4799, 398.6699)	997723.8748 (997583.5152, 998155.6837)
EPO	399.5135 (399.5051, 399.5348)	399.0884 (398.998, 399.1743)	399.9885 (399.9883, 399.9889)	999988.4297 (999984.4808, 999989.8338)
HV maximization	399.5823†‡ (399.5619, 399.6005)	**399.3795***†‡ **(399.3481, 399.4039)**	**399.9954***†‡ **(399.9927, 399.9957)**	**999999.7069 (999999.4543, 999999.8266)***†‡

5.4.2. NEURAL STYLE TRANSFER

We further considered the MO optimization problem of neural style transfer as defined in [11] (we reused and adjusted Pytorch's neural style transfer implementation [16]), where pixels of an image are optimized to minimize content loss (semantic similarity with a target image) and style loss (artistic similarity with a style image) simultaneously. We performed experiments with 25 image pairs (image sources as in [5]), obtained by combining 5 content and 10 style images to generate 6 solutions on the Pareto front. The reference point in our approach was chosen as (100, 10000) based on preliminary runs.

Figure 5.4 shows the obtained Pareto front estimates for 25 image sets by each approach. Linear scalarization (a) and EPO (c) determine solutions close to or on the chosen user preferences which, however, do not diversely cover the range of possible trade-offs. Pareto MTL (b) achieves sets of clustered and partly dominated solutions, which do not cover trade-offs with low content loss. On the other hand, HV maximization (d) returns Pareto front estimates that broadly cover diverse trade-offs between style and content loss across different image sets without having to specify user preferences. This is also reflected in the significantly larger median HVs reported in Table 5.1. Figure 5.5 shows the images generated by each approach for one of the image sets. This case was manually selected for its aesthetic appeal.[8] The images seen

[8] Generated images for all image sets are available at https://github.com/timodeist/multi_objective_learning.

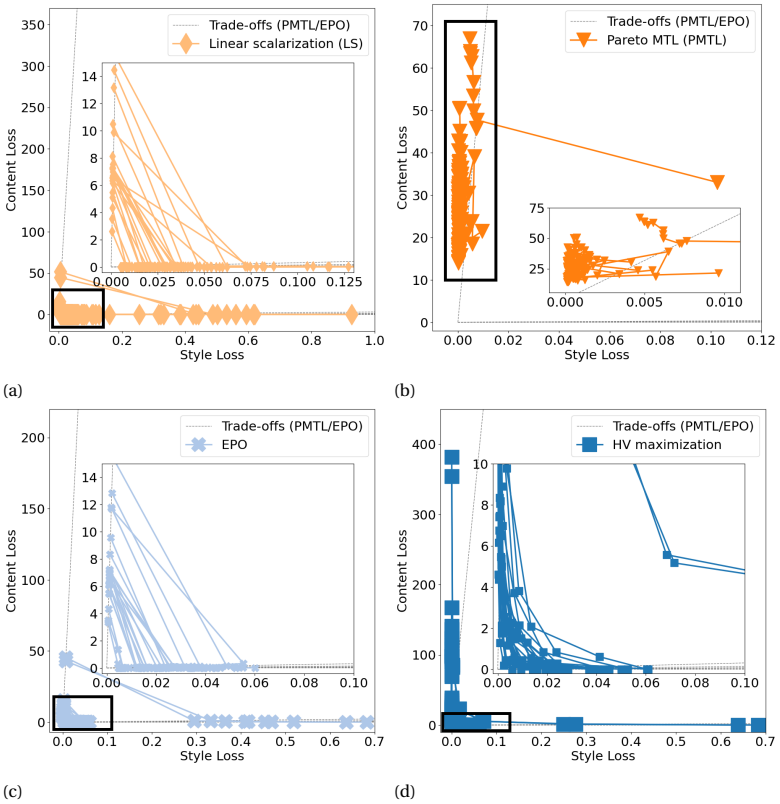


Figure 5.4: Pareto front estimates in loss space by different approaches for neural style transfer using four approaches: (a) Linear scalarization (b) Pareto MTL, (c) EPO, and (d) HV maximization. Sections within the black frames are magnified.

here match observations from Figure 5.4, e.g., Pareto MTL's images show little diversity in style and content, many images by linear scalarization of EPO have too little style match ('uninteresting' images), and images by HV maximization show most interesting diversity.

Figure 5.5: Neural style transfer example by all four approaches for one image set.

5.5. DISCUSSION

We have proposed an approach to train a set of neural networks such that they jointly predict Pareto front approximations for each sample during inference, without requiring user-specified trade-offs. Our approach translates the concept of gradient-based HV maximization from MO optimization to MO learning. We provide experimental comparisons with existing approaches that require a priori specification of the trade-offs. The results highlight the advantage of our HV maximization approach, especially in MO problems that exhibit asymmetric Pareto front.

Our HV maximization based approach does not require specifying p trade-offs a priori (based on the number of predictions, p, required on the Pareto front), which essentially are $p(n-1)$ hyperparameters of the learning process for n losses. Choosing these trade-offs well requires knowledge of the Pareto front shapes, which is often not known a priori. HV maximization, however, introduces the n-dimensional reference point r and thus n additional hyperparameters. However, choosing a reference point such that the entire Pareto front gets approximated is not complex. It often suffices to use losses of randomly initialized networks rescaled by a factor ≥ 1 as the reference point. If only a specific section of the Pareto front is relevant and this is known a priori, the reference point can be chosen so that the Pareto front approximation only spans the chosen section.

HV-based training for sets of neural networks can, in theory, be applied to any number of networks, p, and loss functions, n. In practice, the time complexity of exact HV (exponential in n, [10]) and HV gradient (quadratic in p with $n \leq 4$, [8]) computations is limiting but may be overcome by algorithmic improvements using, e.g., HV approximations. Further, we train a separate network corresponding to each prediction. This increases computational load linearly if more predictions on the Pareto front are desired. We train separate networks instead of one multi-headed network for the sake of simplicity in experimentation and clarity when demonstrating our approach. It is expected that the HV maximization formulation would work similarly if the parameters of some of the neural network layers are shared, which would decrease computational load.

In conclusion, we describe MO training of neural networks using HV maximization for learning-based a posteriori MO decision-making. Our approach provided the desired well-spread Pareto front approximations on artificial MO regression problems. On the MO style transfer problem, our method yielded encouraging results that emphasize its usefulness for a posteriori decision-making.

BIBLIOGRAPHY

[1] Isabela Albuquerque et al. “Multi-objective training of Generative Adversarial Networks with multiple discriminators”. In: *arXiv preprint arXiv:1901.08680* (2019).

[2] Brendan Avent et al. “Automatic Discovery of Privacy–Utility Pareto Fronts”. In: *Proceedings on Privacy Enhancing Technologies* 2020.4 (2020), pp. 5–23.

[3] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. “Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization”. In: *arXiv preprint arXiv:2006.05078* (2020).

[4] Kalyanmoy Deb et al. “A fast and elitist multiobjective genetic algorithm: NSGA-II”. In: *IEEE Transactions on Evolutionary Computation* 6.2 (2002), pp. 182–197.

[5] Timo M Deist et al. “Multi-objective learning to predict Pareto fronts using hypervolume maximization”. In: *arXiv preprint arXiv:2102.04523* (2021).

[6] Timo M. Deist et al. “Multi-objective Optimization by Uncrowded Hypervolume Gradient Ascent”. In: *International Conference on Parallel Problem Solving from Nature*. Springer. 2020, pp. 186–200.

[7] Jingda Deng and Qingfu Zhang. “Approximating hypervolume and hypervolume contributions using polar coordinate”. In: *IEEE Transactions on Evolutionary Computation* 23.5 (2019), pp. 913–918.

[8] Michael Emmerich and André Deutz. “Time complexity and zeros of the hypervolume indicator gradient field”. In: *EVOLVE-A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation III*. Springer, 2014, pp. 169–193.

[9] Mark Fleischer. “The measure of Pareto optima applications to multi-objective metaheuristics”. In: *International Conference on Evolutionary Multi-Criterion Optimization*. Springer. 2003, pp. 519–533.

[10] Carlos M Fonseca, Luıs Paquete, and Manuel López-Ibánez. “An improved dimension-sweep algorithm for the hypervolume indicator”. In: *2006 IEEE International Conference on Evolutionary Computation*. IEEE. 2006, pp. 1157–1163.

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. “Image style transfer using convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2414–2423.

[12] Daniel Golovin et al. “Random Hypervolume Scalarizations for Provable Multi-Objective Black Box Optimization”. In: *arXiv preprint arXiv:2006.04655* (2020).

[13] Maoguo Gong et al. “A multiobjective sparse feature learning model for deep neural networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 26.12 (2015), pp. 3263–3277.

[14] C-L Hwang and Abu Syed Md Masud. *Multiple objective decision making—methods and applications: a state-of-the-art survey*. Vol. 164. Springer Science & Business Media, 2012.

[15] Md Shahriar Iqbal et al. "FlexiBO: Cost-Aware Multi-Objective Optimization of Deep Neural Networks". In: *arXiv preprint arXiv:2001.06588* (2020).

[16] Alexis Jacq. *Neural style transfer using Pytorch*. https://pytorch.org/tutorials/advanced/neural_style_tutorial.html. 2017.

[17] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations*. 2015. URL: http://arxiv.org/abs/1412.6980.

[18] Patrick Koch et al. "Efficient multi-criteria optimization on noisy machine learning problems". In: *Applied Soft Computing* 29 (2015), pp. 357–370.

[19] Xi Lin et al. "Controllable Pareto Multi-Task Learning". In: *arXiv preprint arXiv:2010.06313* (2020).

[20] Xi Lin et al. "Pareto multi-task learning". In: *Advances in Neural Information Processing Systems*. 2019, pp. 12060–12070.

[21] Xiao Lin et al. "A Pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation". In: *Proceedings of the 13th ACM Conference on Recommender Systems*. 2019, pp. 20–28.

[22] Pingchuan Ma, Tao Du, and Wojciech Matusik. "Efficient continuous Pareto exploration in multi-task learning". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6522–6531.

[23] Debabrata Mahapatra and Vaibhav Rajan. "Multi-task learning with user preferences: Gradient descent with controlled ascent in Pareto optimization". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6597–6607.

[24] Yuren Mao et al. "Tchebycheff procedure for multi-task text classification". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4217–4226.

[25] Stefanus C Maree et al. "Evaluation of bi-objective treatment planning for high-dose-rate prostate brachytherapy—A retrospective observer study". In: *Brachytherapy* 18.3 (2019), pp. 396–403.

[26] Conrado S Miranda and Fernando J Von Zuben. "Single-solution hypervolume maximization and its use for improving generalization of neural networks". In: *arXiv preprint arXiv:1602.01164* (2016).

[27] B. Müller et al. "Multicriteria plan optimization in the hands of physicians: a pilot study in prostate cancer and brain tumors". In: *Radiation Oncology* 12 (2017).

[28] Aviv Navon et al. "Learning the Pareto Front with Hypernetworks". In: *arXiv preprint arXiv:2010.04104* (2020).

[29] Akira Oyama and Meng-Sing Liou. "Multiobjective optimization of rocket engine pumps using evolutionary algorithm". In: *Journal of Propulsion and Power* 18.3 (2002), pp. 528–535.

[30] Adam Paszke et al. "Automatic differentiation in PyTorch". In: *Advances in Neural Information Processing Systems-W*. 2017. URL: https://github.com/pytorch/pytorch.

[31] Michael Ruchte and Josif Grabocka. “Efficient Multi-Objective Optimization for Deep Learning”. In: *CoRR* abs/2103.13392 (2021). arXiv: 2103.13392. URL: https://arxiv.org/abs/2103.13392.

[32] Theodor Stewart et al. “Real-world applications of multiobjective optimization”. In: *Multiobjective Optimization* (2008), pp. 285–327.

[33] Sara Tari et al. “Automatic Configuration of a Multi-objective Local Search for Imbalanced Classification”. In: *International Conference on Parallel Problem Solving from Nature*. Springer. 2020, pp. 65–77.

[34] Kristof Van Moffaert and Ann Nowé. “Multi-objective reinforcement learning using sets of Pareto dominating policies”. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3483–3512.

[35] David A Van Veldhuizen and Gary B Lamont. “Multiobjective evolutionary algorithms: Analyzing the state-of-the-art”. In: *Evolutionary Computation* 8.2 (2000), pp. 125–147.

[36] Hao Wang et al. *Code repository: Hypervolume Indicator Gradient Ascent Multi-objective Optimization*. https://github.com/wangronin/HIGA-MO.

[37] Hao Wang et al. “Hypervolume indicator gradient ascent multi-objective optimization”. In: *International Conference on Evolutionary Multi-Criterion Optimization*. Springer. 2017, pp. 654–669.

[38] Jie Xu et al. “Prediction-Guided Multi-Objective Reinforcement Learning for Continuous Robot Control”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 10607–10616.

[39] Eckart Zitzler and Lothar Thiele. “Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach”. In: *IEEE Transactions on Evolutionary Computation* 3.4 (1999), pp. 257–271.

6

MULTI-OBJECTIVE LEARNING FOR DEFORMABLE IMAGE REGISTRATION

Deformable image registration (DIR) involves optimization of multiple conflicting objectives, however, not many existing DIR algorithms are multi-objective (MO). Further, while there has been progress in the design of deep learning algorithms for DIR, there is no work in the direction of MO DIR using deep learning. In this paper, we fill this gap by combining a recently proposed approach for MO training of neural networks with a well-known deep neural network for DIR and create a deep learning based MO DIR approach. We evaluate the proposed approach for DIR of pelvic magnetic resonance imaging (MRI) scans. We experimentally demonstrate that the proposed MO DIR approach – providing multiple DIR outputs for each patient that each correspond to a different trade-off between the objectives – has additional desirable properties from a clinical use point-of-view as compared to providing a single DIR output. The experiments also show that the proposed MO DIR approach provides a better spread of DIR outputs across the entire trade-off front than simply training multiple neural networks with weights for each objective sampled from a grid of possible values.

The content of this chapter is based on the following publication: **Grewal, M.**, Westerveld, H., Bosman, P. A. N., & Alderliesten, T. (2024, February). *Multi-Objective Learning for Deformable Image Registration.* In Medical Imaging with Deep Learning.

6.1. INTRODUCTION

Deformable image registration (DIR) refers to the task of finding a non-linear transformation that aligns two images. The non-linear transformation is characterized by a deformation vector field (DVF), that maps each location in the target image (also referred to as fixed or reference image) to a location in the source image (also referred to as moving image). The source image is then warped by resampling from the mapped locations. Some of the potential applications of DIR in medical imaging are dose accumulation in radiation treatment, contour propagation, tumor growth tracking, and creating a digital atlas [12, 16, 20, 17].

DIR involves optimization of a parameterized DVF to maximize the similarity between two images. However, optimizing only for maximizing image similarity may yield a highly irregular or sometimes physically implausible DVF due to model overfitting. Therefore, an additional objective penalizing irregularity in the DVF is often used, which inherently conflicts with the objective of maximizing image similarity [10, 3, 5]. Further, an additional guidance objective (either maximizing the similarity between organ contours or minimizing the distance between corresponding landmarks) is often utilized in challenging DIR problems [3, 7]. Intuitively, improvement in the additional guidance objective should always lead to improvement in the image similarity objective. However, in practice, the additional guidance objective may still conflict with the image similarity objective. This is often caused when the optimization gets overfitted to the regions where additional guidance is provided, deteriorating performance in other image regions [3]. Another cause for conflict between the image similarity objective with the additional guidance can be the uncertainty in the additional guidance, which, in turn, could be caused by either inter/intra-observer variance in case of manual annotation or modeling error in case of automatic generation of additional guidance. Therefore, DIR is essentially a multi-objective (MO) problem [6], which involves two or more conflicting objectives. This implies that fundamentally an MO approach is appropriate for DIR, where multiple DIR outputs corresponding to a diverse range of trade-offs between the conflicting objectives are provided to the clinicians to a posteriori choose the best solution. Although the notion of DIR being multi-objective is well accepted and discussed, not many DIR approaches have been developed with this perspective. [1] provided a proof-of-concept study for MO DIR of 2D images. [15] used an evolutionary algorithm to tune the corresponding weights of different objectives for each 3D breast MRI pair and perform a single objective DIR multiple times. [13] formulated DIR as MO problem by partitioning the template image into several overlapping regions. [2] presented the first integral approach to MO DIR that could be used for 3D volumetric scans using an MO optimization algorithm.

With the advent of deep learning in the past few years, multiple deep learning based DIR approaches have been proposed [3, 19, 10, 11, 17, 16], which provide the possibility to predict the DVF for an entire volumetric scan within seconds. However, to the best of our knowledge, there is no work done in the direction of MO DIR using deep learning. In this chapter, we fill this gap and provide a novel approach for MO DIR using deep learning. To this end, we employed a well-known deep neural network for DIR, VoxelMorph [3], and combined it with the proposed technique for training neural

networks multi-objectively in Chapter 5. Our main contributions are the following:

- We develop a deep learning based approach for MO DIR so that multiple DIR outputs corresponding to different trade-offs between multiple objectives can be presented to the clinical experts for a posteriori decision-making [8].
- We demonstrate MO DIR for a challenging real-world registration task: DIR of female pelvic magnetic resonance imaging (MRI) scans and highlight its potential benefits.

6.2. APPROACH

We first provide a brief background on the concepts of MO optimization that we apply to deep learning based DIR. MO optimization refers to minimizing[1] a vector of n objectives simultaneously. The goal is to find a set (often referred to as 'approximation set') of p solutions that are both close to as well as diversely-spread along the Pareto front – the set of all Pareto optimal solutions in objective space. A solution is Pareto optimal if none of the objectives can be improved without a simultaneous detriment in performance in at least one of the other objectives [18].

Our deep learning based MO DIR implementation consists of a DIR network within the MO learning framework proposed in the previous chapter (Chapter 5). We selected VoxelMorph [3] for DIR because it is a well-known neural network for DIR. VoxelMorph uses an encoder-decoder style neural network for predicting a DVF, which is a basis for many deep learning based DIR approaches proposed afterwards. We selected the MO learning framework proposed in Chapter 5 for two reasons: a) it achieves MO training of neural networks through hypervolume (HV) maximization - a process that inherently ensures Pareto optimality[2] and diversity between the solutions, b) it is the only MO approach that allows training neural networks multi-objectively without a priori knowledge of the exact preference between different objectives. It should be noted that the latter is crucial in the task of DIR. This is because earlier literature suggests that the exact preference between different objectives may be different between different image pairs, which may only be known a posteriori after inspecting multiple solutions [15].

In this chapter, we aim to minimize p loss vectors (corresponding to p solutions or DIR outputs in the approximation set), each comprising of three losses: $L_{ImageSimilarity}$, $L_{DVFSmoothness}$, and $L_{SegSimilarity}$. Here, for $L_{ImageSimilarity}$, we used normalized cross-correlation loss. $L_{DVFSmoothness}$ is the squared sum of spatial gradients of the predicted DVF in all directions, and $L_{SegSimilarity}$ is the Dice loss between the fixed image's organ mask and the moving image's organ mask warped by the predicted DVF (refer to Balakrishnan et al. [3] for details). In the original formulation of MO learning in Chapter 5, p neural networks are required corresponding to p solutions in the approximation set. Due to the memory intensive nature of training a 3D DIR network, this poses a challenge due to limited GPU memory. To tackle this, we modified the original implementation by sharing the weights of the encoder between p

[1] In this chapter, we assume minimization as objectives correspond to losses in deep learning.
[2] If HV is maximal, all the solutions are Pareto optimal.

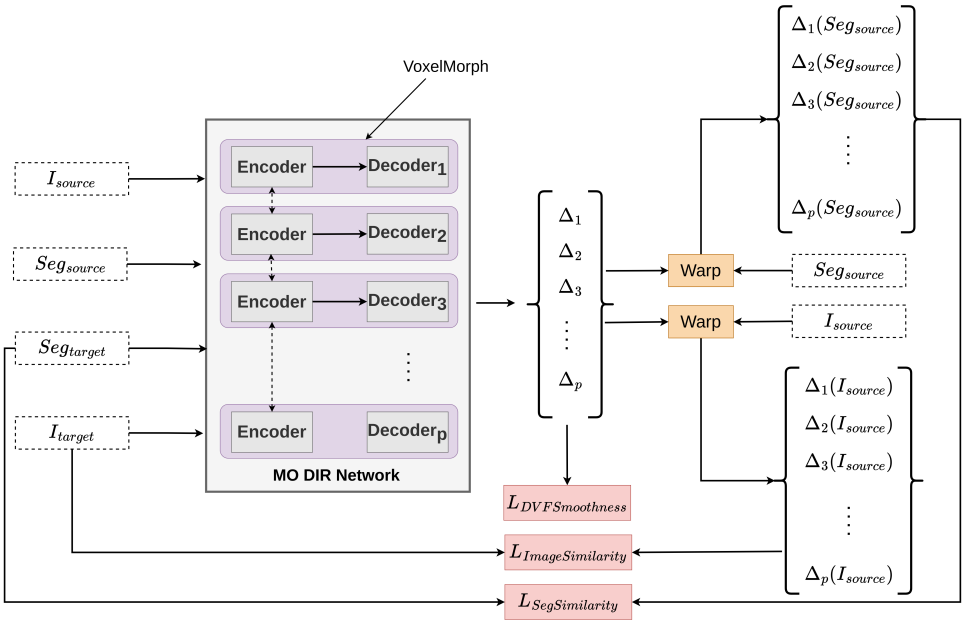


Figure 6.1: Illustration of the proposed deep learning based MO DIR approach. I_{source}: source image, I_{target}: target image, Seg_{source} and Seg_{target}: organ segmentation masks for source and target image, respectively. The weights of the encoder are shared among p DIR networks, which output p DVFs (Δ_1, Δ_2, ..., Δ_p) to warp I_{source} and Seg_{source}. The network is trained to simultaneously minimize p loss vectors $[L_{ImageSimilarity}, L_{DVFSmoothness}, L_{SegSimilarity}]$ using MO learning.

DIR networks as shown in Figure 6.1. We confirmed through preliminary experiments (shown in Appendix 6.6.1) that sharing the weights in the encoder did not affect the performance adversely.

The DIR network predicts p DIR outputs (DVFs). This is followed by calculation of p loss vectors, which are used in the MO learning framework. The parameters of the DIR network are updated using a dynamic loss formulation, that, for each DIR output is defined as:

$$L^i = w_1^i L_{ImageSimilarity} + w_2^i L_{DVFSmoothness} + w_3^i L_{SegSimilarity} \quad \forall i \in \{1, \ldots, p\} \tag{6.1}$$

Here, the weights w_1^i, w_2^i, w_3^i are calculated in each iteration using the HV maximization described in the previous chapter (Chapter 5). This ensures that at the end of the training the DIR outputs (which are used to calculate the p loss vectors) are close to, and diversely distributed along the Pareto front of the three objectives.

MO DIR as described above can be understood as training p DIR networks simultaneously, each with different weights for the loss terms, and the weights being selected automatically such that the HV is maximal. That said, MO DIR is fundamentally different from the traditional single DIR following hyperparameter search for the loss weights. In the traditional set up, the selection of a weight (which translates to a trade-off on the approximation front) for each loss is done a priori based on quantitative comparison of a single aggregated (on a validation set) performance metric. Whereas in MO DIR, the selection is done a posteriori by clinical experts based on qualitative evaluation of multiple criteria specific to each patient.

6.3. DATA

We retrospectively used data from cervical cancer patients who received brachytherapy treatment at Leiden University Medical Center (LUMC), The Netherlands. We received 136 MRI scan pairs (along with associated contours generated for clinical use of four organs at risk: bladder, bowel bag, rectum, and sigmoid) corresponding to two fractions of brachytherapy treatment in anonymized form after approval from the medical ethics committee. The original resolution of the MRI scans was 0.5 mm × 0.5 mm × 4 mm. We resampled the MRI scans to isotropic voxel spacing of 1 mm × 1 mm × 1 mm because the convolution kernels, downsampling, and upsampling operations in VoxelMorph are symmetric. We used randomly cropped patches of size 192 × 192 × 32 as an input to the neural network. We separated the scans at patient level based on their chronological order of acquisition into train and validation (126 scan pairs), and test (10 scan pairs) splits. On the test scans, a radiation therapy technologist annotated 23 anatomical landmarks (details in Appendix 6.6.2), which were selected by a radiation oncologist on the basis of their importance in brachytherapy treatment for cervical cancer patients. The placement of landmarks was cross-checked by another radiation oncologist.

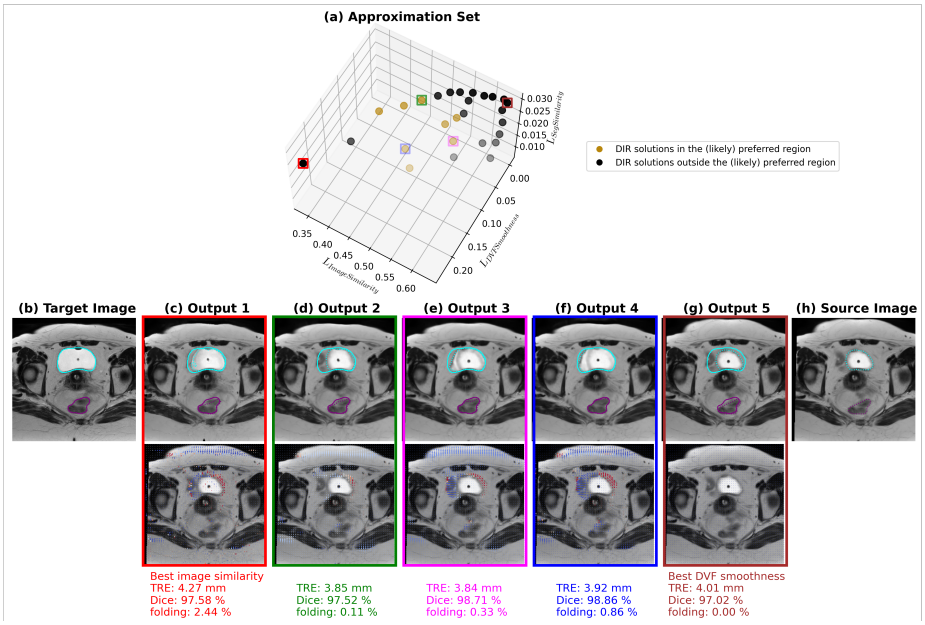


Figure 6.2: (a) Approximation set. (b) and (h): A transverse slice from the target and source image, respectively. (c)-(g) top row: Warped source images corresponding to five solutions (highlighted with matching frame color) in the set with bladder and rectum contours in cyan and magenta colors, respectively. Solid contours represent the contour in the target image and dashed contours represent the warped source image contour. (c)-(g) bottom row: DVFs overlaid on the source image. Displacement in the x-y plane is represented by direction and scale, and in the z-direction by color (red for cranial, and blue for caudal motion) of arrows.

6.4. EXPERIMENTS AND RESULTS

We implemented[3] our proposed approach using Python and PyTorch. The training hyperparameters were: number of solutions p = 27, initialization = Kaiming He, optimizer = Adam, learning rate (lr) = $1e^{-4}$, number of training iterations = 20K, reference point for HV calculation = (1, 1, 1) (details in Appendix 6.6.3). For each experimental setting, we trained 5 models, each corresponding to a different data split. We report their performance on the test set without model selection. To assess the DIR performance, we calculated target registration errors (TREs) of the 23 manually annotated landmarks by transforming the landmarks in the target image with the predicted DVF and calculating the Euclidean distance with the corresponding landmarks in the source image. We also calculated the percentage of voxels with a negative determinant of the spatial Jacobian of the DVF, as an indication of folding in the transformation.

[3]The implementation is available at https://github.com/monikagrewal/DL-MODIR/tree/public.

6.4.1. COMPARISON OF MO DIR WITH SINGLE DIR OUTPUT

Contrary to traditional DIR, in MO DIR, the decision maker (in our case a clinical expert) is provided with multiple DIR solutions spread across a range of trade-offs between conflicting objectives. This is demonstrated in Figure 6.2 (a). The figure shows that there are multiple possible ways to align the two images. In DIR, the solutions at the extremes of the approximation set are likely not interesting because they might be overfitted to a single objective and consequently may yield sub-optimal performance in other objectives. For example, the solution highlighted in the red frame (Output 1) corresponds to minimum $L_{ImageSimilarity}$, but maximum $L_{DVFSmoothness}$ causing a lot of folding in the DVF. Similarly, the solution highlighted in brown (Output 5) corresponds to no deformation at all. To assist the a posteriori decision-making, such uninteresting solutions can be filtered out by setting acceptance thresholds on each objective. The region of interest in the objective (loss) space where all the acceptance criteria are met, could be considered the preferred region. In Figure 6.2 (a), we show this region with arbitrarily selected acceptance thresholds ($L_{ImageSimilarity} < 0.55$, $L_{DVFSmoothness} < 0.1$, and $L_{SegSimilarity} < 0.025$).

Within a preferred region of interest, one solution cannot be selected over another based on quantitative comparison of performance metrics as demonstrated in Figure 6.2. The solution highlighted in green (Output 2) has minimum folding in the DVF, magenta (Output 3) has minimum mean TRE of landmarks, and blue (Output 4) has maximum Dice similarity between organ masks while other metrics are worse. While Output 2 and Output 3 have less folding in the DVF and smaller mean TRE between landmarks, the warped bladder contours (dashed cyan color) considerably deviate from the target bladder contours (solid cyan color) as compared to Output 4. This is due to MO training of the DIR neural network, which ensures that the obtained DIR solutions are all (close to) Pareto optimal i.e., no solution is better than another in any objective without a simultaneous detriment on other objectives. In such a scenario, the most appropriate DIR output can only be selected after visual inspection of the DIR outputs in the preferred region of interest and considering other clinical criteria. For example, the visual inspection of the DVF from Output 4 may reveal that the folding occurs in regions not relevant for brachytherapy treatment. Further, the alignment of the bladder may be more important than the alignment of some landmarks in other regions. Therefore, a clinical expert may prefer Output 4 over Output 3 despite a larger mismatch between landmarks and more folding in the DVF in this test scan pair. Whereas, in another test scan pair, the characteristics of the DVF may be different and the clinical preference may be reversed. Moreover, it is already known from previous research that the weights, which translate to a given trade-off between objectives on the approximation front and the quantitative value of the performance metrics are different in different scan pairs [15]. This means that the preferred region of interest corresponds to different solutions in the approximation sets from different scan pairs.

Because multiple solutions are provided with MO DIR that are spread in objective space, the clinical expert can navigate through these solutions and select an appropriate trade-off based on the underlying clinical scenario. In contrast, with traditional single DIR, only one of these solutions is provided to the clinical expert. Therefore, the

opportunity to evaluate other possibilities and make an informed decision specifically tuned to each patient is lost.

COMPARISON OF COMPUTATIONAL OVERHEAD

In the case of single DIR, a DIR network is trained multiple times with different weight combinations for each loss function following a certain strategy. The weights yielding the best aggregated performance on a validation set are used for final training. In MO DIR, multiple neural networks (in our case a single DIR network with multiple decoders) are trained. Therefore, the training overhead of MO DIR in terms of runtime is similar to that of single DIR. However, in MO DIR, the training is done in parallel, requiring more memory. In our implementation, training for $p = 27$ required ~39 GB and ~32 GB without and with a shared encoder, respectively, as compared to ~3.5 GB required for training a single DIR network.

6.4.2. COMPARISON OF PROPOSED MO DIR WITH LINEAR SCALARIZATION

In the proposed MO DIR, we used HV maximization to dynamically find the weights for each loss term such that the differently weighted loss training of different neural network heads yields their outputs diversely spread across the approximation front. It may be speculated that a similar diversity of outputs can be trivially obtained by training the different neural networks with uniformly distributed weights for different losses. Such an approach is called 'linear scalarization'. In Chapter 5, we compared linear scalarization with HV maximization for different shapes of Pareto fronts. We observed that the translation of the weights to a location on the front is dependent on the shape of the Pareto front, and is as such non-trivial. To investigate this in the case of MO DIR, we compared the proposed HV maximization based MO DIR approach with linear scalarization based MO DIR. To simulate the MO DIR set up with linear scalarization, we trained the different heads of our MO DIR neural network with weights corresponding to diversely distributed points in a grid. We used 27 grid points by enumerating over all the possible combinations for $w_1 \in \{0, 0.5, 1\}$, $w_2 \in \{0, 0.1, 0.5, 1\}$, and $w_3 \in \{0, 0.5, 1\}$ and omitting redundant (e.g., $\{0, 0.5, 0.5\}$ and $\{0.5, 0.5, 0.5\}$). It should be noted that this process of selecting linear scalarization weights is already slightly better than naive linear scalarization.

The approximation sets obtained from linear scalarization vs HV maximization based MO DIR are shown in Figure 6.3. It is apparent upon visual inspection of the figure that even though the weights used for linear scalarization were diversely distributed, still the obtained solutions are clustered along two edges of the expected triangle-like approximation front. There is a void of solutions in the center region of the expected triangle-like approximation front. This observation corroborates the results in the previous chapter (Chapter 5) - the diverse spread of solutions across the approximation front cannot be obtained trivially through linear scalarization - in the case of DIR as well. In contrast, visual inspection of the solutions in the approximation set obtained using HV maximization based MO DIR, shows a rough triangle-like shape with diversely distributed points in the center as well. This is because HV maximization ensures not only proximity to the Pareto front but also diversity across the approximation front.

Figure 6.3: Approximation sets obtained for the first four test scan pairs by linear scalarization (red circles) and the proposed HV-based MO DIR (blue triangles). The approximation sets from five different models trained with different training data splits are shown with slight variations in the color saturation to give an indication of model variance.

6.4.3. QUANTITATIVE COMPARISON OF DIR PERFORMANCE

Although TRE is a sparse metric and affected by inter- and intra-observer variation in the placement of landmarks, it is often used to quantitatively assess the performance of a DIR method. In this section, we compare the linear scalarization and proposed MO DIR approach in terms of mean TRE of 23 landmarks. First, we automatically select a single DIR solution from each approximation set. For this, we assume that a clinical expert would a posteriori select the DIR solution corresponding to minimum mean TRE of 23 landmarks. The underlying idea is that even if the TRE is not explicitly computed, the expert intuitively looks for solutions where landmarks that they are familiar with are well-aligned. In Table 6.1, we report the mean and standard deviation of this TRE value from 5 models, each trained on a different training data split to provide an estimate of model variance. We also report the associated folding in the DVF of the selected DIR solution. Although it is difficult to derive any clinical conclusions without inspecting the underlying DVFs, it can be observed that both linear scalarization and HV based MO DIR find quantitatively similar trade-offs between the best TRE values and associated DVF folding. This is not entirely surprising, given that the underlying DL architecture for DIR is the same for both methods.

One might notice a trend of higher TRE values and lower image folding in the selected solutions from HV maximization based MO DIR. However, it is important to realize that the training approach may play a role in this and that training for MO DIR and linear scalarization proceeds differently. Training neural networks with HV maximization is more complex as compared to using fixed weights as in the case with linear scalarization. This is because of the dynamically changing gradients for each network head as a consequence of the HV maximization goal. Therefore, if the exact weights corresponding to the desired trade-off between each objective are known a priori, linear scalarization may yield non-dominated solutions faster. For a fair comparison, we trained the networks in both the linear scalarization and the MO DIR approach with the same number of iterations. It may be possible that this was not the saturation point for both procedures. Ideally, upon saturation, we would expect both linear scalarization and HV maximization to obtain solutions with the same proximity to the Pareto front. However, obtaining the same diversity of solutions (for a given p) along the front is not guaranteed for linear scalarization. As demonstrated in Section 6.4.2, this is because the translation from scalarization weights to a well distributed set of solutions along the approximation front is not trivial. Therefore, achieving a diverse spread of solution through linear scalarization would require trying many more combinations. On the other hand, with the HV maximization based MO DIR approach, it can be achieved in a single go.

6.4.4. MO DIR WITHOUT AND WITH ADDITIONAL GUIDANCE

We aimed to gain insights into the effect of additional guidance from organ masks on the DIR performance. To this end, we compared the following two settings: a) MO DIR using $L_{ImageSimilarity}$, and $L_{DVFSmoothness}$ (no additional guidance), b) MO DIR using $L_{ImageSimilarity}$, $L_{DVFSmoothness}$, and $L_{SegSimilarity}$ (additional guidance). In Figure 6.4, the obtained approximation sets on test scan pairs from both settings are shown in the objective space of $L_{ImageSimilarity}$, $L_{DVFSmoothness}$, and $L_{SegSimilarity}$.

Table 6.1: Mean TRE and associated % folding in DVF of the 'best' solution in the approximation set obtained by linear scalarization and MO DIR, respectively for each test scan pair. In each approximation set, the solution corresponding to minimum mean TRE of 23 landmarks is assumed 'best' for the sake of quantitative comparison. Mean ± standard deviation from 5 models trained on different training data splits is reported without model selection.

Test scan	TRE before	**Linear Scalarization**		**MO DIR**	
		TRE	% folding	TRE	% folding
1	3.97	3.63 ± 0.04	0.29 ± 0.19	3.74 ± 0.03,	0.05 ± 0.03
2	4.71	4.53 ± 0.11	3.45 ± 0.38	4.66 ± 0.07,	2.00 ± 1.23
3	8.21	8.04 ± 0.10	1.33 ± 1.18	8.12 ± 0.06,	1.07 ± 1.64
4	9.07	8.18 ± 0.07	0.12 ± 0.15	8.58 ± 0.17,	0.47 ± 0.39
5	4.46	4.01 ± 0.06	0.80 ± 0.96	4.08 ± 0.07,	1.36 ± 1.01
6	5.55	4.52 ± 0.09	1.31 ± 0.17	4.69 ± 0.09,	0.76 ± 0.32
7	5.99	5.90 ± 0.03	0.26 ± 0.18	5.93 ± 0.02,	0.29 ± 0.13
8	4.39	3.96 ± 0.05	2.72 ± 0.88	4.06 ± 0.05,	1.72 ± 1.31
9	5.73	5.06 ± 0.06	0.87 ± 0.24	5.24 ± 0.13,	0.82 ± 0.97
10	3.80	3.72 ± 0.03	0.20 ± 0.28	3.70 ± 0.03,	0.11 ± 0.13
Mean ± SD across patients	5.59 ± 1.71	5.15 ± 1.63	1.14 ± 1.21	5.28 ± 1.69,	0.87 ± 1.04

Table 6.2: Maximum mean percent Dice score of four organs at risk (bowel bag, bladder, rectum, and sigmoid), and associated % folding for approximation sets obtained from MO DIR without and with guidance from organ masks, for each test scan pair. Mean ± standard deviation from 5 models from 5-fold cross-validation is reported.

Test scan	**No Guidance**		**Guidance**	
	% Dice	% folding	% Dice	% folding
1	97.63 ± 0.04	1.24 ± 0.26	99.28 ± 0.06,	0.77 ± 0.22
2	92.75 ± 0.09	1.03 ± 0.28	95.66 ± 0.26,	1.38 ± 0.28
3	96.25 ± 0.07	0.64 ± 0.37	98.99 ± 0.10,	0.93 ± 0.17
4	96.56 ± 0.04	0.69 ± 0.18	98.53 ± 0.13,	0.87 ± 0.28
5	94.58 ± 0.05	0.03 ± 0.07	98.24 ± 0.09,	0.66 ± 0.07
6	96.49 ± 0.13	1.36 ± 0.27	98.73 ± 0.13,	1.00 ± 0.37
7	96.93 ± 0.02	0.93 ± 0.53	99.01 ± 0.11,	0.96 ± 0.49
8	97.56 ± 0.09	0.63 ± 0.16	99.07 ± 0.07,	0.64 ± 0.11
9	95.63 ± 0.03	1.89 ± 0.44	98.01 ± 0.11,	1.09 ± 0.42
10	95.04 ± 0.01	0.67 ± 0.17	97.48 ± 0.12,	1.02 ± 0.26

The figure shows that training MO DIR with the additional guidance from organ masks, some solutions are obtained in the region corresponding to lower $L_{SegSimilarity}$ loss but higher $L_{ImageSimilarity}$ loss. These solutions underline the conflict between $L_{ImageSimilarity}$ and $L_{SegSimilarity}$, whose nature and causes could only be known after exploring the DIR outputs corresponding to these solutions. It is worth noting that with MO DIR, such an exploratory analysis is possible and straight-forward.

Furthermore, in Table 6.2, the maximum mean Dice score and % folding in the associated DVF of an approximation set is reported for each test scan pair. Similar to Figure 6.4, Table 6.2 also shows that by training DIR with additional guidance from organ masks, higher similarity between organ masks (indicated by high Dice scores) can be achieved without compromising with % folding in the DVFs. It is important to state here that the best solutions in the approximation sets according to Dice score (reported in Table 6.2) are not same as the best solutions according to TRE values (reported in Table 6.1), highlighting the nuances of evaluating a DIR outcome. Further, it is difficult to make clinically relevant performance comparisons solely based on quantitative values due to two reasons: a) mean Dice score is biased towards large organs, b) the solution corresponding to maximum Dice score may be overfitted to $L_{SegSimilarity}$ loss.

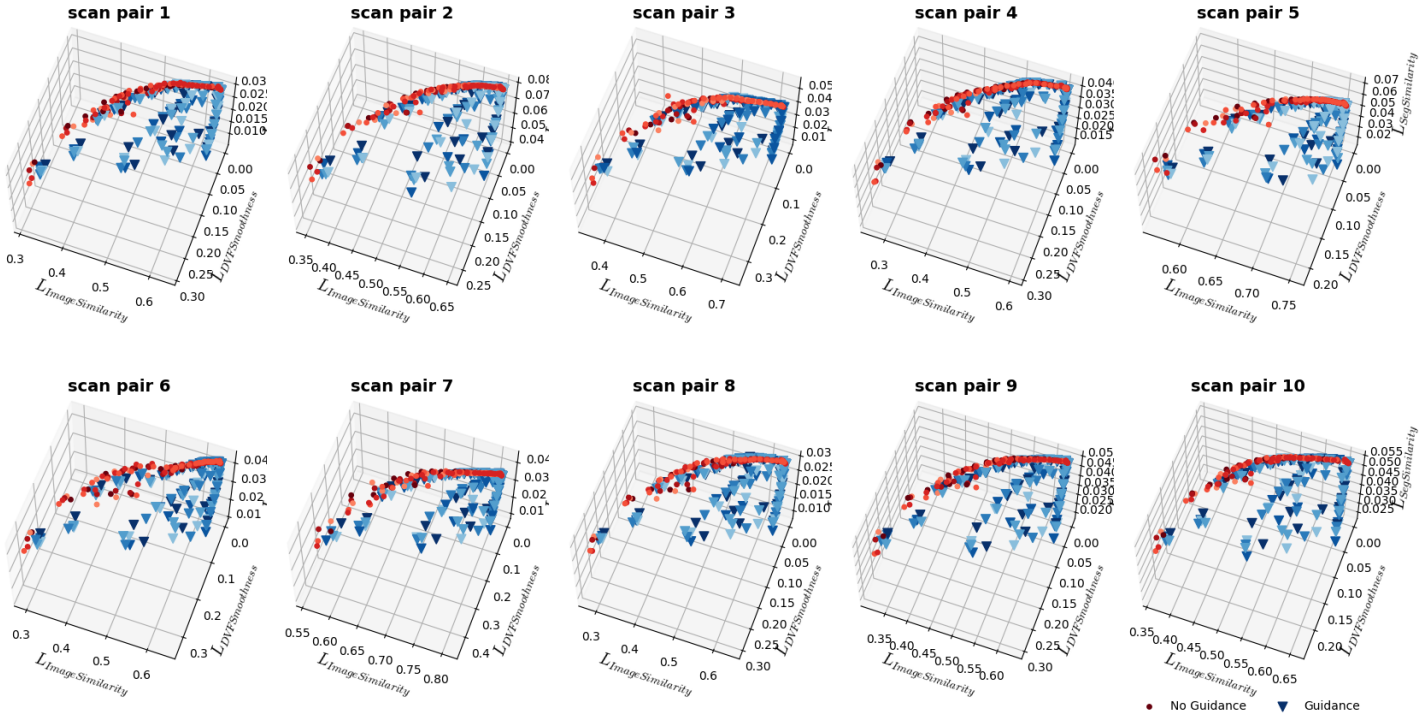


Figure 6.4: Effect of additional guidance. filled circles: MO DIR without additional guidance from organ contours, triangles: MO DIR with additional guidance from organ contours. $p = 27$. Approximation sets obtained from 5 models of 5-fold cross-validation are shown.

6.5. CONCLUSIONS AND DISCUSSION

We propose the first deep learning approach for MO DIR, which provides multiple DIR solutions diversely spread across the trade-off front between conflicting objectives. With such an approach, clinicians can evaluate multiple DIR solutions that are of potential interest and select the preferred one according to patient-specific and/or treatment-specific clinical criteria. While the prospect of clinicians having to review multiple DIR solutions may seem burdening, in a previous study using a dedicated user interface to navigate MO DIR solutions obtained through optimization (as opposed to deep learning as in this chapter), clinicians were positive, considering the use of MO DIR to be insightful [14]. We also demonstrated that a diverse spread of solutions across the approximation front such as obtained by the proposed MO DIR approach can not be trivially obtained by linear scalarization with diversely distributed weights. Although the potential utility of deep learning based MO DIR is evident from experimental results, the presented work is still only a proof-of-concept. Some of the limitations, open questions, and possible future research directions are as follows:

- HV maximization provides a straightforward way to distribute the solutions diversely on the approximation front without requiring any manual tuning. In future work, it would be interesting to investigate the use of the weighted HV [21] metric in MO DIR to steer the solutions to a desired region (if such a region can be defined clearly a priori). It is also important to investigate which part of the approximation front is more desired by involving clinicians as a posteriori decision-makers.

- In Figure 6.2 and Figure 6.3, the solutions seem more clustered in the region where $L_{ImageSimilarity}$ and $L_{SegSimilarity}$ are large and $L_{DVFSmoothness}$ is small. This could be because solutions in this region of the front are easy to obtain due to no or little deformation, or because of the corresponding shape of and local density along the Pareto front. It is known that setting the reference point differently can impact this [9] (also see Appendix 6.6.3). It is interesting to investigate this further in the future.

- In our proof-of-principle, we made certain choices e.g., number of objectives, number of solutions in the approximation set, type of additional guidance, type of neural network for DIR, in an effort to create a baseline deep learning based MO DIR approach. That said, the current approach leaves multiple improvement possibilities open in order to realize the complete potential of the MO perspective for DIR. For example, it can be improved by using a more sophisticated neural network for DIR, multi-resolution registration, constraints on tissue types, and diffeomorphism. All of these aspects are independent from the general idea and framework proposed in this chapter.

- The presented MO DIR work provides more insights than traditional single DIR approaches by showcasing the trade-offs between different objectives and how these trade-offs differ between scan pairs. However, the objectives are still average values per pair of scans. Practically, the DIR performance will likely not be uniform across the entire scan. Additionally, it is possible that clinically a

solution in the vicinity of a provided discrete solution on the approximation front is more desired. It is therefore essential to research in the direction of intuitively visualizing the DVFs and navigating across (and in the local neighborhood of) different solutions.

6.6. APPENDIX

6.6.1. EFFECT OF PARAMETER SHARING IN THE ENCODER

In Figure 6.5, 5 approximation sets obtained from 5 models after 5-fold cross-validation, by training the MO DIR approach with $p = 5$ for $L_{ImageSimilarity}$, and $L_{DVFSmoothness}$ losses without (filled circles) and with parameter sharing (triangles) in the encoder are shown for all the test scan pairs. The figure shows that parameter sharing does not impact the distribution of solutions on the front.

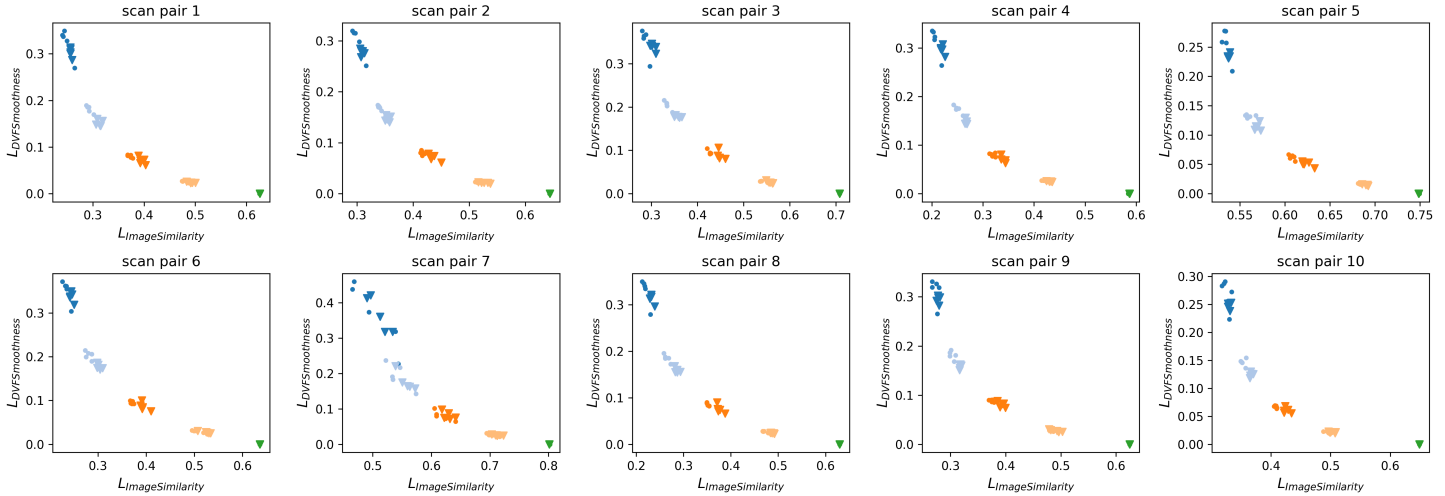
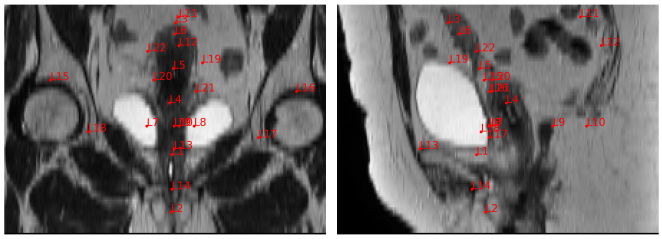


Figure 6.5: Effect of parameter sharing in the Encoder. filled circles: MO DIR without parameter sharing in the encoder, triangles: MO DIR with parameter sharing in the encoder. $p = 5$, $n = 2$. Approximation sets obtained from 5 models trained on different data splits are shown. Each color represents a DIR solution corresponding to a specific trade-off between $L_{ImageSimilarity}$ and $L_{DVFSmoothness}$.

6.6.2. DESCRIPTION OF LANDMARKS



coronal view sagittal view

L1	Internal urethral ostium
L2	External urethral ostium
L3	Uterus top
L4	Cervical ostium
L5	Isthmus
L6	Intra-uterine canal top
L7	Right ureteral ostium
L8	Left ureteral ostium
L9	Internal anal sfincter
L10	Os coccygis
L11	Most ventral intersections of S2-S3
L12	Most ventral intersections of S3-S4
L13	Anterior superior border sympysis (ASBS)
L14	Posterior inferior border sympysis (PIBS)
L15	Right femur head
L16	Left femur head
L17	Left acetabulum
L18	Right acetabulum
L19	Left ligament rotundum
L20	Right entrance of uterine artery to cervix
L21	Left entrance of uterine artery to cervix
L22	Right ligament rotundum
L23	Most ventral intersections of S1-S2

Figure 6.6: Description of landmarks. The landmarks are projected on a coronal (left) and sagittal (right) slice. L23 is not visible in this scan.

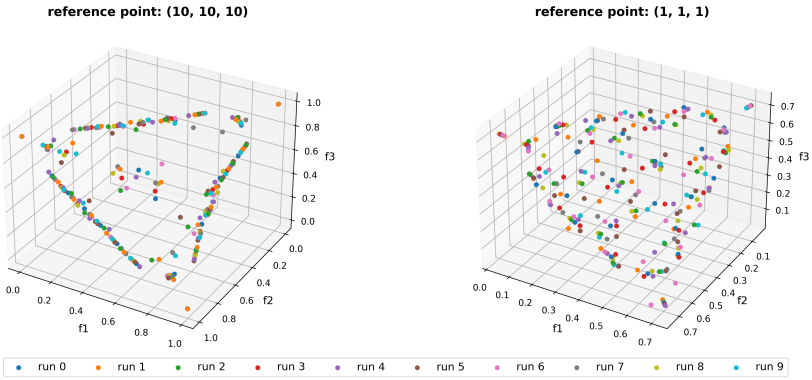


Figure 6.7: Effect of the location of reference point on the GenMED [4] benchmark problem. The Pareto front was approximated using 25 points. The solutions from 10 runs are shown for two different locations of the reference point.

6.6.3. EFFECT OF SELECTING REFERENCE POINT

The calculation of the HV (and consequently its gradients) is sensitive to the choice of the reference point [9], which, in turn, affects the spread of the solutions on the front. This is particularly the case for three or more objectives. In Figure 6.7, this phenomenon is illustrated with experiments on the convex GenMED problem with three objectives [4]. Briefly, in the GenMED problem, the n objectives (in our case, $n = 3$ i.e., f1, f2, f3 are the sum of square distances from n unit vectors. When the reference point is far away, the final solutions tend to cluster on the edges of the Pareto front. The spread of the points becomes more uniform across the Pareto front when the reference point is moved closer. Based on these empirical observations, we tuned the reference point for MO DIR training. We considered the following choices: (10, 10, 10), (1, 1, 1), (1, 1, 0.2), (0.5, 1, 1) based on observing the worst loss values after training. For experiments in this chapter, we selected (1, 1, 1) as the reference point because it provided well distributed points across the front based on visual inspection on validation set.

BIBLIOGRAPHY

[1] Tanja Alderliesten, Peter A. N. Bosman, and Arjan Bel. "Getting the most out of additional guidance information in deformable image registration by leveraging multi-objective optimization". In: *Medical Imaging 2015: Image Processing*. Vol. 9413. Proc. SPIE. International Society for Optics and Photonics. 2015, 94131R.

[2] Georgios Andreadis, Peter A. N. Bosman, and Tanja Alderliesten. "MOREA: a GPU-accelerated Evolutionary Algorithm for Multi-Objective Deformable Registration of 3D Medical Images". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. 2023, pp. 1294–1302.

[3] Guha Balakrishnan et al. "VoxelMorph: A Learning Framework for Deformable Medical Image Registration". In: *IEEE Transactions on Medical Imaging* 38.8 (2019), pp. 1788–1800. DOI: 10.1109/TMI.2019.2897538.

[4] Peter A. N. Bosman. "On gradients and hybrid evolutionary algorithms for real-valued multiobjective optimization". In: *IEEE Transactions on Evolutionary Computation* 16.1 (2011), pp. 51–69.

[5] Bob D De Vos et al. "A deep learning framework for unsupervised affine and deformable image registration". In: *Medical Image Analysis* 52 (2019), pp. 128–143.

[6] Kalyanmoy Deb, Karthik Sindhya, and Jussi Hakanen. "Multi-objective optimization". In: *Decision Sciences*. CRC Press, 2016, pp. 161–200.

[7] Alessa Hering et al. "CNN-based lung CT registration with multiple anatomical constraints". In: *Medical Image Analysis* 72 (2021), p. 102139.

[8] C-L Hwang and Abu Syed Md Masud. *Multiple objective decision making—methods and applications: a state-of-the-art survey*. Vol. 164. Springer Science & Business Media, 2012.

[9] Hisao Ishibuchi et al. "How to specify a reference point in hypervolume calculation for fair performance comparison". In: *Evolutionary Computation* 26.3 (2018), pp. 411–440.

[10] Hongming Li and Yong Fan. "Non-rigid image registration using self-supervised fully convolutional networks without training data". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1075–1078.

[11] Hongming Li, Yong Fan, and for the Alzheimer's Disease Neuroimaging Initiative. "MDReg-Net: Multi-resolution diffeomorphic image registration using fully convolutional networks with deep self-supervision". In: *Human Brain Mapping* 43.7 (2022), pp. 2218–2231. DOI: https://doi.org/10.1002/hbm.25782. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.25782. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.25782.

[12] R. Mohammadi et al. “Evaluation of deformable image registration algorithm for determination of accumulated dose for brachytherapy of cervical cancer patients”. In: *Journal of Contemporary Brachytherapy* 11 (5 2019), pp. 469–478.

[13] Takumi Nakane, Haoran Xie, and Chao Zhang. “Image Deformation Estimation via Multiobjective Optimization”. In: *IEEE Access* 10 (2022), pp. 53307–53323.

[14] Kleopatra Pirpinia et al. “Simplex-based navigation tool for a posteriori selection of the preferred deformable image registration outcome from a set of trade-off solutions obtained with multiobjective optimization for the case of breast MRI”. In: *Journal of Medical Imaging* 5.4 (2018), p. 045501. DOI: 10.1117/1.JMI.5.4.045501. URL: https://doi.org/10.1117/1.JMI.5.4.045501.

[15] Kleopatra Pirpinia et al. “The feasibility of manual parameter tuning for deformable breast MR image registration from a multi-objective optimization perspective”. In: *Physics in Medicine & Biology* 62.14 (2017), p. 5723.

[16] Bastien Rigaud et al. “Deformable image registration for dose mapping between external beam radiotherapy and brachytherapy images of cervical cancer”. In: *Physics in Medicine & Biology* 64.11 (2019), p. 115023.

[17] Mohammad Salehi et al. “Deep Learning-based Non-rigid Image Registration for High-dose Rate Brachytherapy in Inter-fraction Cervical Cancer”. In: *Journal of Digital Imaging* (2022), pp. 1–14.

[18] David A Van Veldhuizen and Gary B Lamont. “Multiobjective evolutionary algorithms: Analyzing the state-of-the-art”. In: *Evolutionary Computation* 8.2 (2000), pp. 125–147.

[19] Bob D. de Vos et al. “End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by M. Jorge Cardoso et al. Cham: Springer International Publishing, 2017, pp. 204–212. ISBN: 978-3-319-67558-9.

[20] Tiandi Zhao et al. “Evaluating the accumulated dose distribution of organs at risk in combined radiotherapy for cervical carcinoma based on deformable image registration”. In: *Brachytherapy* (2022). ISSN: 1538-4721. DOI: https://doi.org/10.1016/j.brachy.2022.09.001. URL: https://www.sciencedirect.com/science/article/pii/S1538472122001635.

[21] Eckart Zitzler, Dimo Brockhoff, and Lothar Thiele. “The Hypervolume Indicator Revisited: On the Design of Pareto-compliant Indicators Via Weighted Integration”. In: *Evolutionary Multi-Criterion Optimization*. Ed. by Shigeru Obayashi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 862–876. ISBN: 978-3-540-70928-2.

7

DISCUSSION

7.1. KEY TAKEAWAYS

In this thesis, we developed several deep learning algorithms with a particular focus on DIR for radiation treatment of locally advanced cervical cancer. We can present the following take-home messages from this thesis.

7.1.1. LOOK BEYOND SUPERVISED DEEP LEARNING

A major shortcoming of the common practices of deep learning research in the medical imaging domain is restricting to the domain of supervised learning i.e., relying on fully annotated datasets. This translates to requiring a lot of time from clinical experts to annotate data for deep learning research, which is counter-productive because the motivation of deep learning research is to help clinical experts save time. In this thesis, we made conscious efforts to reduce the need for clinical experts to annotate data specifically for deep learning research. In Chapter 2 and Chapter 3, we proposed an approach to learn thousands of landmark correspondences in medical images without needing manual annotations. Similarly, in Chapter 4, we proposed an approach for OARs segmentation by making efficient use of clinically available images and annotations. Furthermore, in Chapter 6, we used unsupervised learning to develop a deep learning method for DIR.

The results in Chapter 2 and Chapter 4 showed that the proposed algorithms outperformed their traditional counterparts, indicating that it is possible to work outside the domain of supervised learning without compromising on performance. In fact, the results in Chapter 4 indicated that making use of partially labeled data with semi-supervised learning led to better performance than learning from a small dataset using a purely supervised learning setup.

Overall, the work in this thesis demonstrated different ways to exploit non-traditional learning approaches (e.g., unsupervised learning, semi-supervised learning, and self-supervised learning) to get the most out of clinically available data. The work in this thesis provides a strong argument to shift the focus of deep learning research for medical imaging towards more unconventional learning paradigms.

7.1.2. MORE DATA IS BETTER, BE IT IMPERFECT

Along with not restricting to the supervised learning paradigm, we also embraced the imperfections of clinical data. Specifically, in Chapter 4, we used the clinically available images and annotations from patients who were treated for abdominal cancer to train a deep learning model to be used for scans from cervical cancer patients. In this process, the data had image- and label-inhomogeneity and partial annotations. Through experiments, we showed that dealing with imperfect data was worth it and yielded better results than working with a clean but smaller dataset.

7.1.3. ADDITIONAL GUIDANCE IS HELPFUL IN DIR

Using additional guidance through corresponding landmarks or through contours is an intuitive way to improve DIR performance. There exists plenty of literature supporting this idea [15, 5, 12, 1, 6]. The work presented in Chapter 3 provides further empirical evidence in this direction for the case of DIR in the pelvic region. The results showed an improvement in DIR performance when automatically identified landmarks are incorporated into the DIR approach. The chapter also provides some insights on what attributes of the automatic landmarks are more beneficial to DIR.

In Chapter 6, the masks of OARs have been used to provide additional guidance to DIR. The results show that with the use of additional guidance from the masks of OARs, DIR solutions exhibiting better alignment of the contours of OARs and reduced image folding can be obtained. However, the results also indicate that these DIR solutions may not necessarily provide the minimum target registration errors between manually annotated landmarks. The results highlight the nuance of evaluating DIR outcomes, and hence the importance of further research in the direction of qualitative evaluation of DIR.

7.1.4. KEEPING HUMAN-IN-THE-LOOP

While on the one hand deep learning is achieving tremendous success across a variety of cognitive tasks, on the other hand there is an increasing concern regarding trustworthiness of deep learning solutions for real-world applications [8]. Keeping a human as a final gatekeeper is a viable solution that can increase the trust in deep learning solutions. In Chapter 5, we developed a novel approach to enable a posteriori decision making with deep learning. In Chapter 6, we used this approach for MO DIR. With such an approach, deep learning is used to identify a diverse subset of all possibilities and the users (in case of Chapter 6, clinical experts) can evaluate the possible outcomes qualitatively to make an informed decision.

7.2. SCIENTIFIC AND SOCIETAL IMPLICATIONS

In this thesis, deep learning approaches were developed for the automatic detection of corresponding landmarks, the segmentation of OARs, and MO DIR. During the development, the primary focus was on medical images pertaining to cervical cancer radiation treatment. However, many of the developed approaches in this thesis can be used for broader applications. The approach for corresponding landmarks could potentially be used for the automatic validation of DIR [10, 4]. The corresponding landmarks identified by the proposed approach in Chapters 2 and 3 do not pertain to any particular semantic feature description. This makes the corresponding landmarks, in principle, suitable for applications where manually identifying the anatomical landmarks is not feasible, for example, automatic bowel motion tracking [14, 3]. Similarly, the semi-supervised learning approach proposed in Chapter 4 can be used to efficiently make use of partially-labeled clinically available data from a different modality or use-case. Intuitively, the approaches developed in chapters 2, 3, 4, and 6 are relevant to the broader field of radiation treatment concerning cancer in other anatomical regions, e.g., head and neck, prostate, bladder, ovaries, pancreas. Similarly, the application of the MO learning approach proposed in Chapter 5 extends to all real-world applications that utilize machine learning and involve multiple conflicting objectives.

In regard to the radiation treatment workflow mentioned in the introduction chapter (Chapter 1), the approaches developed in this thesis can potentially improve the automatic contouring of OARs, and improve the performance of DIR through additional guidance from corresponding landmarks and contours of OARs. With the deep learning based MO DIR, multiple DIR solutions, each representing a different trade-off between given performance metrics, can be presented to clinicians. The clinicians can then select the most appropriate solution while taking into account patient-specific criteria that were not part of DIR. This will improve patient-specific quality assurance of DIR, potentially increasing the clinical adoption of DIR. Overall, the deep learning approaches presented in this thesis may contribute to a more efficient and effective radiation treatment workflow.

7.3. LIMITATIONS AND CHALLENGES

While the approaches developed in this thesis demonstrate potential benefits, there are certain limitations. The first and foremost limitation is that even though the deep learning models in the thesis were developed with real-world data and for a real-world application, they are not good enough to be deployed in the hospital as is. While the developed models showed good generalization on unseen data, the test data still came from the same hospital or demographic region as the training data. Since the models were not tested on datasets from multiple sites, we can not say much about the generalization potential of the developed models outside the hospital of which the data was used for development. The inclusion of each new site would require retraining as well as validation, which is burdening in terms of time and efforts.

The second limitation of the work presented in this thesis is that it did not regard various open challenges in the DIR of pelvic anatomy. These challenges include a) sliding tissue at the organ boundary, b) content mismatch due to gas pockets, insertion of an applicator, and tumor shrinkage, c) large deformations of the bladder, and d) ensuring bio-mechanical plausibility. It is crucial to model these challenges explicitly in the deep learning paradigm before considering the clinical deployability of DIR solutions based on deep learning for cases with complex changes in anatomy. It is also vital to investigate the developed approaches for their applicability and added benefits in cross-modality DIR. Furthermore, the current thesis leaves scope for improvements in the direction of building more robust deep learning models e.g., robustness against variations in scanner types, image quality, and imaging protocols. In Chapter 2 and 3, it is noted that the performance of automatic landmark detection may be improved with a different neural network architecture for feature extraction. Similarly, the approaches mentioned in Chapter 4 and 6 are applicable for a different neural network architecture, which has not been investigated. In the thesis, conscious efforts have been put to efficiently use clinically available datasets. However, most of the works focus on utilizing either only a single modality (CT or MRI) or data from a single hospital. This leaves the possibility for developing more generalized models by utilizing multiple modalities or data from multiple sites.

Another major limitation of the presented work is the lack of a validation study to investigate the feasibility and the true potential of the MO perspective for DIR. An earlier study using an optimization based MO DIR approach reported positive perception of the usability of MO DIR in a clinical setting for breast MRI scans[11]. A similar validation study for applications in cervical cancer radiation treatment would be beneficial to judge the true potential and added value of the MO DIR approach in a clinical setting.

An obvious next question would be *whether it is only a matter of a few more Ph.D. theses on the topic of extending the proposed approaches in this thesis to more capabilities and conducting validation studies on bigger demographic regions.* The answer is that it is not so straightforward. Apart from ensuring good performance of deep learning solutions, several other challenges need resolving e.g., integration in the clinical workflow and ensuring significant reduction in clinical workload [7, 13]. Further, on the path of having a research output with a potential for clinical use and its actual use in the hospitals, there exist complexities, hurdles, and nuances such as the financial cost of acquiring intellectual property, patenting, quality assurance, clinical trials, and getting approvals from regulatory bodies. In the end, what makes it into clinical software is decided by the executives of software companies by judging what will bring the most sales. Most of these factors can not be controlled from within an academic research group.

What can be done to improve the societal utilization of applied research such as this thesis? First, we should understand that research advancement of any sort is still a step forward, however far away it be from its societal utilization. Such small leaps are an essential building block towards realizing a final product with a capability of direct societal utilization. There exist some endeavors on bridging the gap between the applied research and its societal utilization. For example, NVIDIA and King's college,

London initiated project-MONAI[1], which provides a collaborative framework for accelerating applied deep learning research and its clinical translation. Other initiatives include gathering, anonymizing, annotating, and open-sourcing big medical imaging datasets for collaborative and reproducible research e.g., the Stanford AIMI dataset[2], RadImageNet[3], and the medical segmentation decathlon [2]. We believe that more collaborative research between research institutes, hospitals, and industry stakeholders will help bridge the gap between applied research and clinical deployment at scale.

7.4. FUTURE DIRECTIONS

One of the motivations for using deep learning for DIR in this thesis was that with deep learning, DIR solutions can be obtained within a few seconds for a complete 3D scan. This makes deep learning up to a thousand times faster than traditional approaches and hence increases the potential for clinical adoption in some cases where time is critical. However, a key caveat of relying only on deep learning for obtaining DIR solutions is its sub-optimal performance due to the generalization gap between the training and testing data. On the other hand, traditional approaches for DIR perform optimization with the given pair of images at the time of inference and as a result have more potential to yield better solutions. Of all existing optimization approaches, EAs are perhaps most noteworthy in light of this thesis due to their capability to work well for MO problems. Therefore, it is very interesting to investigate a combination of both an EA and deep learning for DIR, to get the best out of these techniques. One way to do so could be to use the predictions from the deep learning based DIR method as a starting point for an EA optimizer so that optimization is faster while final DIR solutions are better. Another way could be to plug in the methods for landmark correspondence detection and OARs segmentation to an EA-based DIR approach to use the additional guidance from landmarks and OARs segmentations in a multi-objective manner.

Alternatively, one could develop the methods in this thesis with a different perspective. The methods developed in this thesis have focused on effective training and not so much on the improvement of the neural network architecture. It can be speculated that some results, e.g., similar performance on the segmentation of OARs (refer to Chapter 4) could be achieved by a single training but with a much better (data-efficient) architecture. Additionally, as noted in Chapter 2, the results on the automatic identification of corresponding landmarks might be improved with the use of a more sophisticated architecture. It might be overkill and deviating from the main aim, but nonetheless interesting, to investigate the use of neural architecture search for the tasks mentioned in this thesis [9, 19, 16, 18, 17].

Further, while the concept of multiple DIR solutions representing the trade-off front of different objectives has better potential for clinical adoption than a single-objective perspective, the MO approach proposed in this thesis may still not be comprehensive. In a practical scenario, it may be required to reduce the number of choices in order to reduce the time spent on evaluating them. A clinical expert might

[1] https://monai.io/index.html
[2] https://aimi.stanford.edu/shared-datasets/shared-datasets/shared-datasets/shared-datasets
[3] https://www.radimagenet.com

also want to explore other DIR solutions on the trade-off front (that are not part of the provided set of solutions) in the vicinity of a provided trade-off. Therefore, the proposed MO learning approach in combination with approaches on either Pareto front exploration or interpolation between different points from the Pareto set will potentially provide a more comprehensive solution.

As a final note, we should remember the following.

Rome was not built in a day.

John Heywood

It always seems impossible until it's done.

Nelson Mandela

BIBLIOGRAPHY

[1] Tanja Alderliesten, Peter A. N. Bosman, and Arjan Bel. "Getting the most out of additional guidance information in deformable image registration by leveraging multi-objective optimization". In: *Medical Imaging 2015: Image Processing*. Vol. 9413. Proc. SPIE. International Society for Optics and Photonics. 2015, 94131R.

[2] Michela Antonelli et al. "The medical segmentation decathlon". In: *Nature Communications* 13.1 (2022), p. 4128.

[3] Danique L.J. Barten et al. "A 3D cine-MRI acquisition technique and image analysis framework to quantify bowel motion demonstrated in gynecological cancer patients". In: *Medical Physics* 48.6 (2021), pp. 3109–3119. DOI: https://doi.org/10.1002/mp.14851. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.14851. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.14851.

[4] Guillaume Cazoulat et al. "Detection of vessel bifurcations in CT scans for automatic objective assessment of deformable image registration accuracy". In: *Medical Physics* 48.10 (2021), pp. 5935–5946.

[5] Zeinab Ghassabi et al. "An efficient approach for robust multimodal retinal image registration based on UR-SIFT features and PIIFD descriptors". In: *EURASIP Journal on Image and Video Processing* 2013.1 (2013), p. 25. ISSN: 1687-5281. DOI: 10.1186/1687-5281-2013-25. URL: https://doi.org/10.1186/1687-5281-2013-25.

[6] Dong Han et al. "Robust anatomical landmark detection with application to MR brain image registration". In: *Computerized Medical Imaging and Graphics* 46 (2015), pp. 277–290. ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.compmedimag.2015.09.002. URL: http://www.sciencedirect.com/science/article/pii/S089561111500124X.

[7] Thomas C Kwee and Robert M Kwee. "Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence". In: *Insights into imaging* 12 (2021), pp. 1–12.

[8] Bo Li et al. "Trustworthy AI: From Principles to Practices". In: *ACM Computing Surveys* 55.9 (Jan. 2023). ISSN: 0360-0300. DOI: 10.1145/3555803. URL: https://doi.org/10.1145/3555803.

[9] Li Liu et al. "DKNAS: A Practical Deep Keypoint Extraction Framework Based on Neural Architecture Search". In: *2022 International Conference on Robotics and Automation (ICRA)*. 2022, pp. 5643–5649. DOI: 10.1109/ICRA46639.2022.9812101.

[10] Chiara Paganelli et al. ""Patient-specific validation of deformable image registration in radiation therapy: Overview and caveats"". In: *Medical Physics* 45.10 (2018), e908–e922. DOI: https://doi.org/10.1002/mp.13162. eprint:

https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.13162. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.13162.

[11] Kleopatra Pirpinia et al. "Simplex-based navigation tool for a posteriori selection of the preferred deformable image registration outcome from a set of trade-off solutions obtained with multiobjective optimization for the case of breast MRI". In: *Journal of Medical Imaging* 5.4 (2018), p. 045501. DOI: 10.1117/1.JMI.5.4.045501. URL: https://doi.org/10.1117/1.JMI.5.4.045501.

[12] Thomas Polzin et al. "Combining automatic landmark detection and variational methods for lung CT registration". In: *Fifth International Workshop on Pulmonary Image Analysis*. 2013, pp. 85–96.

[13] European Society of Radiology (ESR) communications@ myesr. org Becker Christoph D. Kotter Elmar Fournier Laure Martı-Bonmatı Luis. "Current practical experience with artificial intelligence in clinical radiology: a survey of the European Society of Radiology". In: *Insights into Imaging* 13.1 (2022), p. 107.

[14] F Slevin et al. "Inter and intra-fraction bowel motion during abdomino-pelvic stereotactic ablative radiotherapy". In: *Radiotherapy and Oncology*. Vol. 133. Supplement 1. Elsevier BV. 2019, S1082–S1083.

[15] Eliana M. Vásquez Osorio et al. "Accurate CT/MR vessel-guided nonrigid registration of largely deformed livers". In: *Medical Physics* 39.5 (2012), pp. 2463–2477. DOI: 10.1118/1.3701779. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1118/1.3701779. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3701779.

[16] Yu Weng et al. "NAS-Unet: Neural architecture search for medical image segmentation". In: *IEEE access* 7 (2019), pp. 44247–44257.

[17] Jiong Wu and Yong Fan. "HNAS-reg: hierarchical neural architecture search for deformable medical image registration". In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2023, pp. 1–4.

[18] Qihang Yu et al. "C2FNAS: Coarse-to-Fine Neural Architecture Search for 3D Medical Image Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2020.

[19] Zhuotun Zhu et al. "V-NAS: Neural Architecture Search for Volumetric Medical Image Segmentation". In: *2019 International Conference on 3D Vision (3DV)*. 2019, pp. 240–248. DOI: 10.1109/3DV.2019.00035.

ACKNOWLEDGEMENTS

A PhD is often described as a long and lonely journey of perseverance. I began mine with this expectation, but I was fortunate to have a strong support system that kept me afloat throughout.

At the heart of this support were my promotor, **Peter Bosman**, and copromotors, **Tanja Alderliesten** and **Henrike Westerveld**. Peter, with his infectious enthusiasm and vast expertise, engaged in deep discussions and problem-solving like a fellow PhD student rather than a professor. Tanja's keen eye for detail, scientific rigor, and ability to maintain a perfect balance between meticulousness and pragmatism ensured precision and clarity in my work. Henrike not only guided me through the clinical aspects but also provided encouragement as I neared the finish line. I am deeply grateful for their mentorship and for the solid foundation Peter and Tanja built with the MODIR project before my journey even began.

I was also privileged to share this journey with wonderful colleagues. **Timo** was the best officemate one could ask for, always ready for conversations - scientific or otherwise. I will always cherish the coffee breaks, discussions, and conference trips shared with **Hoang**, **Marco**, **Arkadiy**, and **Anton**. I am also thankful to my colleagues from AMC, **Ziyuan**, **Marjolein**, **Stef**, **Jeroen**, and **Jan**, for fruitful discussions and collaboration. A special thanks to **Alex** for being ready to get nerd-sniped. I am also thankful to other members of the Evolutionary Intelligence group, MODIR teammates, including **Cedric**, **Georgios**, and colleagues from industry partners (Xomnia and Elekta) for making my foreign work experience truly memorable. Additionally, I am grateful to **NWO**, **Dutch Ministry of Economic Affairs**, **Elekta**, and **Xomnia** for funding the project and to **CWI**, **Amsterdam UMC**, **LUMC**, and **TU Delft** (especially the graduate school) for their administrative and institutional support.

Beyond academia, I am thankful for the many inspiring women in my life—my sisters **Anjali**, **Mamta**, my sisters-in-law **Sunil**, **Sunita**, and my sisters-in-life **Xanthate**, **Shammi**, and **Sumiti**. Though we have been separated by oceans and sometimes observed each other's lives from a distance, the resilience with which each of them has faced life's challenges has been a constant source of inspiration for me. Their strength has encouraged me to persevere through my own struggles. I can never thank Xanthate and Shammi enough for being more than family members for my daughter and for ensuring that I celebrated every small milestone in my life. I am thankful to **Subhasree**, whose straightforwardness and pragmatism mirror my own. Our coffee conversations about everything under the sun and beyond were truly refreshing. I am also grateful to my friends, in India, the Netherlands and around the world, whose presence gives me a sense of security and fills my life with joy, laughter, and colors.

Coming from a small farming village in India, this journey was far from conventional. I owe everything to my parents and parents-in-law, who supported me unconditionally despite not fully understanding my work.

Finally, I owe my deepest gratitude to **Kuldeep** — my partner, my personal manager, my boss, and my biggest support. You have managed everything from my time, health, and mood to my work performance, expectations from life, and overall sanity. Thank you for knowing exactly what to say when I needed it the most, for lifting me up when I felt low — whether by reminding me how far I have come or joking that you would help me fake a degree if all else failed. Thank you not only for your unwavering support but also for being the strict boss who nudged me to work when procrastination took over, even on late nights and weekends. Most importantly, thank you for giving me the confidence that I can achieve anything I set my mind to, while also making me feel I already have everything I could ever wish for.

CURRICULUM VITÆ

Monika GREWAL

1989	Born in India	

EDUCATION

2006–2010	B.Tech, ECE	TIT&S, Bhiwani, India
2010–2012	M.Tech, ECE	JIIT, Noida, India

EXPERIENCE

2012–2016	National Brain Research Centre, India	Neuroimaging & neurospectroscopic research
2016–2018	Paralleldots, Inc., India	AI and medical imaging research
2019–2023	Centrum Wiskunde & Informatica, The Netherlands	Deep learning, evolutionary algorithms, medical imaging research
2024–Present	Nicolab, The Netherlands	AI and medical imaging research

PROJECTS

Automatic Gaze Encoding from Eye-tracking Videos. Developed an AI pipeline to map eye-gaze locations in eye-tracking videos to static images of stimuli.

Medical Image Diagnostics using Deep Learning. Developed novel deep learning methods for brain hemorrhage detection in Computed Tomography (CT) scans, lung disease detection in X-Ray, tooth cavity segmentation in dental images.

Neural connectivity differences for processing of happy and sad facial expressions. Designed and executed a functional Magnetic Resonance Imaging (fMRI) study to identify the different brain regions responsible for processing happy and sad facial expressions.

LIST OF PUBLICATIONS

10. **Grewal, M.**, Westerveld, H., Bosman, P., & Alderliesten, T. (2024, June). *Multi-Objective Learning for Deformable Image Registration.* In Medical Imaging with Deep Learning.

9. **Grewal, M.**, van Weersel, D., Westerveld, H., Bosman, P., & Alderliesten, T. (2024, February). *Learning Clinically Acceptable Segmentation of Organs at Risk in Cervical Cancer Radiation Treatment from Clinically Available Annotations.* In Medical Imaging with Deep Learning (pp. 260-273). PMLR.

8. Deist, T. M.*, **Grewal, M.***, Dankers, F. J., Alderliesten, T., & Bosman, P. A. (2023, March). *Multi-Objective Learning using HV Maximization.* In International Conference on Evolutionary Multi-Criterion Optimization (pp. 103-117). Cham: Springer Nature Switzerland. **Authors contributed equally.*

7. **Grewal, M.**, Wiersma, J., Westerveld, H., Bosman, P. A., & Alderliesten, T. (2023, January). *Automatic Landmark Correspondence Detection in Medical Images with an Application to Deformable Image Registration.* Journal of Medical Imaging, 10(1), 014007-014007.

6. Bosma, M. M. A., Dushatskiy, A, **Grewal, M.**, Alderliesten, T., Bosman, P. A. (2022, April). *Mixed-block Neural Architecture Search for Medical Image Segmentation.* In Medical Imaging 2022: Image Processing, 120320S. SPIE.

5. **Grewal, M.**, Deist, T. M., Wiersma, J., Bosman, P. A., & Alderliesten, T. (2020, March). *An End-to-End Deep Learning Approach for Landmark Detection and Matching in Medical Images.* In Medical Imaging 2020: Image Processing (Vol. 11313, pp. 548-557). SPIE.

4. **Grewal, M.**, Srivastava, M.M., Kumar, P. & Varadarajan, S., (2018, May) *RADnet: Radiologist Level Accuracy using Deep Learning for Hemorrhage Detection in CT Scans.* In IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, pp. 281-284.

3. Kumar, P., **Grewal, M.**, Srivastava, M.M. (2018, June). *Boosted Cascaded Convnets for Multilabel Classification of Thoracic Diseases in Chest Radiographs.* In: Image Analysis and Recognition. ICIAR 2018. Lecture Notes in Computer Science, vol 10882. Springer, Cham.

2. **Grewal, M.**, Dabas, A., Saharan, S., Barker, P.B., Edden, R.A.E. and Mandal, P.K. (2016, June), *GABA Quantitation using MEGA-PRESS: Regional and Hemispheric differences.* Journal of Magnetic Resonance Imaging, 44: 1619-1623.

1. *A Method for Metabolite Signal Quantitation for Magnetic Resonance Spectroscopy Data* 2017. International patent: WO2017125800 A1

SIKS DISSERTATION SERIES

2016
01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UvA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UvA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UvA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media

28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning

30 Ruud Mattheij (TiU), The Eyes Have It

31 Mohammad Khelghati (UT), Deep web content monitoring

32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example

34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment

35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation

36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry

38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design

39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice

44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration

45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control

46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

---

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UvA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TU/e), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UvA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VUA), Logics for causal inference under uncertainty

23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VUA), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VUA), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TU/e), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TU/e), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TU/e), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech

30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data

06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VUA), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming

36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OU), Learning visually grounded and multilingual representations

---

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TU/e), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OU), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation Methods for Long-Tail Entity Recognition Models

12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst

31 Gongjin Lan (VUA), Learning better – From Baby to Better

32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

---

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

22 Sihang Qiu (TUD), Conversational Crowdsourcing

23 Hugo Manuel Proença (UL), Robust rules for prediction and description

24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing

25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications

26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables

27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You

28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

---

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games

02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey

03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare

04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework

05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization

06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding

07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics

08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design

09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach

10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines

11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring

12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases

13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge

14 Michiel Overeem (UU), Evolution of Low-Code Platforms

15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining

16 Pieter Gijsbers (TU/e), Systems for AutoML Research

17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation

19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media

21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations

23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents

24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values

25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences

27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences

28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems

29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality

30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays

31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management

32 Cezara Pastrav (UU), Social simulation for socio-ecological systems

33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management

34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change

35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions

02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts

03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations

04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval

05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications

06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment

07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning

08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning

09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing

11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications

12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries

13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation

14 Selma Čaušević (TUD), Energy resilience through self-organization

15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models

16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters

17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight

18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation

19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals

20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning

21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain

22 Alireza Shojaifar (UU), Volitional Cybersecurity

23 Theo Theunissen (UU), Documentation in Continuous Software Development

24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning

25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs

26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour

27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions

28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts

29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

---

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education

02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems

03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis

04 Mike Huisman (UL), Understanding Deep Meta-Learning

05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair

06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems

08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation

09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks

10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science

11 withdrawn

12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning

13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence

14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling

15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health

16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems

17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design

18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations

19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior

20 Ritsart Anne Plantenga (UL), Omgang met Regels

21 Federica Vinella (UU), Crowdsourcing User-Centered Teams

22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts

23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution

24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour

25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics

26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework

27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning

28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs

29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases

30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting

31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows

32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining

33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method

35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach

36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs

37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models