



Delft University of Technology

PedVision

A manual-annotation-free and age scalable segmentation pipeline for bone analysis in hand X-ray images

Homayounfar, Morteza; Bierma-Zeinstra, S. M.A.; Zadpoor, Amir A.; Tümer, Nazli

DOI

[10.1016/j.bspc.2025.108569](https://doi.org/10.1016/j.bspc.2025.108569)

Publication date

2026

Document Version

Final published version

Published in

Biomedical Signal Processing and Control

Citation (APA)

Homayounfar, M., Bierma-Zeinstra, S. M. A., Zadpoor, A. A., & Tümer, N. (2026). PedVision: A manual-annotation-free and age scalable segmentation pipeline for bone analysis in hand X-ray images. *Biomedical Signal Processing and Control*, 112, Article 108569. <https://doi.org/10.1016/j.bspc.2025.108569>

Important note

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

Copyright

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

Takedown policy

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



PedVision: A manual-annotation-free and age scalable segmentation pipeline for bone analysis in hand X-ray images

Morteza Homayounfar^{a,b,*}, S.M.A. Bierma-Zeinstra^b, Amir A. Zadpoor^{a,1}, Nazli Tümer^{a,1}

^a Department of Biomechanical Engineering, Delft University of Technology (TU Delft), Delft, Netherlands (the)

^b Department of General Practice, University Medical Centre Rotterdam (Erasmus MC), Rotterdam, Netherlands (the)

ARTICLE INFO

Keywords:

Pediatric image analysis
Visual foundation model
Hand bone segmentation
X-rays
Manual-annotation-free
Human-in-the-loop training

ABSTRACT

Medical image analysis often involves time-consuming annotation processes. Pediatric image analysis introduces additional complexity due to the scarcity of data, noise, and growth-related anatomical variations, particularly in bone analysis, where bone structures evolve more slowly compared to other organs. This study aims to develop a segmentation model that scales across different age groups, reduces annotation effort, and ensures high accuracy, particularly in low-quality images.

To address these challenges, we propose a segmentation pipeline (PedVision) that first uses a Region of Interest (ROI) network to identify relevant regions, followed by a foundation model that translates each region into meaningful instances. These instances are then mapped to segmentation classes through an instance classifier (IC) network. To initiate rounds of the training of ROI and IC networks, we developed a fast, semi-automated annotation framework that leverages foundation models to annotate a subset of images using an object-level approach. In subsequent rounds, a human discriminator selects promising predictions from the last round, which are fed by unseen data, progressively enriching the model's training dataset for further fine-tuning of the networks. The networks are expanded from low-parameter to high-parameter models across rounds, incorporating a curriculum learning approach to capture increasingly complex features.

We evaluated PedVision on 552 hand X-ray images of children, retrieved from the publicly available Radiological Society of North America (RSNA) and Digital Hand Atlas (DHA) datasets, which represent a diverse range of ages and racial backgrounds. PedVision performed segmentation of 19 hand bones, grouped in five classes, and was compared against U-Net and DeepLabV3+ models using ResNet34 and ResNet101 backbones, as well as the SegFormer model with four different encoder variants. For pediatric cases (*i.e.*, 0–7 years), the PedVision pipeline outperforms the best-performing models, achieving an 11.08 % improvement in Dice score over U-Net in the RSNA dataset and a 7.68 % improvement in the DHA dataset. When compared to DeepLabV3+, the improvements are even more substantial, with gains of 14.43 % in RSNA and 14.78 % in DHA. Additionally, PedVision shows notable advantages over the best SegFormer model, with improvements of 8.16 % in RSNA and 1.91 % in DHA. The project is open source at github.com/mohofar/PedVision.

1. Introduction

The latest developments in deep learning (DL), particularly in visual foundation models (VFM) [1–4], have transformed the field of computer vision for medical image analysis. These powerful models excel at tasks involving medical images of adults due to the vast amount of available training data with fully mature anatomical structures. However, applying these models to pediatric medical images presents significant challenges due to the development of children's anatomy and

open growth plates. For example, the epiphyses of the phalanges and metacarpals typically complete the fusion between the ages of 13 and 16 for females and males [5]. This ongoing developmental process requires specialized algorithms that account for anatomical variations throughout this crucial growth period while coping with a limited number of samples per age group. This highlights the urgent need for robust and age-agnostic methods in pediatric image analysis.

Existing methods, including those presented in the references [6–8], have struggled with the inherent variability in pediatric images, such as

* Corresponding author at: Department of Biomechanical Engineering, Delft University of Technology (TU Delft), Delft, Netherlands (the).

E-mail address: m.homayounfar@tudelft.nl (M. Homayounfar).

¹ Authors contributed equally.

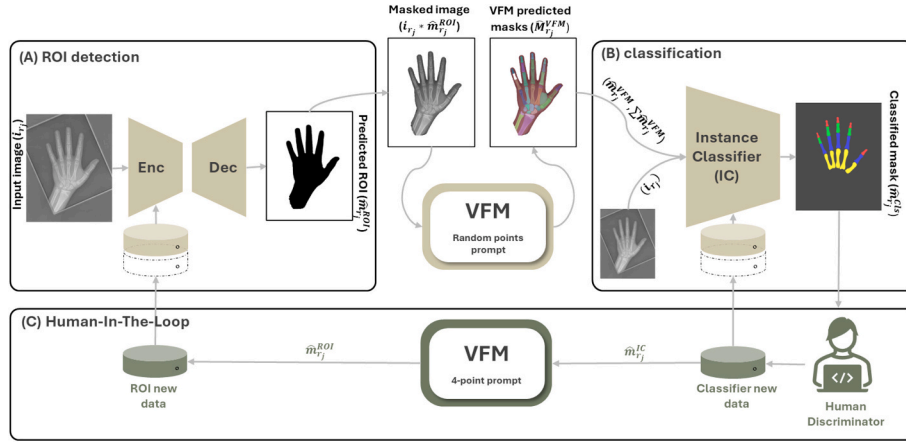


Fig. 1. Segmentation pipeline for all rounds. A. ROI detection part receives raw images and gives a binary mask to exclude irrelevant parts. B. The instance classifier uses raw images and VFM-predicted masks to classify each instance (i.e., bone). C. The human discrimination part selects the most accurate predictions to be included in the next round of training. This part is only used during the training process. To understand the mathematical notation, see Algorithm 1.

density variations and potential fractures. In this regard, Noguchi *et al.* [9] employed different data augmentation techniques, such as mixing random parts of cropped body images into a single image, to cope with this problem. However, the effectiveness of each augmentation strategy in capturing the specific variations and pathologies remains unclear [10]. While Liu *et al.* [11] demonstrated the benefits of leveraging large datasets and human expertise for training, this approach may not be feasible for pediatric medical image processing tasks due to data scarcity. This scarcity is due to the ionizing radiation associated with X-ray-based imaging modalities, such as planar X-rays and CT scans, and the difficulty of obtaining high-quality images at earlier ages when limiting the subject movements during the imaging process [12] is challenging.

While the studies mentioned above [6–11] highlight the limitations of DL models in pediatric musculoskeletal imaging, some other studies offer promising directions in solving pediatric challenges. For example, Boutillon *et al.* [13] presented a DL model for segmenting the ankle and shoulder, customized for pediatric magnetic resonance images (MRIs). This targeted approach acknowledges the unique challenges of pediatric anatomical growth. However, their work covers ages 5 to 17 years and lacks evaluation of younger age groups (under five years), a crucial age period of significant skeletal development. Additionally, they did not report age-based results to evaluate the effects of early bone growth on their segmentation task.

Another study by Boutillon *et al.* [14] segmented ankle, knee, and shoulder joints from pediatric MRIs of multi-domain datasets. While addressing diverse populations is valuable, their model struggled with unseen data from a new domain (e.g., a new modality). This highlights the ongoing challenge of developing generalizable models that can adapt to new datasets and imaging protocols. By addressing these challenges, DL has the potential to revolutionize pediatric musculoskeletal image analysis, leading to more accurate diagnoses and improved patient outcomes.

Unlike other areas that struggle with limited pediatric data, hand bone segmentation benefits from datasets covering various age groups, including newborns [15]. This rich data source facilitates the development of models specifically designed for the details of pediatric hand anatomy. Several studies highlight the effectiveness of DL in this domain. For example, Deshmukh *et al.* [16] employed a U-Net-based model for segmenting hand bones from X-ray images, achieving superior performance in age estimation compared to traditional segmentation techniques. Liu *et al.* [17] also utilized a U-Net model for segmenting hand bones from X-rays but identified potential issues with the appearance of rectangular irrelevant regions on pediatric images. Their proposed method addressed this specific challenge in pediatric hand bone segmentation. Du *et al.* [18] presented a two-step approach

involving bone detection and segmentation, achieving promising results through data pre-processing and model selection (OSA-YOLOv5 for detection and GRU-U-Net for segmentation). Ding *et al.* [19] focused on lightweight architectures designed explicitly for pediatric hand bone segmentation, demonstrating the potential of U-Net variations in this domain. Nagaraju *et al.* [20] explored the application of DeepLabV3+ [21] based models with different backbones for efficiently segmenting five hand bones across various age groups. Tay *et al.* [22] focused on different deep DL-based approaches for segmentation of pediatric hand phalanges that explicitly account for anatomical growth. Their work emphasizes the importance of growth-aware models and highlights DeepLabV3+'s suitability for handling anatomical variability in pediatric imaging.

Accurate segmentation is fundamental to bone age assessment of the hand, as it directly impacts the reliability of evaluating skeletal maturity in children by ensuring precise delineation of bone boundaries and structures. For instance, Jia *et al.* [23] developed a DL model using a recursive feature pyramid network, achieving precise segmentation of fine-grained bone structures and improving bone age assessment. Similarly, Spampinato *et al.* [24] employed a convolutional neural network with a BoNet architecture, reporting an improvement in the Dice score for hand bone segmentation, which in turn enhanced bone age assessment consistency across diverse age groups. However, challenges such as variations in X-ray quality and anatomical differences due to developmental stages can cause inaccurate boundary delineation and segmentation errors, as noted by Halabi *et al.* [15]. These studies highlight the significant progress in pediatric hand bone segmentation using DL.

Despite advancements in pediatric hand bone segmentation, current DL models encounter several challenges, including age dependency due to limited available training data, sensitivity to image quality (e.g., noise), and the need for manual segmentation of images from different age groups is a major bottleneck, as annotated datasets are often scarce or require substantial preparation. These factors hinder the generalization of DL models across diverse age groups and degrade segmentation performance.

This paper presents a novel segmentation pipeline (i.e., PedVision) that addresses critical challenges in medical image analysis, with a particular focus on pediatric cases characterized by limited training samples. Unlike conventional deep networks, our approach eliminates the need for manual annotation by leveraging a fast annotation framework for initialization and integrating ultralow-effort human feedback through an efficient approval/disapproval mechanism during training. By introducing object-level processing, the pipeline offers a new perspective on segmentation, achieving superior performance in

Table 1

Dataset characteristics. DHA dataset ages are presented as years, with starred cases indicating the minimum and maximum based on months to align with the RSNA dataset.

	No. samples	Manual segmentation	Age Min (month)	Age Max (month)	Age Mean (month)	Age SD (month)
Train RSNA-first round	88	No	6	228	111.23	47.83
Train RSNA-last round	407	No	6	228	108.95	50.12
Validation RSNA	13	No	10	168	88.92	65.24
Test RSNA-NSM group	200	Yes	169	228	186.59	10.28
Test RSNA-Pediatric group	200	Yes	4	84	50.16	22.95
Test DHA-total test	152	Yes	>0*	<228*	120	65.94
Test DHA-NSM group	40	Yes	>168*	<228*	204	17.18
Test DHA-Pediatric group	56	Yes	>0*	<84*	48	24.21

pediatric groups, significantly surpassing traditional models in data-scarce scenarios while maintaining robust segmentation across all age ranges. The design leverages the capabilities of VFMs to generalize across ages, mitigate the effects of noise and artifacts, and adapt effectively to pediatric imaging challenges. Its scalability, robustness, and adaptability position it as a transformative solution for pediatric medical imaging, ensuring consistent performance even in diverse and resource-constrained settings.

2. Methods

Fig. 1 illustrates the main components of the proposed pipeline and the data flow. The pipeline incorporates (1) a VFM, (2) a region of interest (ROI) network, (3) an instance classifier (IC) network, and (4) a human discriminator. The VFM converts pixel-level information into meaningful instances, such as bone segmentation masks, while the ROI and IC networks supervise the VFM inputs and outputs. Specifically, the ROI network focuses on regions containing only hand bones in images to prepare the input of VFM, while the IC network classifies the instances generated by the VFM into five classes, including four different hand bone types (*i.e.*, metacarpals and proximal, middle, and distal phalanges) and an irrelevant instances class. These two networks were trained and fine-tuned over multiple rounds using an initial dataset with limited annotated samples. Using the VFM at the pipeline's core eliminates the need for a rich dataset with annotated samples, reducing the demanding workload for human annotators. Instead, the human discriminator supports the enrichment of the training dataset for the ROI and IC networks by performing the ultralow effort task of approving or disapproving the segmentations predicted by the VFM for unseen images. In each round of training, we fine-tuned the ROI and IC networks and used the human discriminator to enrich the training dataset.

In the following subsections, we provide details of the dataset used and pipeline components, along with pseudocode, to clarify the steps followed. Additionally, we detail the experimental studies conducted to evaluate the pipeline's performance and compare it with different models.

2.1. Dataset

We used two publicly available datasets for our study. The first dataset, provided by the Radiological Society of North America (RSNA) for the 2017 Pediatric Bone Age Machine Learning Challenge [15], contains 12,611 hand X-ray images without bone segmentation masks. These images, taken from both male and female left or right hands, exhibit variability in size and resolution and were acquired from individuals aged 4 to 228 months, reflecting a wide range of growth variations in hand bones. We used this dataset to train our networks and test the performance of the proposed pipeline.

To evaluate the pipeline's ability to generalize in segmenting hand bones from X-rays acquired from different sources and race groups, we used a second dataset, the Digital Hand Atlas (DHA)[25]. The DHA

dataset comprises 1,390 digitized left-hand radiographs from children aged 0 to 18 years, evenly distributed across four race groups (Caucasian, Asian, African American, and Hispanic) and both genders. Each image is annotated with demographic information and bone age assessments provided by pediatric radiologists.

2.2. Data preparation

PedVision efficiently annotates data to initiate ROI and IC model training in the first round of training. In subsequent rounds, it self-annotates new samples. The following details provide more information about the first round.

For the first-round training (r_1) and validation of the ROI and IC networks, we retrieved 101 radiographs with arbitrary image sizes from the RSNA database. These radiographs were manually selected to capture a range of image resolutions, artifacts (*e.g.*, irrelevant objects), and ages, ensuring diverse representation in the dataset. The radiographs were divided into training (88 samples) and validation (13 samples) sets. The validation set is used solely to save the best model during training.

Training the ROI and IC networks requires paired images (I) and masks (M). For the first round of training (r_1), two sets of paired images and masks, one for the ROI network ($I_{r_1}, M_{r_1}^{ROI}$) and one for the IC network ($I_{r_1}, M_{r_1}^{IC}$), were generated using the VFM-based annotation frameworks that we developed. In the ROI annotation framework, each image was paired with a selected set of 2D point coordinates (x, y) as the VFM prompt, which was used as input of the VFM to predict the hand region ($m_{r_1}^{ROI}$). Following this step, the VFM network used 32^2 random number of points coordinates on the masked image ($i_{r_1} * \hat{m}_{r_1}^{ROI}$) to predict a set of mask instances ($\hat{M}_{r_1}^{VFM}$). We used the highest predicted mask score as the hand region. Using the IC annotation framework, each of the VFM-predicted masks was assigned to one of the 5 classes (*i.e.*, $\hat{m}_{r_1}^{IC}$) corresponding to (i) the distal phalanges, (ii) the intermediate phalanges, (iii) the proximal phalanges, (iv) the metacarpals, and (v) regions other than related hand bones. The IC and ROI annotation frameworks were employed only in the first training round. In the subsequent training rounds, the ROI and IC masks for the unlabeled images ($M_{r_{j>1}}^{ROI}, M_{r_{j>1}}^{IC}$) were automatically generated by the pipeline.

The images used to train the ROI and IC networks were resized to 256×256 . However, since the VFM model operates with images of arbitrary sizes, resizing raw images to a constant size may lead to information loss. Thus, we used the results from the IC network to select the relevant masks and create the final mask based on each image's original size and resolution.

To test the performance of the pipeline, we randomly retrieved 400 radiographs from the RSNA database, ensuring that half of them were from individuals up to 7 years of age (hereafter referred to as the Pediatric group) and the other half from individuals aged between 14 years and 21 years (hereafter referred to as near-skeletally-mature, NSM, group). This was done to study the age dependency of the proposed

Inputs:

I_{r_1} : Input of images of the first round
 I_u : Unlabeled samples
 $M_{r_1}^{ROI} \leftarrow$ ROI annotation framework
 $M_{r_1}^{IC} \leftarrow$ IC annotation framework
 $(I_{r_1}, (M_{r_1}^{ROI}, M_{r_1}^{IC}))$ # Inputs and masks of the first round

Procedure:

Initial round
 $\hat{M}_{r_1}^{ROI} \leftarrow$ Train ROI($I_{r_1}, M_{r_1}^{ROI}$)
 $\hat{M}_{r_1}^{VFM} \leftarrow$ VFM($\hat{M}_{r_1}^{ROI} * I_{r_1}$) # Get all masks from the foundation model
 $\hat{M}_{r_1}^{IC} \leftarrow$ Train IC($(\hat{M}_{r_1}^{VFM}, \sum \hat{M}_{r_1}^{VFM}, I_{r_1}), M_{r_1}^{IC}$)

Next rounds
for r_j **in** Rounds:
 $I_{u_s} \leftarrow$ subset(I_u)
 for $i_{u_s} \in I_{u_s}$:
 # Prediction of unlabeled samples
 $\hat{m}_{u_s}^{ROI} \leftarrow$ Predict ROI(i_{u_s})
 $\hat{M}_{u_s}^{VFM} \leftarrow$ VFM($\hat{m}_{u_s}^{ROI} * i_{u_s}$)
 $\hat{m}_{u_s}^{IC} \leftarrow$ Predict IC($(\hat{M}_{u_s}^{VFM}, \sum \hat{M}_{u_s}^{VFM}, i_{u_s})$)
 if $\hat{m}_{u_s}^{IC}$ **is correct**: # human discriminator
 $I_{r_j} \leftarrow I_{r_{j-1}} \cup i_{u_s}$
 $M_{r_j} \leftarrow M_{r_{j-1}} \cup \hat{m}_{u_s}^{IC}$
 else:
 continue
 end if
 end for
 $\hat{M}_{r_j}^{ROI} \leftarrow$ Finetune ROI($I_{r_j}, M_{r_j}^{ROI}$)
 $\hat{M}_{r_j}^{VFM} \leftarrow$ VFM($\hat{M}_{r_j}^{ROI} * I_{r_j}$)
 $\hat{M}_{r_j}^{IC} \leftarrow$ Finetune IC($(\hat{M}_{r_j}^{VFM}, \sum \hat{M}_{r_j}^{VFM}, I_{r_j}), M_{r_j}^{IC}$)
end for

pipeline. To test the generalizability of the pipeline, we used the DHA dataset, randomly selecting a test sample from each combination of age, gender, and race to cover the full diversity of hand radiographs. With 19 age groups (0–18 years, in one-year increments), two genders (male and female), and four racial categories (Caucasian, Asian, African American, and Hispanic), we utilized a total of 152 images from the DHA dataset.

All the test images were resized to 1024×1024 pixels to ensure high-quality images, and all of them were manually segmented using the open-source software LabelMe [26]. These manual segmentations served as the gold standard for evaluating the segmentations predicted by the trained models. The details of the training, validation, and test datasets are outlined in Table 1.

2.3. Networks and algorithms

2.3.1. ROI network

To train the ROI network, we used PyTorch segmentation models [27] with the Efficientnet-B0 [28] backbone as the encoder, pre-trained on the ImageNet dataset. We utilized a U-Net-based architecture [29] to complete the ROI network's decoder part. Over the training rounds, the ROI model was fine-tuned using the pairs of images and ROI masks ($I_{r_1}, M_{r_1}^{ROI}$) obtained as explained in Section 2.2.

To enhance the generalization of the ROI network, various augmentation techniques were applied during each training epoch, with

a 50 % probability of replacing actual data in each training epoch. These techniques included random horizontal flipping, zooming by a random factor between 0.4 and 1.2 times the original size, translation of up to 64 pixels in both translational directions, rotation by a random degree between -90 and 90° , and brightness adjustment by a random factor ranging from 0.3 to 1.5 times of the original image intensity. Additionally, random rectangular shapes of varying sizes and transparency levels (ranging between 0 and 0.8) were superimposed on the original images.

The ROI network's weights were updated using Dice-based loss and the Adam optimizer, which has a learning rate of 1×10^{-4} . The validation set was used to identify the best-performing fine-tuned ROI network during each training epoch. The model with the highest validation Dice score over 50 epochs in each training round was saved for future use.

2.3.2. VFM

Segment Anything Model (SAM) [1], trained on an extensive dataset of eleven million images and over one billion masks, was used as the VFM model. We employed the huge model weights (*i.e.*, ViT-H) with 32^2 points over images, an IoU prediction threshold of 0.9, and a stability score threshold of 0.9. All other parameters of the VFM were kept at the default settings of the SAM model [1]. The default settings were preserved to achieve a balance between performance and computational

feasibility. Modifying VFM parameters requires retraining the IC network to ensure alignment with VFM for each parameter adjustment. This process is time-consuming and yields only minor performance variations. In the training of the pipeline, we used masked images ($\hat{m}_{r_j}^{ROI * i_r}$) to feed the VFM and obtain different instances of each ROI-masked image as VFM masks (\hat{M}_r^{VFM}).

2.3.3. IC network

In each training round, after completing the training of the ROI network, we froze its weights to pass the data through it and prepare the data for training the IC network. The IC network used pre-trained models (e.g., MobileNet_V2) with three input channels. We used input images (I_r), an instance of predicted masks by VFM (\hat{M}_r^{VFM}), and summation of all predicted masks ($\sum \hat{M}_r^{VFM}$) as three channels of the IC network. During each training epoch, augmentation techniques were applied to the images with a 50 % probability of replacement with actual images. These techniques included random horizontal flipping, zooming by a random factor between 0.4 and 1.2 times the original size, random translation up to 128 pixels in both translational directions and random rotation by an angle between -20° and 20° . We utilized a curriculum learning approach to overcome overfitting, which a limited number of training samples can cause. This was achieved by employing various pre-trained networks, ranging from low-parameter to high-parameter networks, across different training rounds. The networks used included MobileNet_V2 [30], Efficientnet-B1 model [28], and EfficientNet-B5 [28]. When transitioning from one model (e.g., MobileNet_V2) to the next (e.g., Efficientnet-B1), we initiated training for the next model using ImageNet-based weights.

We used cross-entropy loss and the Adam optimizer with a learning rate of 1×10^{-3} . At each training round of 50 epochs, the best-performing model was selected based on the highest average Dice

score across all classes using the validation set.

2.3.4. Human-in-the-loop

After each training round, which involved training the ROI network, getting the masks from VFM, and training the IC network, we froze all the networks. Then, we used a random subset of unseen and unlabeled images (I_{us}) to predict segmentation masks (\hat{m}_{us}^{IC}). The role of the human discriminator was to identify cases where the segmentation masks were predicted correctly. This task involved accepting or rejecting the predicted masks, which is an ultralow effort task in contrast to manual annotation. After providing an arbitrary number of promising segmentation cases, the pipeline added them to the training set of the IC network. For the ROI network, we chose four random points within the bone areas predicted by the IC network as the prompt of the VFM (Fig. 1. C). The VFM results with the highest probability were saved as the ROI masks (i.e., hand region) in the ROI training set. After all this procedure, the pipeline was ready for the next round of training, which included fine-tuning ROI and IC networks, followed by the human discriminator to add more promising samples to the training set from new unseen and unlabeled data.

An overview of the pipeline and training procedure is presented in Algorithm 1. For further clarification of the pseudocode, we refer readers to the [Supplementary material](#).

2.4. Performance evaluation of the pipeline

The performance of the proposed pipeline was compared to several architectures: DeepLabV3+ [21] and U-Net [29] using ResNet34 [31] and ResNet101 [31] encoders, and SegFormer [32] with mit-b0, b1, b2, and b3 encoders. Given that the primary difference between the hand radiographs of the Pediatric and NSM groups lies in nonlinear transformations of the bones (e.g., variations in shape and size), DeepLabV3+ is considered one of the most suitable options for performance comparison [22]. We considered U-Net as a standard convolutional architecture. We also included the recently developed SegFormer models, which are state-of-the-art transformer-based architectures that have demonstrated superior performance across various segmentation tasks [32,33].

These models were evaluated using Dice, IoU, precision, and Hausdorff distance (HD) scores. The average for each class was computed using the previously mentioned metrics. Mean and 95 % confidence interval (95 % CI) values were then calculated by averaging across all classes, excluding the background class. Confidence intervals were derived using the standard error of the mean and the corresponding critical value from the z-distribution. The margin defines the range around the meaning within which the true value is expected to lie with 95 % confidence, expressed as $mean \pm margin$. While Dice and IoU capture overall overlap, precision explicitly reflects the model's ability to avoid false positives, ensuring that only truly relevant pixels are included, which is crucial in clinical tasks such as avoiding over-treatment areas. HD complements these by quantifying the worst-case boundary error, highlighting cases where the model fails to capture fine anatomical details. We used the directed HD from the SciPy library in Python, whereas high HD values particularly in small pediatric structures, often suggest critical margin deviations due to anatomical complexity or image contrast variability.

All the models were trained using training images and masks (i.e., images of 1024×1024 pixels) employed in training the ROI and IC networks during the first training round. However, cases with missing segmentation parts in a mask (e.g., a part of bone missed by VFM) were excluded. This exclusion was necessary because label inconsistencies, such as missing parts, could negatively impact the performance of the compared models [34]. We trained all the models for 50 epochs, saving the best-performing model based on the highest Dice score on the

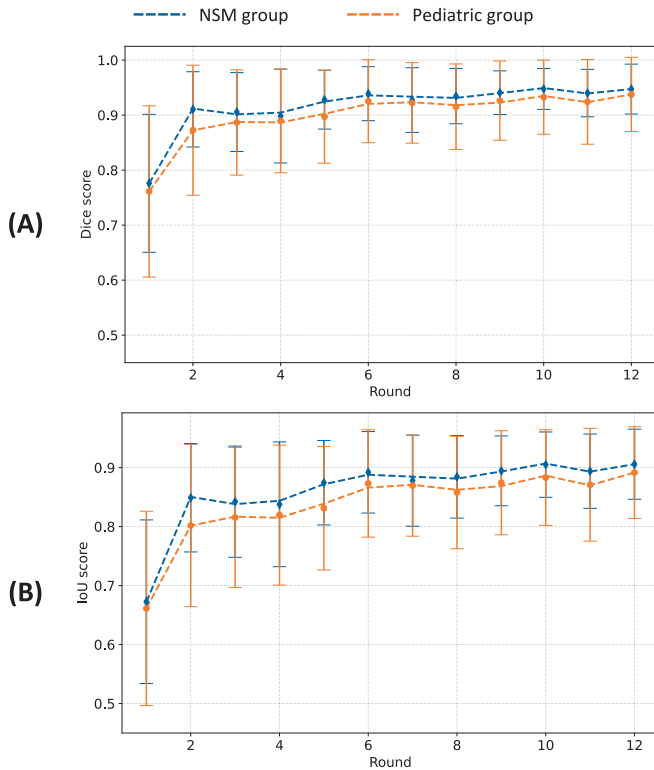


Fig. 2. Test set performance (mean \pm SD) during training for the Pediatric and NSM groups using RSNA dataset over rounds of training. A. Dice score. B. IoU score.

Table 2

Segmentation comparison on 200 NSM group and 200 Pediatric group images. The 95 % confidence interval (CI 95 %) margins are presented in parentheses alongside all metrics.

Models	RSNA dataset							
	Pediatric group				NSM group			
	Dice	IoU	Precision	HD	Dice	IoU	Precision	HD
U-Net (Res34)	83.05 (2.40)	74.84 (2.72)	78.69 (1.97)	97.22 (14.08)	94.59 (0.43)	90.00 (0.68)	91.70 (0.57)	87.85 (1.05)
U-Net (Res101)	74.64 (2.64)	64.19 (2.94)	72.40 (2.09)	–	87.85 (1.05)	79.96 (1.38)	85.77 (0.47)	106.55 (9.67)
DeepLapV3+ (Res34)	77.09 (1.77)	65.05 (1.98)	66.97 (1.68)	–	88.54 (0.37)	79.86 (0.56)	80.48 (0.48)	–
DeepLapV3+ (Res101)	79.87 (1.95)	68.98 (2.19)	73.75 (1.53)	–	90.78 (0.41)	83.46 (0.64)	84.68 (0.56)	24.50 (3.69)
SegFormer (mit_b0)	75.89 (2.09)	64.18 (2.28)	71.08 (1.71)	–	87.95 (0.52)	78.96 (0.79)	80.77 (0.69)	153.13 (12.21)
SegFormer (mit_b1)	84.04 (1.61)	74.61 (1.95)	80.07 (1.35)	–	90.94 (0.47)	83.75 (0.75)	86.08 (0.65)	108.58 (11.27)
SegFormer (mit_b2)	85.97 (1.41)	76.96 (1.82)	80.18 (1.38)	177.23 (11.18)	91.91 (0.44)	85.30 (0.71)	87.17 (0.63)	135.89 (9.36)
SegFormer (mit_b3)	83.86 (1.84)	74.59 (2.27)	77.58 (1.93)	–	92.79 (0.44)	86.83 (0.71)	87.91 (0.68)	93.85 (7.65)
PedVision	94.13 (0.54)	89.46 (0.78)	94.80 (0.58)	18.40 (3.31)	94.80 (0.50)	90.56 (0.75)	96.45 (0.30)	27.75 (4.11)

models	DHA dataset							
	Pediatric group				NSM group			
	Dice	IoU	Precision	HD	Dice	IoU	Precision	HD
U-Net (Res34)	87.57 (4.59)	81.71 (5.01)	88.99 (3.02)	–	91.58 (1.84)	85.34 (2.37)	91.21 (0.67)	86.34 (22.09)
U-Net (Res101)	72.75 (6.13)	63.12 (5.97)	76.46 (3.97)	–	80.18 (3.41)	69.47 (3.75)	80.58 (0.99)	144.85 (22.65)
DeepLapV3+ (Res34)	79.66 (5.32)	70.39 (5.20)	75.63 (4.08)	–	86.47 (2.56)	77.40 (2.85)	81.15 (1.54)	66.47 (10.76)
DeepLapV3+ (Res101)	80.47 (5.64)	72.06 (5.71)	81.18 (4.20)	–	88.30 (2.34)	80.32 (2.82)	86.77 (0.73)	59.78 (15.34)
SegFormer (mit_b0)	90.09 (2.46)	83.27 (3.13)	87.68 (0.88)	60.46 (14.04)	91.72 (0.64)	84.91 (1.01)	88.19 (0.62)	78.95 (13.57)
SegFormer (mit_b1)	90.87 (2.01)	84.25 (2.58)	88.74 (0.85)	60.94 (13.62)	91.05 (0.76)	83.94 (1.13)	88.11 (0.56)	87.05 (14.62)
SegFormer (mit_b2)	92.49 (1.49)	86.74 (2.10)	90.52 (0.71)	42.96 (13.84)	92.93 (0.56)	87.02 (0.87)	89.89 (0.52)	90.95 (13.60)
SegFormer (mit_b3)	93.34 (0.99)	87.87 (1.52)	89.79 (0.97)	30.01 (11.88)	93.86 (0.33)	88.54 (0.57)	90.08 (0.48)	65.47 (14.24)
PedVision	95.25 (0.46)	91.17 (0.78)	95.94 (0.49)	33.14 (10.08)	94.98 (0.81)	90.74 (1.28)	96.01 (1.12)	52.33 (11.17)

validation set. Finally, all models were evaluated using the same test samples from the RSNA and DHA datasets. A more detailed comparison of the PedVision pipeline with the other models was conducted on the DHA dataset, with results stratified by age, race, and gender. This analysis enabled us to assess the influence of demographic factors on the performance in segmenting hand bones, as detailed in the Results section.

In addition, we performed four experiments, detailed in the Ablation study presented in the Results section, to assess how the inclusion of the ROI model and variations in the VFM parameters affect the performance of the proposed pipeline in segmenting hand bones from X-rays in the RSNA dataset. In the first experiment, we varied the number of points used in the VFM (*i.e.*, 16^2 and 64^2 , as well as 32^2 points, as the default setting) without fine-tuning the networks in the pipeline. The second experiment involved replacing the ViT-H model with smaller models, including the ViT-L and ViT-B models, which have fewer parameters. In the third experiment, we studied the effects of removing the ROI model entirely from the pipeline.

In addition, the quality of segmentations produced by the PedVision was systematically analyzed to gain a detailed understanding of its performance. Segmentation inaccuracies were categorized into four types: 1) extra objects (predicted regions that do not correspond to any target bone), 2) misclassifications (a bone incorrectly labeled as another), 3) missing parts (one or more bones not detected), and 4)

minor inaccuracies.

3. Results

3.1. Progression of PedVision performance over training rounds

To show how the model improves its segmentation performance over time, we present Fig. 2, which shows the pipeline's progress across 12 training rounds using the RSNA dataset. As additional training samples were introduced in each round, the Dice score for the NSM group improved from 0.78 in the first round to 0.95 in the final round, while the Pediatric group saw an increase from 0.76 to 0.94. Similarly, the IoU score for the Pediatric group increased from 0.66 to 0.89, and for the NSM group, it rose from 0.67 to 0.91. These performance gains were also accompanied by a reduction in standard deviation (SD) as shown in (Fig. 2), indicating more consistent and stable model predictions over time.

3.2. Evaluation of PedVision and benchmark models

3.2.1. Age-based comparison

For the RSNA dataset, PedVision demonstrated generalization across different age groups (Table 2). PedVision achieved a Dice score of 94.13 % (95 % CI margin: 0.54) for the Pediatric group and 94.80 % (95 % CI

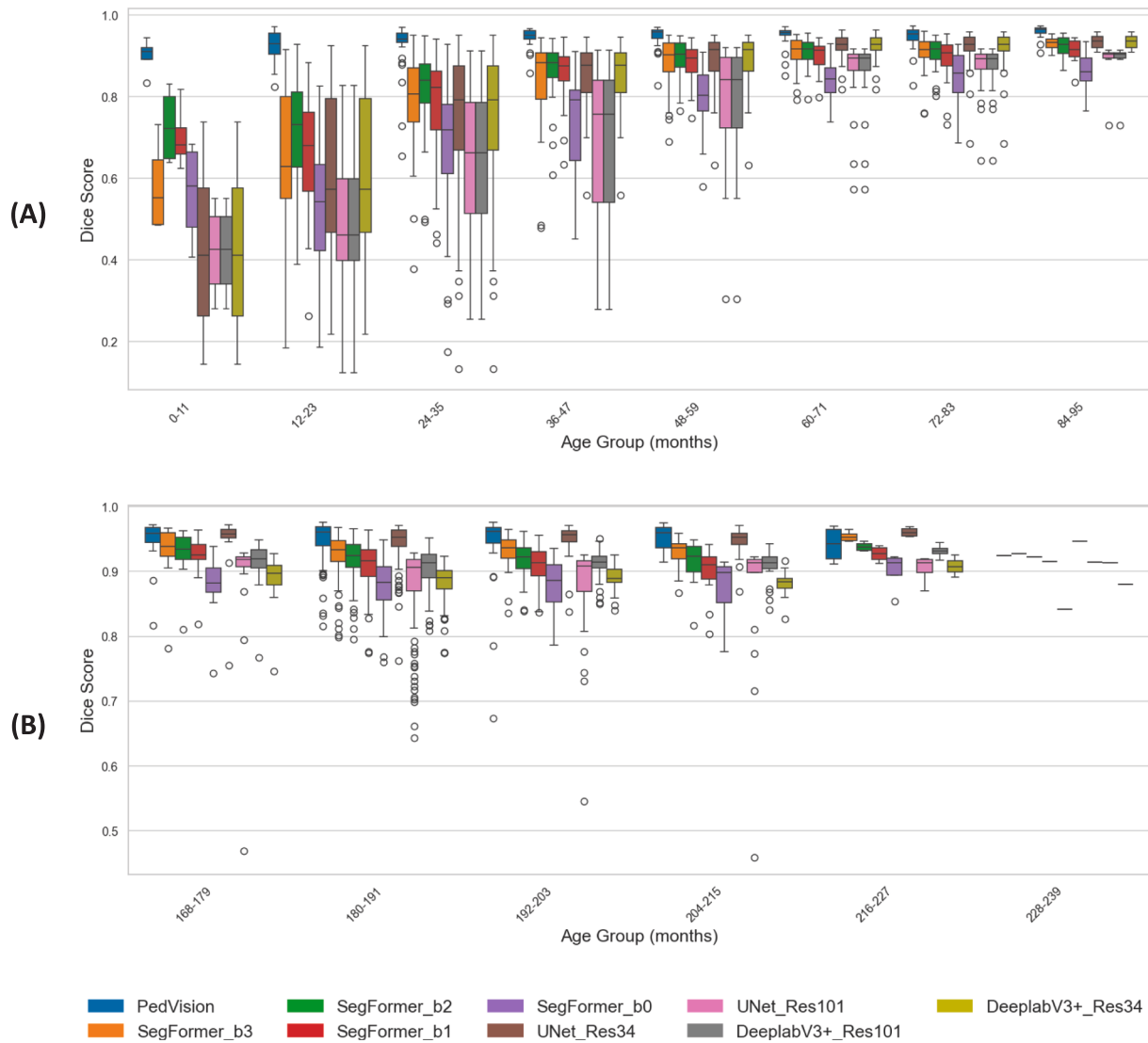


Fig. 3. Age-based comparison of the dice scores for the RSNA dataset. A. Pediatric group (the last age group, 84–96, only includes 84-month cases). B. NSM group (the last age group, 228–239, includes one sample that is 228 months).

margin: 0.50) for the NSM group. Other models did not achieve similarly high scores or consistent results across both the Pediatric and NSM groups. For example, the U-Net model with a ResNet34 encoder achieved a Dice score of 94.59 % (95 % CI margin: 0.43) for the NSM group but experienced an 11.54 % reduction in performance for the Pediatric group, resulting in a Dice score of 83.05 % (95 % CI margin: 2.40). Fig. 3 shows the Dice score and its stability for all models in more detail. This figure indicates that the Dice scores achieved with PedVision are higher and more stable for both the Pediatric and NSM groups. In contrast, other models exhibit this stability only for the NSM group.

On the DHA dataset, PedVision demonstrated consistent performance, similar to what was observed on the RSNA dataset. In contrast to the RSNA results, the stability of the benchmark models showed a relative improvement (Table 2). The SegFormer (mit b3) model, in particular, achieved results comparable to those of PedVision. With a Dice score of 95.25 % (95 % CI margin: 0.46), PedVision exhibited a 1.91 % higher Dice score compared to the SegFormer (mit-b3) model for the Pediatric group. The difference for the NSM group was a 1.12 % improvement in PedVision score. A more detailed breakdown of Dice scores across smaller age subgroups is provided in Fig. 4.C, which shows that PedVision outperformed other models in almost all ages. The larger Dice score gaps were observed in individuals younger than 3 years,

while for older subjects, some models, such as SegFormer, showed comparable results to PedVision (Fig. 4C).

Fig. 5 depicts the predictions made by PedVision and the best-performing benchmark model (i.e., SegFormer with the mit-b3 backbone) for both Pediatric and NSM group cases of the DHA dataset. The visualizations show that our pipeline achieved more precise bone boundary delineation, fewer inter-class misclassifications, and less missing bone coverage compared to SegFormer models. Additional segmentation results can be found in Fig. S3 of the Supplementary material for readers interested in further visual comparisons.

3.2.2. Gender and race-based comparison

Fig. 4 compares various models on the DHA dataset across races (Fig. 4.A) and genders (Fig. 4.B). Fig. 4.A shows minimal racial differences, with the African American (BLK) group deviating slightly more than other models. The SD for this group is 0.028. BLK male group show larger SD than females with an SD of 0.035, while all other groups show an SD of 0.018. The U-Net and DeeplabV3+ models exhibit more deviations compared to PedVision and SegFormer (Fig. 4). For other races, model results are generally similar, with PedVision consistently achieving a higher average Dice score. Fig. 4.B shows that gender-based comparison showing that all models have similar results for both male

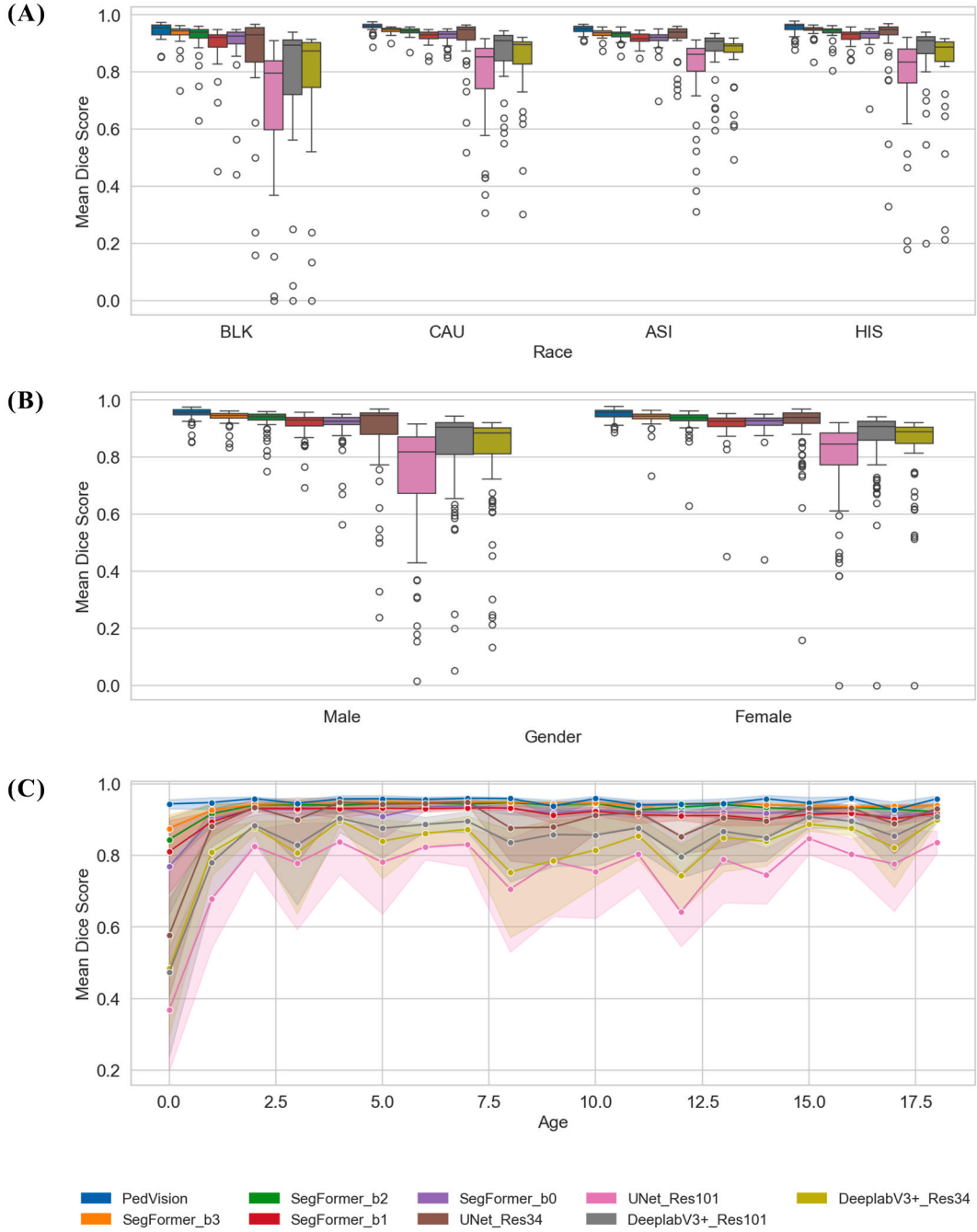


Fig. 4. Comparative results of (A) Race, (B) Gender, and (C) Age, for the DHA dataset.

and female.

3.3. Ablation study

The results mentioned in the previous subsections were achieved using the default settings of SAM model, *i.e.*, 32^2 points for the huge version of the SAM (ViT-H), as indicated in [1]. In the first experiment presented in Table 3, we varied the number of points in the VFM to 16^2 and 64^2 without fine-tuning networks in the pipeline. The variation in the number of points did not noticeably influence the outcome of the NSM group cases. However, there was a 3.7 % decrease in the Dice score for the Pediatric group cases (Table 3).

In the second experiment, as shown in Table 3 with 32^2 -point SAM, we replaced the largest SAM-based model (ViT-H) with other versions (*i.e.*, ViT-L and ViT-B). This change had a more significant impact on the

NSM group when using the smallest model (ViT-B) compared to the other model (ViT-L), which has fewer parameters than ViT-H but is larger than ViT-B.

The third experiment, presented in Table 3, explored the effects of the ROI model by removing it from the pipeline. The most significant impact was observed in the Pediatric group population, with a 5 % reduction in the Dice score, whereas this reduction in the NSM group was less than 2 %. At the same time, there was a maximum of 4 % increase in the SD of the results.

In the final experiment, we used the final-round training set generated by our pipeline to train the SegFormer B3 model, which outperformed the other benchmark models included in the study. Fig. S1 in the Supplementary material displays the number of samples added for each age subgroup during the 12 training rounds of PedVision. Fig. S2 presents the Dice scores achieved for each age subgroup using both the

















Age (y)	Input image	Ground Truth	PedVision	SegFormer (mit-B3)
NSM group				
ID: 5291, Age: 18				
Pediatric group				
ID: 4706, Age: 0				
ID: 4737, Age: 1				
ID: 7025, Age: 5				

Fig. 5. Visualization of the NSM and Pediatric group cases for the DHA dataset.

PedVision and SegFormer B3-based models on the RSNA and DHA datasets.

3.4. Evaluation of PedVision's segmentation outcomes

To improve the transparency of the segmentation error analysis, we have summarized the common failure types observed in PedVision's outputs in Table 4. In 67 % of the test cases, the segmentations were highly accurate, with only minor boundary discrepancies that may

reflect imperfect ground truth annotations or human error. Among the remaining cases, the most frequent segmentation issue was the missing of bone parts, affecting 17 % of the samples. Extra objects and misclassifications were each observed in 5 % of cases, while minor inaccuracies appeared in 6 %. Examples of these four types of segmentation errors are illustrated in Fig. 6.

Table 3

Effect of changing the number of points, core model, and usage of ROI model in the pipeline.

Experiments	VFM Points	VFM type	ROI model	Dice mean	Dice SD	IoU mean	IoU SD	Dice mean	Dice SD	IoU mean	IoU SD
				Pediatric group				NSM group			
Default	32	ViT-H		0.9374	0.0664	0.8907	0.0773	0.9484	0.0355	0.9062	0.0533
1	16	ViT-H		0.9006	0.0753	0.8388	0.0879	0.9338	0.0480	0.8841	0.0670
	64	ViT-H		0.9366	0.0664	0.8899	0.0771	0.9459	0.0360	0.9023	0.0539
2	32	ViT-L		0.9316	0.0702	0.8827	0.0828	0.9280	0.0587	0.8754	0.0800
	32	ViT-B		0.9282	0.0742	0.8778	0.0844	0.8915	0.1029	0.8251	0.1296
3	32	ViT-H	—	0.8826	0.0893	0.8168	0.1041	0.9317	0.0479	0.8804	0.0639

Table 4

Categorization and frequency of segmentation errors in PedVision.

Error type	Description	Frequency (%)	Error part in the pipeline (possible causes)
Extra Objects	Predicted areas that do not correspond to any anatomical bone	5	IC network error between background class and bone classes (underfitting [*])
Misclassifications	Bone segments labeled as the wrong class	5	IC network error between bone classes (underfitting [*])
Missing Parts	Absence of one or more bone regions	17	VFM error (VFM config or unclear bone boundaries ^{**})
Minor Inaccuracies	Slight boundary deviations or labeling mismatches	6	VFM error (VFM config or unclear bone boundaries ^{**})
Highly Accurate	Minimal or no visible errors	67	—

^{*} Underfitting may result from using too few samples in training.^{**} VFM parameter configuration may change the likelihood of identifying bone regions depending on image quality and prompt points.

3.5. Computational time

The time required for first-round annotations, which were semi-automated, depends on the computational resources used (e.g., GPU speed), image size, and the number of classes to segment. In our case, experiments were conducted using a Nvidia GeForce RTX 3080 Ti Laptop GPU with 16 GB RAM. In the first round of the ROI annotation, the average time required was 18 s per image, including the VFM prediction and user interaction time, based on a sample of 10 random images. For the classifier annotation in the first round, the average time was 75 s per image for identifying the classes of each predicted mask, excluding VFM prediction time, which was performed once for all samples before user interaction began. The human-in-the-loop strategy, including VFM prediction time, averaged 20 s per image of varying size. The inference time was like the human-in-the-loop strategy time, averaging 20 s per image. After the first round, there was no need to spend the mentioned annotation time, as all processes were completed automatically within the training.

From a model complexity standpoint, the number of trainable parameters varies substantially across all compared models. U-Net variants include approximately 24.4 M (Res34) and 51.5 M (Res101) parameters, while DeepLabV3+ has 22.4 M (Res34) and 45.7 M (Res101). SegFormer models range from 3.7 M (mit-b0) to 44.6 M (mit-b3). In contrast, PedVision includes a large VFM with 641 M parameters, which is used in a frozen state and thus not trained. Only its lightweight components, including the ROI model (6.2 M) and the IC network (28.3 M), are optimized during training.

4. Discussion

The presented results demonstrate that our proposed pipeline, PedVision, offers a robust and generalizable solution for pediatric medical image segmentation, addressing critical challenges related to anatomical variability across different age groups and the limited number of annotated samples. While previous models, such as those developed by Boutillon *et al.* [13,14] for ankle, shoulder, and knee, are not doing the same experiments as ours, they have successfully segmented pediatric images for children aged 5–17. However, their work does not fully address younger age groups, particularly infants under 12 months, where skeletal structures are rapidly developing. In contrast, PedVision demonstrated high and consistent Dice scores across all age groups (Fig. 3 for RSNA dataset and Fig. 4.C for DHA dataset). Similarly, U-Net-based models, such as those proposed by Deshmukh *et al.* [16] and Liu *et al.* [17] for hand X-ray images, have achieved significant progress in pediatric segmentation yet faced challenges with age-dependent variability and irrelevant objects in images. By incorporating an ROI model, our approach has been able to mitigate these issues and improve segmentation accuracy.

The results of this study highlight the effectiveness and stability of our proposed pipeline in segmenting pediatric medical images across diverse age groups as compared to the U-Net, DeepLabV3+ and SegFormer models with different encoder backbones. PedVision consistently demonstrated high Dice scores across all ages, particularly in younger children, where all compared models struggle. For instance, as shown in Fig. 3.A, PedVision achieved a Dice score of 0.90 for age groups under 11 months, whereas the other models performed notably worse. This level of stability, even with limited training samples (e.g., based on Fig. S1.B, only four samples were provided for ages less than ~11 months) in these subgroups, is a significant achievement, as pediatric image segmentation often suffers from data scarcity.

Table 2 presents a comprehensive comparison of segmentation performance across Pediatric and NSM groups on the RSNA and DHA datasets. While numerical metrics and their associated 95 % confidence intervals (CI) quantify performance, several clear trends and outliers merit discussion.

Across both datasets (Table 2), PedVision consistently achieved superior performance with narrow CIs across Dice, IoU, and Precision, indicating both high accuracy and model stability. For example, in the RSNA Pediatric group, PedVision achieved a Dice score of 94.13 (CI \pm 0.54), significantly outperforming the next best model, SegFormer (mit-b2), which had a Dice score of 85.97 (CI \pm 1.41). The observed performance gap, together with PedVision's lower HD, indicates improved boundary delineation and reduced segmentation variability, which result from its VFM-based design and training strategies aimed at generalization (e.g., across pediatric anatomies).

When comparing the performance of the pediatric and NSM groups, several models demonstrated higher accuracy for the NSM group, especially in terms of Dice and Precision (e.g., U-Net Res34: Dice 83.05 in the pediatric group vs. 94.59 in the NSM group, RSNA dataset). This discrepancy may reflect increased anatomical variability and smaller organ sizes in pediatric scans, which can challenge general-purpose models trained primarily on adult data. PedVision, by contrast,






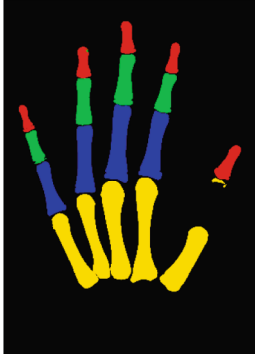


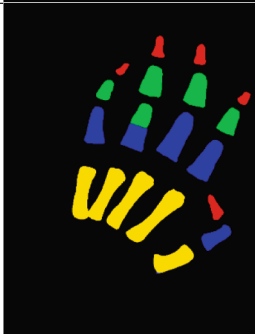



Mistake s	Input image	Ground Truth	PedVision
Minor inaccuracy			
Missing parts			
Misclassification			
Extra objects			

Fig. 6. Different types of segmentation errors.

maintained nearly identical performance between Pediatric and NSM groups (Dice 94.13 vs 94.80), with similarly tight CIs, suggesting better generalizability across age groups.

The confidence intervals also help reveal the consistency of the model. Narrow CIs (e.g., PedVision’s Precision of 96.45 ± 0.30 in the RSNA NSM group) imply stable performance across test samples. In contrast, wider CIs (e.g., U-Net Res101 Dice 72.75 ± 6.13 on DHA

Pediatric) suggest greater sensitivity to sample variability. This instability is especially prominent in models like U-Net Res101 and DeepLabV3+, likely due to underfitting resulting from a limited number of trained samples or a lack of capacity to capture the diverse anatomical structures in pediatric imaging.

Notably, outliers in HD values were observed, particularly in SegFormer (mit_b2) on the RSNA Pediatric data (HD: 177.23 ± 11.18), and

U-Net Res101 on the DHA NSM data (HD: 144.85 ± 22.65). These abnormally high HD values imply extreme boundary errors in a subset of test cases. These outliers may stem from failure to segment small or low-contrast structures, especially in pediatric images where organ boundaries are less distinct. In contrast, PedVision achieved consistently low HD across datasets, again highlighting its robustness.

Fig. 4A shows that the BLK group, representing African American individuals, contributes to increased SD in Dice scores across all models. This aligns with studies [35–37] indicating differences in bone growth between African Americans and others. This may be attributed to biological growth variations within this group, which can impact segmentation performance.

Fig. 4B displays results stratified by gender, demonstrating no major differences in performance between male and female subjects for the compared models. This suggests that the anatomical variations between male and female hands are accurately captured by all models.

Fig. 4C highlights the superior performance of PedVision on the DHA dataset, particularly for ages under 3 years, where it outperforms other models. While some models exhibit stable performance in certain age groups, their results are inconsistent across the entire age range, limiting their reliability. For instance, SegFormer-based models achieve Dice scores that are 2–5 % lower than PedVision for individuals above 3 years of age (Table 2) but show significantly larger performance gaps for individuals under 3 years.

Introducing the ROI model in PedVision was crucial for achieving this stability and removal of unnecessary elements, directing the VFM to concentrate on the primary segments of the image rather than the entire image. Utilizing ROI increases the number of relevant predicted masks generated by VFM. Consequently, this creates a more balanced training set for the IC network, with the background class (majority class) having a relatively lower number of samples compared to scenarios where ROI is not employed.

The Dice score reduction observed when the ROI model was excluded, particularly in the Pediatric group cases (Table 3, Experiment 3), underscores the importance of this step in creating an age-agnostic segmentation pipeline. This observation aligns with similar studies, such as those by Boutillon *et al.* [14], which emphasize the challenges in pediatric segmentation when the entire image is processed without focusing on the most relevant areas.

Furthermore, the IC network in PedVision enhanced segmentation accuracy by utilizing a three-channel input, which combines real image data, single-instance masks, and aggregated instance masks. This strategy allowed the classifier to focus on the essential features of each instance, such as bone shapes and spatial relationships, rather than relying solely on pixel values, as done in pixel-based models like DeepLabV3+. Studies like Deshmukh *et al.* [16] have shown the limitations of pixel-based approaches in handling the anatomical variability of pediatric images, further supporting our choice to implement instance-level segmentation.

Regarding training, while PedVision benefited from iterative rounds of semi-automated annotation, other models, such as SegFormer, would require substantially more manually annotated data to achieve similar performance improvements, as shown in Fig. S2. This is a notable advantage of our approach, especially in pediatric imaging, where acquiring annotated datasets is both challenging and time-consuming due to factors like limited sample sizes and the challenges of image acquisition in younger children [12].

The PedVision pipeline addresses this challenge by producing promising initial results using VFM. Combined with the human-in-the-loop procedure, PedVision can enhance its performance during training rounds. This self-supervised method is not feasible for non-VFM based models such as SegFormer. Consequently, both PedVision and other compared models were trained on and evaluated using the same initial dataset. However, the results shown in Fig. S2 indicate that SegFormer b3, identified as the best model of trained models based on the first round data, does not exceed PedVision in performance on

pediatric cases when trained with PedVision generated last round data.

PedVision shows some recurring prediction errors (Fig. 6). The most common issue is the absence of a bone in the segmentation output. This often results from image quality limitations or VFM configuration parameters, such as the number of points used in the image. Increasing the number of points can help recover these missed regions, though at the cost of higher computational demand. Another observed error involves the appearance of extra objects and misclassification, which may occasionally mislead the IC network. Although these mistakes are less frequent, their impact can be mitigated by training the IC network with additional irrelevant objects, enabling it to better distinguish between true and false positives.

Despite all the promising results, this study has limitations that should be acknowledged. Using VFM with various configurations demonstrated its flexibility; however, the reduced performance with fewer points (Table 3) indicates that selecting VFM parameters is critical. Furthermore, model size appears to play a role in segmentation performance. The ablation study showed that smaller VFMs may fail to separate some meaningful instances, such as bones, especially when boundaries are unclear. As a result, smaller models are more prone to missing parts of bones than larger models (ViT-H). Future work could investigate how this model adapts to different configurations and compare its performance to other studies that utilize varying numbers of points in their VFM configurations.

5. Conclusions

In this study, we introduced a novel segmentation pipeline with several beneficial features for pediatric medical image processing. The model operates at the instance level, allowing it to make independent decisions for each instance within an image while simultaneously considering the context of other instances. This unique feature of including human-in-the-loop strategy and visual foundation models facilitated the development of an ultralow-effort annotation tool for initiating or providing feedback while training the pipeline, enabling the pipeline to be initiated without manually annotated datasets, thus achieving precise segmentation models rapidly. Moreover, PedVision demonstrated robust performance in handling growth-related bone changes and addressing image quality variations, including irrelevant objects. These capabilities make PedVision particularly effective for the complexities inherent in pediatric medical imaging, contributing to improved segmentation accuracy and reliability in clinical applications.

CRedit authorship contribution statement

Morteza Homayounfar: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **S.M.A. Bierma-Zeinstra:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Amir A. Zadpoor:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **Nazli Tümer:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is part of the HIPSTAR project (grant nr. ERC-adv. 101054778).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.bspc.2025.108569>.

Data availability

The used dataset is publicly available.

References

- [1] A. Kirillov, et al., Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.
- [2] M. Caron, et al., Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.
- [3] J. Ma, et al., Segment anything in medical images, *Nat. Commun.* 15 (1) (2024) 654.
- [4] L. Ke, et al., Segment anything in high quality, *Adv. Neural Inf. Proces. Syst.* 36 (2024).
- [5] V. Gilsanz, O. Ratib, *Hand Bone Age: A Digital Atlas of Skeletal Maturity*, Springer, 2005.
- [6] H. Lamecker, et al., A 3D statistical shape model of the pelvic bone for segmentation. *Medical Imaging 2004: Image Processing*, SPIE, 2004.
- [7] J. Jeuthe, et al., Semi-automated 3D segmentation of pelvic region bones in CT volumes for the annotation of machine learning datasets, *Radiat. Prot. Dosim.* 195 (3–4) (2021) 172–176.
- [8] X. Liu, et al., Fully automated pelvic bone segmentation in multiparametric MRI using a 3D convolutional neural network, *Insights Imaging* 12 (2021) 1–13.
- [9] S. Noguchi, et al., Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques, *Comput. Biol. Med.* 121 (2020) 103767.
- [10] E. Goceri, Medical image data augmentation: techniques, comparisons and interpretations, *Artif. Intell. Rev.* 56 (11) (2023) 12561–12605.
- [11] P. Liu, et al., Deep learning to segment pelvic bones: large-scale CT datasets and baseline models, *Int. J. Comput. Assist. Radiol. Surg.* 16 (2021) 749–756.
- [12] P. Ciet, et al., Magnetic resonance imaging in children: common problems and possible solutions for lung and airways imaging, *Pediatr. Radiol.* 45 (2015) 1901–1915.
- [13] A. Boutillon, et al., Multi-structure bone segmentation in pediatric MR images with combined regularization from shape priors and adversarial network, *Artif. Intell. Med.* 132 (2022) 102364.
- [14] A. Boutillon, et al., Generalizable multi-task, multi-domain deep segmentation of sparse pediatric imaging datasets via multi-scale contrastive regularization and multi-joint anatomical priors, *Med. Image Anal.* 81 (2022) 102556.
- [15] S.S. Halabi, et al., The RSNA pediatric bone age machine learning challenge, *Radiology* 290 (2) (2019) 498–503.
- [16] S. Deshmukh, A. Khaparde, Multi-objective segmentation approach for bone age assessment using parameter tuning-based U-net architecture, *Multimed. Tools Appl.* 81 (5) (2022) 6755–6800.
- [17] R. Liu, et al., Coarse-to-fine segmentation and ensemble convolutional neural networks for automated pediatric bone age assessment, *Biomed. Signal Process. Control* 75 (2022) 103532.
- [18] H. Du, et al., Hand bone extraction and segmentation based on a convolutional neural network, *Biomed. Signal Process. Control* 89 (2024) 105788.
- [19] L. Ding, et al., A lightweight U-Net architecture multi-scale convolutional network for pediatric hand bone segmentation in X-ray image, *IEEE Access* 7 (2019) 68436–68445.
- [20] Y. Nagaraju, et al., Efficient hand bone segmentation for medical applications using refined DeepLab Model, *Int. J. Pattern Recognit. Artif. Intell.* (2024).
- [21] L.-C. Chen, et al., Rethinking atrous convolution for semantic image segmentation. *arXiv. arXiv preprint arXiv:1706.05587*, 2017. 5.
- [22] E. Tay, A.A. Zadpoor, N. Tümer, Towards growth-accommodating deep learning-based semantic segmentation of pediatric hand phalanges, *Biomed. Signal Process. Control* 102 (2025) 107338.
- [23] Y. Jia, et al., Fine-grained precise-bone age assessment by integrating prior knowledge and recursive feature pyramid network, *EURASIP J. Image Video Process.* 2022 (1) (2022) 12.
- [24] C. Spampinato, et al., Deep learning for automated skeletal bone age assessment in X-ray images, *Med. Image Anal.* 36 (2017) 41–51.
- [25] A. Gertych, et al., Bone age assessment of children using a digital hand atlas, *Comput. Med. Imaging Graph.* 31 (4–5) (2007) 322–331.
- [26] K. Wada, *Labelme: Image Polygonal Annotation with Python [Computer software]*. github.com, 2018.
- [27] P. Iakubovskii, *Segmentation Models Pytorch*, 2019. Available from: https://github.com/qubvel/segmentation_models.pytorch.
- [28] M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, PMLR, 2019.
- [29] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer, 2015.
- [30] M. Sandler, et al., Mobilenetv2: inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [31] K. He, et al., Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [32] E. Xie, et al., SegFormer: simple and efficient design for semantic segmentation with transformers, *Adv. Neural Inf. Proces. Syst.* 34 (2021) 12077–12090.
- [33] S. Perera, P. Navard, A. Yilmaz, Segformer3d: an efficient transformer for 3d medical image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [34] Ş. Vădineanu, et al., An analysis of the impact of annotation errors on the accuracy of deep learning for cell segmentation. *International Conference on Medical Imaging with Deep Learning*, PMLR, 2022.
- [35] R. Bryant, et al., Racial differences in bone turnover and calcium metabolism in adolescent females, *J. Clin. Endocrinol. Metabol.* 88 (3) (2003) 1043–1047.
- [36] M. Laster, R.C. Pereira, I.B. Salusky, Racial differences in bone histomorphometry in children and young adults treated with dialysis, *Bone* 127 (2019) 114–119.
- [37] R. Wetzsteon, et al., Ethnic differences in bone geometry and strength are apparent in childhood, *Bone* 44 (5) (2009) 970–975.