Towards growth-accommodating deep learning-based semantic segmentation of pediatric hand phalanges

Tay, Edwin; Zadpoor, Amir A.; Tümer, Nazli

DOI
10.1016/j.bspc.2024.107338

Publication date
2025

Document Version
Final published version

Published in
Biomedical Signal Processing and Control

Biomedical Signal Processing and Control 102 (2025) 107338

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Towards growth-accommodating deep learning-based semantic segmentation of pediatric hand phalanges

Edwin Tay [*], Amir A. Zadpoor [1], Nazli Tümer [1]

Department of Biomechanical Engineering, Faculty of Mechanical Engineering, Delft University of Technology (TU Delft), Mekelweg 2, 2628 CD, Delft, the Netherlands

ARTICLE INFO

Keywords:
Deep learning
Semantic segmentation
Hand phalanges
Pediatric medical imaging
Radiographs

ABSTRACT

Existing deep learning (DL) networks are primarily trained on adult datasets and may not always generalize to pediatric populations, where growth plays a major role. Here, we investigated improving semantic segmentation outcomes of pediatric hand phalanges from radiographs without relying on fully pediatric training datasets, which are scarce. First, alternative DL networks (FCN-8, FCN-32, U-Net, Inception U-Net, and DeepLabv3+) were trained with manually segmented radiographs of near-skeletally-mature (NSM) subjects and their performances were evaluated using mean intersection-over-union (Mean IoU) and multiclass Dice scores. DeepLabv3+ and Inception U-Net performed the best for NSM segmentation, with Mean IoU scores of 0.899 ± 0.035 and 0.887 ± 0.062, respectively. These networks were then used to investigate zero pediatric data (scaling-based data augmentation) and minimal pediatric data (incremental pediatric data substitution) approaches to improve age-domain generalizability. The minimal pediatric data approach proved effective, with a 20 % pediatric data inclusion leading to an up to 21.1 % increase in Mean IoU for pediatric subjects compared to networks trained exclusively on NSM subjects. Furthermore, no adverse effects of this approach were found when tested on NSM subjects, and there were even improvements in performance for Inception U-Net. To conclude, we highlight that networks utilizing multi-scale filters perform best for the semantic segmentation of hand phalanges. We further demonstrate that a minimal inclusion of pediatric training data can markedly improve age-domain generalizability for semantic segmentation tasks. This removes the difficult task of gathering large training datasets of pediatric subjects, which is often impractical, if not impossible.

1. Introduction

The prominence of artificial intelligence (AI), enabled by deep learning (DL) techniques, is evident in contemporary medical imaging research [1,2]. Specifically for the task of image segmentation, traditional methods (*e.g.*, thresholding) have been outperformed by DL-based strategies [2,3]. Furthermore, feature maps extracted by DL-based approaches for the purpose of segmentation have been utilized in multi-task frameworks, such as disease diagnosis and/or prognosis [4]. While DL has emerged as a powerful tool for medical image analysis, there is a lack of high-quality, diverse datasets [5,6]. This is particularly apparent in pediatric populations for which there is a general lack of imaging data due to increased radiological health risks [7], difficulty in handling small children during image acquisition [8,9], and ethical concerns [10].

Taken together, these difficulties thus prevent the collection of large datasets of pediatric medical imaging to train DL models. Furthermore, existing AI models, usually trained on adult datasets, rarely perform well when tested on pediatric patient data [11–15]. Nevertheless, the possible applications of AI in pediatric healthcare are innumerable from both prognostic and diagnostic standpoints [16–20]. These applications have the potential to greatly improve patient outcomes and the overall quality of healthcare. Pediatric data paucity, therefore, presents an open challenge for widespread use of AI in healthcare. Specifically, models should be made to be sufficiently age-agnostic even with a lack of existing age-diverse datasets and difficulties in gathering large pediatric training datasets.

Semantic segmentation, wherein anatomical structures are both delineated and labeled, is emerging as the leading direction of research in medical imaging analysis to examine multi-organ images. As compared to previous approaches for multi-body semantic segmentation (*e.g.*, shape model based fitting [21–23]), DL-based semantic

* Corresponding author.
E-mail address: e.w.s.tay@tudelft.nl (E. Tay).
[1] Both authors contributed equally to this study.

https://doi.org/10.1016/j.bspc.2024.107338
Received 14 August 2024; Received in revised form 19 November 2024; Accepted 8 December 2024
Available online 13 December 2024

segmentation methods have become increasingly powerful and consistently achieved record-setting results [24,25]. Despite the increased anatomical significance of 3D medical imaging, 2D medical imaging modalities are more accessible, especially in low-resource settings [26,27]. Therefore, semantic segmentation of structures from 2D radiographs continues to be an active area of research. For example, Ryu *et al.* semantically segmented feet radiographs to quantify flatfootedness [28]. Segmenting chest radiographs into their constituent anatomical structures (*i.e.*, clavicle, heart, lungs, etc.) has also been done both for anatomical studies and to serve as a basis for diagnostic pipelines [29,30]. Liu *et al.* focused specifically on delineating ribs and clavicles for either suppression from radiographs to uncover underlying soft tissue, or to highlight the bony structures themselves [31]. Hand phalange segmentation from radiographs is also useful in bone age prediction as demonstrated by Lv *et al.* and the RSNA Pediatric Bone Age Challenge [32,33]. Hatano *et al.* extensively investigated this task, utilizing network architectures of increasing complexity and also ensemble-like frameworks [34–36]. While semantic segmentation of 2D radiographs continues to be relevant, the lack of diverse age-domain datasets presents a major challenge to pediatric applications of DL-based approaches devised for such tasks.

Several existing works have implemented novel solutions to increase age-domain generalizability and address the issue of pediatric data paucity. Boutillon *et al.* developed a multi-task, multi-domain framework incorporating multi-scale contrastive regularization and multi-joint anatomical priors to improve generalizability of their models for MRI knee, shoulder, and foot bone segmentation in a pediatric population [37]. Their study, however, investigated a relatively small age range (at most from 5 to 17 years of age) and their datasets for training and testing were comparatively small. Furthermore, while effective, relying on learned anatomical priors from multi-anatomy (ankle, knee, shoulder) imaging necessitates collecting said multi-anatomy data, which is difficult in the case of pediatric subjects. Similarly, Rajaraman *et al.* implemented an ensemble-based solution, relying on a stacked ensemble of networks to improve age-domain generalizability of their chest X-ray segmentation models [38]. They demonstrated improved segmentation performance. However, their approach only carried out binary segmentation as opposed to semantic segmentation. Recently, Kumar *et al.* have investigated the effects of pediatric data inclusion on piecewise binary segmentation of differing pelvic and thoracic organs from CT images [39]. They demonstrated that ensuring an age-diverse dataset improves pediatric segmentation outcomes. Nevertheless, their main finding seemingly arose when combining separate training datasets from differing age categories. Therefore, their findings of improved performance could simply be a function of increased training dataset size, and a more granular parametric investigation into increasing age diversity of training data may be warranted. A study by Somasundaram *et al.* demonstrated that transfer learning-based fine-tuning of adult-trained models on pediatric data improves segmentation outcomes [40]. While effective, it, nevertheless, necessitates collating data from multiple sources, which may or may not exist in an open-source setting for specific applications. Taken together, these studies made initial steps to address the issue of age-domain generalizability of DL models. However, they relied on collating images from multiple anatomical structures and sources, were limited in the scope of DL task, or did not adequately separate the effects of age-diverse training data from those of a larger training dataset.

In this study, we aim to tackle the open research problem of age-domain generalizability of DL models in the context of semantic segmentation, a comparatively more complex task than binary segmentation. We aim to provide relatively straightforward guidelines to ensure that practitioners are aware of easy-to-implement methods to improve the age-domain generalizability of their models without necessitating specific network architectures or collecting additional data. Specifically, we aim to investigate two potential methods that require minimal pediatric data to improve the generalizability of DL-based semantic segmentation workflows. As far as the first approach (*i.e.,* zero additional pediatric data) is concerned, a relatively commonplace practice in the field of DL is the use of data augmentation, wherein existing data are modified to increase dataset size and improve generalizability [41,42]. For images, examples of common modifications include rotations, blur, and also scaling. The lattermost augmentation is particularly notable as bones have been determined, generally, to isometrically scale during growth [43]. Thus, one can rationally infer that the use of scaling-based data augmentations can enrich the training data for DL networks, improving age-domain generalizability. Therefore, while most networks already employ a degree of data augmentation, the efficacy of scaling-based augmentation during training to address pediatric data paucity is yet to be elucidated upon. Regarding the second approach (*i.e.,* minimal pediatric data), we aim to study to what extent is the outcome of semantic segmentation dependent on the amount of pediatric data included in training data. We study the efficacy of both above-mentioned approaches within the context of the semantic segmentation of pediatric phalanges from hand radiographs. We first explore various DL network architectures and quantify their effectiveness in semantically segmenting (near) skeletally mature hand phalanges (NSM) from hand radiographs. Then, we use both alternative approaches to determine to what extent they could improve segmentation performance on pediatric subjects. Overall, the results of this study may lead to methodological enhancements or improved guidelines surrounding training dataset construction and pre-processing. These advances could ensure that DL models used in medical image analysis of bones are more generalizable towards pediatric patients in spite of a prevailing paucity of pediatric training data.

2. Materials and methods

2.1. Dataset and pre-processing

In this study, we utilized hand radiographs from the RSNA Paediatric Bone Age Challenge (2017) dataset [44]. We first categorized the data into their developmental stages based on the third version of the Tanner-Whitehouse method (TW3) [45]. We then generalized these data categories further to simplify our analyses. Stages A-D were grouped together as they represent the youngest subjects within the dataset (0–84 months old), and is hereby referred to as the pediatric dataset. Stages G-I on the other hand represent the oldest subjects from the dataset (>169 months old), and is referred to as the NSM dataset (Fig. 1A). 400 radiographs (200 NSM, 200 pediatric) were then randomly selected, pre-processed, and segmentation masks were created (Fig. 1B).

In terms of pre-processing, contrast limited adaptive histogram equalization (CLAHE) was carried out to enhance edges within the dataset and improve segmentation outcomes [46]. Then, segmentation masks highlighting and delineating the metacarpals and the proximal, medial, and distal phalanges were created using the open-source software labelme [47]. All the radiographs were then resized to uniform dimensions (256 x 256 pixels). 10 % of each dataset was extracted and set aside as NSM and pediatric test datasets, NSM-Te and P-Te, respectively. The rest of the data were designated as NSM and pediatric training data, NSM-Tr and P-Tr, respectively. Using a K-folds validation strategy ($K = 5$), NSM-Tr and P-Tr were then split into folds and initially augmented with horizontal flipping and random rotations ($\theta = \pm 20\circ$).

2.2. Improving pediatric generalizability

We investigated the efficacy of two different approaches to pediatric data generalizability of our networks, that is a zero pediatric data and minimal pediatric data approach (Fig. 1C). For the zero pediatric data approach, based on the principle of isometric scaling found in long bones [43], we sought to investigate the use of scaling-based data augmentation in improving performance on pediatric test data. We do so

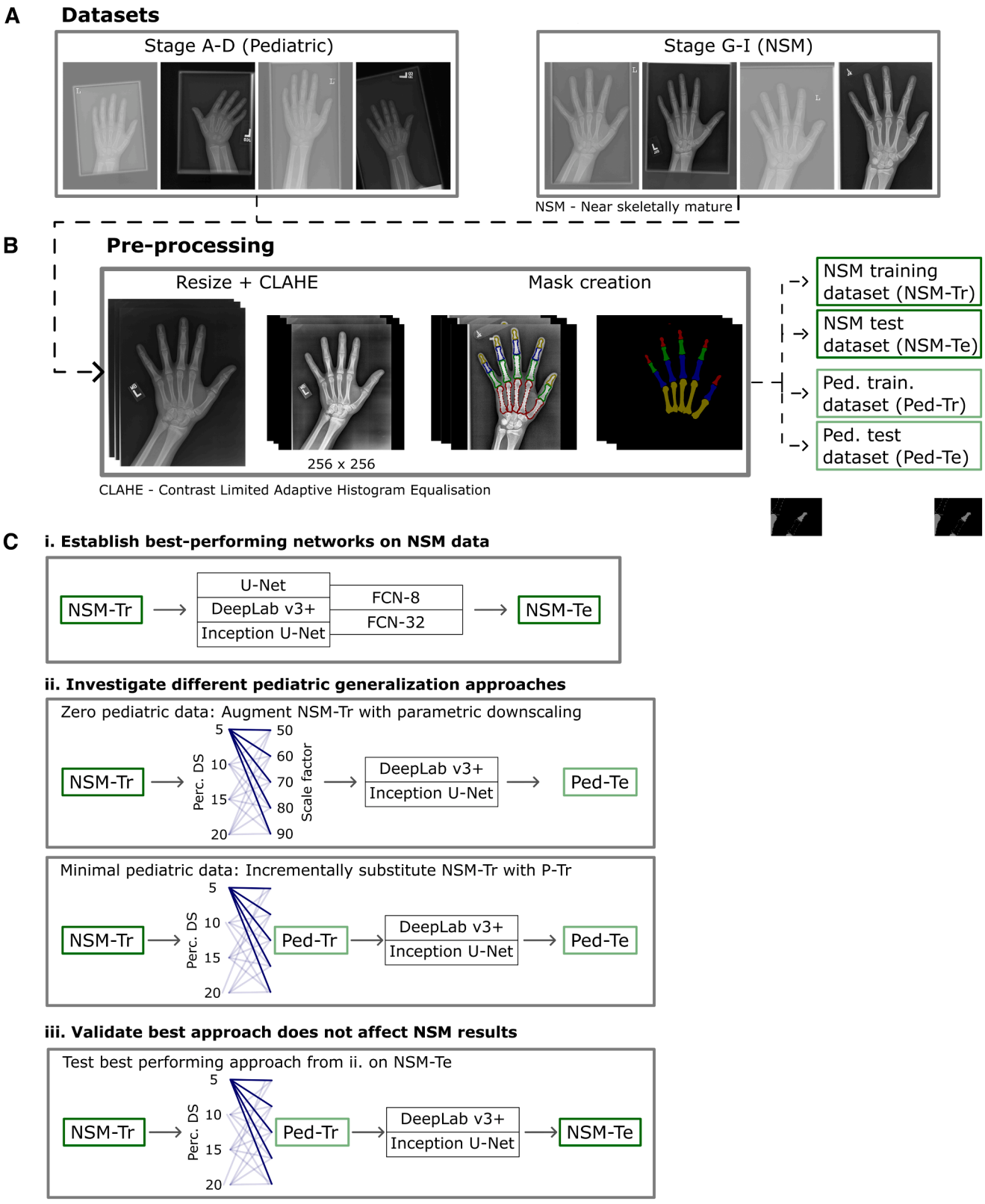

Fig. 1. Illustration of workflow of this study. A) First, the dataset is delineated into developmental stages based on their age according to the TW3 method and grouped as pediatric (Stages A-D) and near skeletally mature (NSM) (Stages G-I) [39]. B) The datasets are then pre-processed with resizing, CLAHE, and segmentation masks are manually created. These datasets are then separated further into testing and training datasets. C) In this study, i) we first determined which network architecture worked best on segmenting NSM subjects. ii) Then, we utilised the best performing networks to investigate a zero pediatric data and a minimal pediatric data approach to improve pediatric generalizability. iii) We then investigated if improvements in pediatric generalizability affected performance on NSM subjects.

as no previous studies have investigated the efficacy of this approach for improving age-domain generalizability specifically. Thus, in this study, varying proportions of NSM-Tr (n = 5, 10, 15, 20 %) were affinely downscaled by varying scaling factors (*s.f.* = 50, 60, 70, 80, 90 %). For the minimal pediatric data approach, we attempted to investigate and quantify the effect of utilizing an age-diverse dataset. Whilst existing studies have demonstrated, generally, the efficacy of including additional pediatric data for improved segmentation outcomes, existing works do not quantitatively investigate training data composition. We did so by taking NSM-Tr and substituting varying proportions (n = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 %) with P-Tr to quantify the effect of utilizing progressively more age-diverse training data. Networks trained via both approaches were then tested on P-Te and compared to networks trained solely on NSM-Tr and P-Tr to quantify any improvements in segmentation outcomes. The best performing approach was then tested on NSM-Te to investigate if it is able to generalize towards pediatric subjects without a loss in performance towards NSM subjects.

2.3. Network architectures

Several network architectures designed for semantic segmentation were utilized in this study (Supplementary S1). Fully convolutional networks (FCNs) were utilized as they represent the first DL networks utilizing convolutional layers for pixel-wise labeling [25,48]. We implemented the FCN-32 and FCN-8 variations in this study, representing the worst and best variations of the FCN networks, respectively, with a VGG-16 encoder as the network backbone. A further development of the FCN principle is U-Net as proposed by Ronneberger *et al.* [49]. U-Net utilizes skip-connections, which integrate both high- and low-level features for enhanced segmentation accuracy. Its popularity owing to its performance has led to the development of many variations [50]. Nevertheless, we utilized the original version proposed by Ronneberger *et al.* in this work for simplicity. Another network architecture we utilized was DeepLabv3+. This architecture is the culmination of continuous improvement upon the principle of multi-scale convolutions for semantic segmentation [51]. Essentially, it relies on a unique spatial pooling layer to encode multi-scale features using parallel atrous convolutions with different rates. Finally, building on a similar principle of multi-scale convolutions, we implemented an Inception U-Net network architecture. Inception modules concatenate the feature map outputs of multi-scale convolution filters to better encapsulate features at multiple length scales [52]. Several works have utilized these modules within 'U-Net'-like encoder-decoder configurations for semantic segmentation tasks [53–55], and we followed a similar configuration.

2.4. Training

The training parameters were set based on initial pilot experiments. First, a learning rate of 1E − 3 and batch size of 5 were set. Then, categorical cross-entropy (CCE) was used as the loss function with an adaptive moment estimation (Adam) optimizer. CCE was utilized as this loss function is the most commonly used for semantic segmentation tasks, and based on initial pilot tests, led to the best training outcomes (Equation (1)). Essentially, it functions as a negative log-likelihood acting over a distribution of n discrete classes, which is minimized during training [25,56,57]. In detail, p represents the true labels and $\widehat{p}$ represents the predicted label.

$$J_{cce} = -\sum_{i=1}^{n} p\log(\widehat{p}) \tag{1}$$

To determine convergence, each network was first trained and tested on NSM data for 500 epochs with the loss at each epoch recorded. Based on these loss plots, epochs of convergence for each network architecture were determined when training and validation losses reached a minima and exhibited minimal further variation (Supplementary S2). Thus, for the rest of the study, all the networks were trained to their respective epochs of convergence. All the networks were created and trained using the Tensorflow (2.10.0) framework in Python (3.10.11). The networks were all trained on the GPU nodes (NVIDIA Tesla V100S, 32 GB) of the DelftBlue High Performance Computing Cluster [58].

2.5. Performance metrics

In this study, we utilized mean intersection-over-union (Mean IoU) and multiclass dice (MC Dice) scores to evaluate the efficacy of our networks. Both of these metrics were calculated and averaged across each fold of our K-folds validation. Mean IoU is commonly used to evaluate the results of semantic segmentation tasks [25]. First, IoU was calculated using the ratio of true positives (*TP*), false positives (*FP*), and false negatives (*FN*) from a confusion matrix (CM) comparing a predicted and true mask (Equation (2)). IoU for each class (*i.e.*, category of phalange) except for the background was determined and the mean was calculated to determine the mean IoU.

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

Similarly, Dice coefficient is a metric commonly used in evaluating segmentation performance [25]. For binary masks, it can be calculated as follows:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{3}$$

In our study, we are concerned with multi-class segmentation outcomes, a Dice coefficient for each class (excluding the background) was first calculated, then a mean was calculated to determine the corresponding multi-class Dice (MC Dice) score. Another metric we utilized was mean pixel accuracy (mPa), which evaluates the ratio of correctly classified pixels against the total number of pixels per class [59]. Similar to the Dice coefficient, mPa is taken for each class (excluding the background), and a mean was calculated to determine the mPa over all the classes:

$$mPa = \frac{TP}{TP + FN} \tag{4}$$

2.6. Statistical analysis

To compare the performance of various network architectures and augmentation strategies, we performed one-way analysis of variance (ANOVA) with the Tukey's *post-hoc* tests for multiple comparisons. Significance levels were set at $p < 0.05$ but were corrected, whenever appropriate, using the Bonferroni correction. All the statistical analyses were performed using SPSS Statistics (v.29.0.0.0; IBM Corp., Armonk, New York, USA).

3. Results

3.1. NSM segmentation

Our initial results revealed a clear disparity in the performance of our implemented network architectures for the semantic segmentation of NSM hand phalanges. When comparing the CMs (Fig. 2B), we can qualitatively observe that Inception U-Net and DeepLabv3 + are much more generalizable and robust than FCN-32, FCN-8, and U-Net. This is evidenced by the consistently high number of TPs identified for all classes across all K-folds, with minimal FPs and FNs. Utilising these CMs, the architectures can then be compared quantitatively (Fig. 3). FCN-32 performed the worst, with a Mean IoU of 0.09 ± 0.177 (mean ± standard deviation) and an MC Dice of 0.123 ± 0.234. This can also be qualitatively observed in sample predicted masks (Fig. 4), wherein the predicted masks using FCN-32 often do not even capture the general shape of the phalanges and are often incomplete. FCN-8 and U-Net,

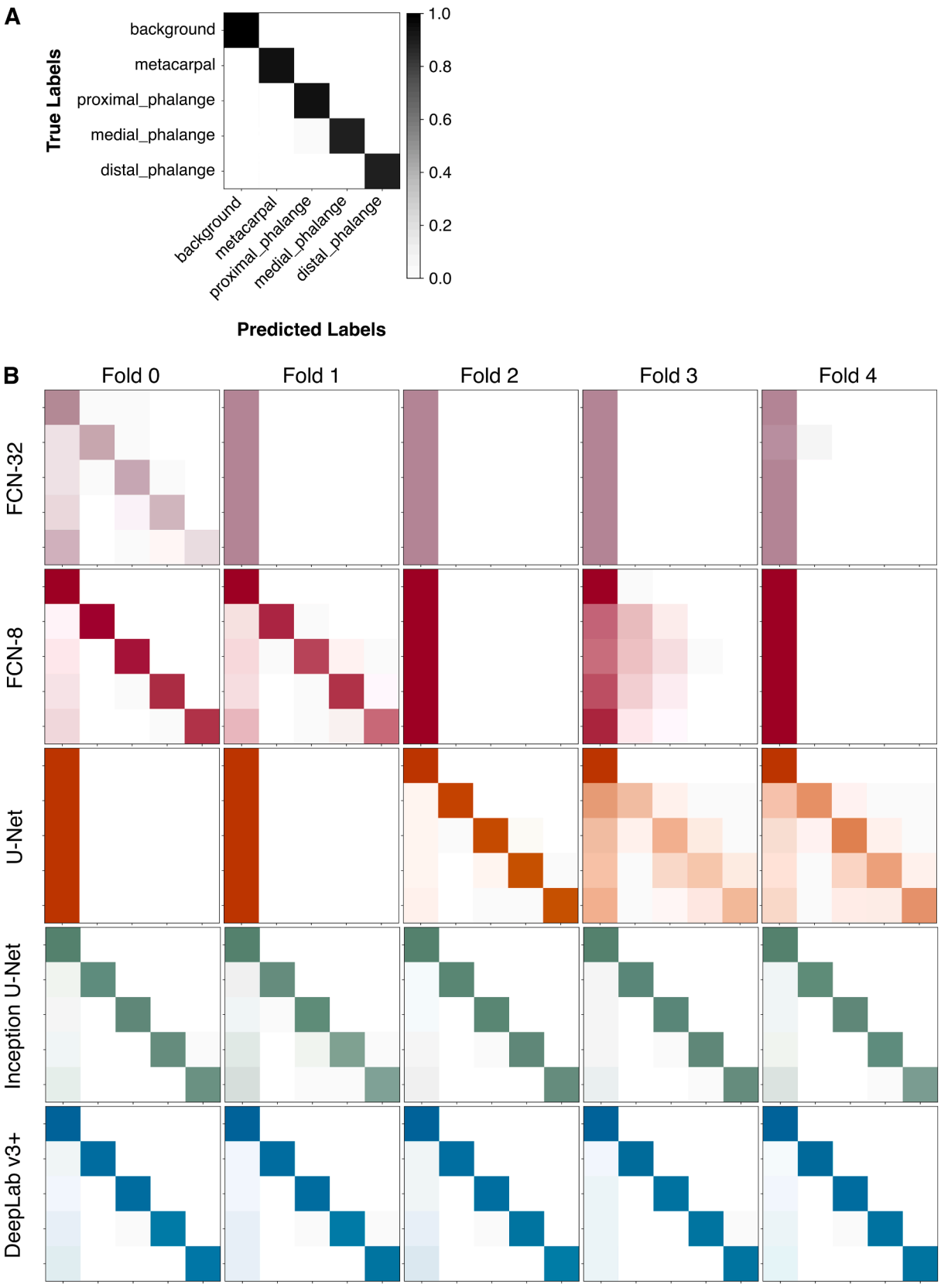

Fig. 2. Confusion matrices comparing the performance of varying network architectures trained and tested on adult data. A) Template confusion matrix detailing the various classes and color gradient scale in greyscale. B) Confusion matrices of all the trained and tested networks across the K-folds validation.

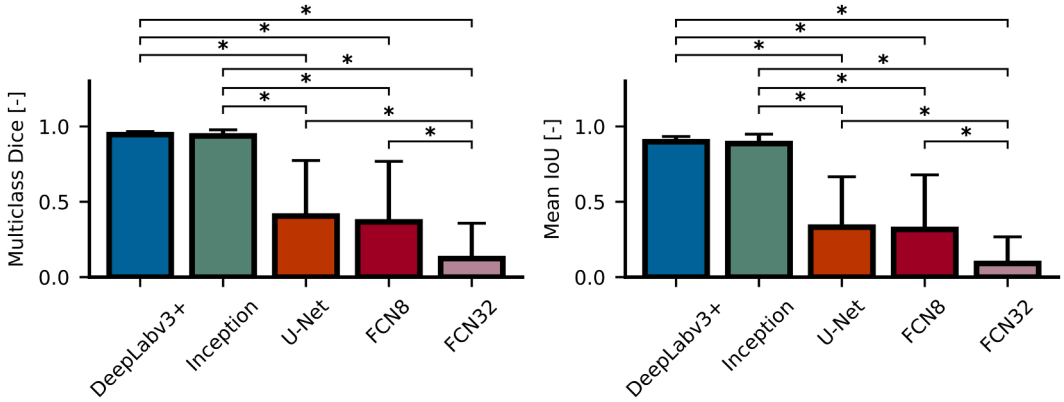

Fig. 3. Metrics from testing the converged networks on adult data, in terms of semantic performance metrics. The significance is marked with *, corresponding to $p < 0.01$. Significance level was determined by adjusting the standard $p < 0.05$ value with a Bonferroni correction. Vertical lines represent standard deviation.

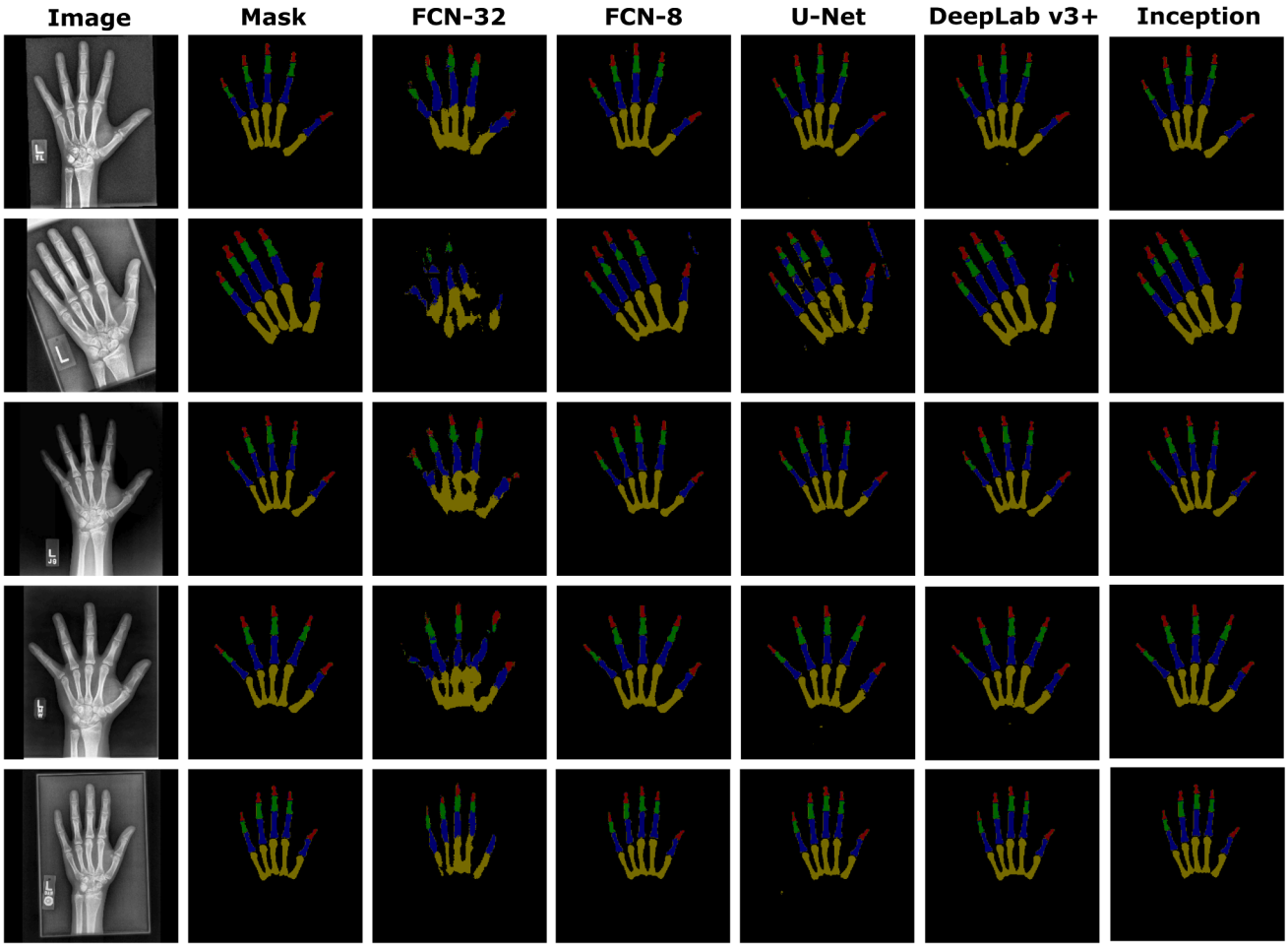

Fig. 4. Predicted NSM segmentation masks from our converged networks.

showed similar performances with the latter peaking at a Mean IoU of 0.332 ± 0.335 and an MC Dice of 0.405 ± 0.369. This is also reflected in the predicted masks (Fig. 4), which show both networks can capture the shape of the phalanges relatively well. However, some small areas of misclassification and false positives are still evident. DeepLabv3 + and Inception emerged as the best performing architectures in terms of high average semantic metrics and low standard deviations. The former performing slightly better at a Mean IoU and MC Dice of 0.899 ± 0.035 and 0.946 ± 0.021, respectively. This is also apparent in the predicted masks (Fig. 4), wherein the shapes of the phalanges are remarkably well captured although with some small areas of misclassification in both architectures. Based on these results, DeepLabv3 + and Inception were chosen for further study on improving pediatric data generalizability.

3.2. *Improving pediatric generalizability*

3.2.1. *Zero pediatric data approach: Scaling-based data augmentation*

We first investigated the use of scaling-based augmentation by

retraining the best performing networks (DeepLabv3 + and Inception U-Net) with various scaled datasets as previously described (Section 2.2, Fig. 1C). Our results for the scaling study demonstrated that the parameters chosen for rational scaling (*i.e.*, dataset percentage and scaling factor) minimally affect the segmentation outcomes for pediatric data (Fig. 5A). This observation was consistent across both network architectures, with relatively unchanged mean and standard deviation values. For Inception, however, a decrease in both Mean IoU and MC Dice at a scaling factor of 50 % and dataset percentage of 20 % can be noted with the caveat of a high standard deviation (Fig. 5A). In fact, a key caveat with our results is that they are marred by high standard deviations across both network architectures and all parameters (see Section 4 for a discussion). Nevertheless, the best performing networks (in terms of mean metrics) for Inception were with 20 % data downscaled by 60 %, whilst for DeepLabv3 + it was 20 % data downscaled by 70 %.

3.2.2. *Minimal pediatric data approach: Pediatric training data substitution*

The alternative approach to scaling-based augmentation was P-Tr substitution. For both network architectures, similar improvements can be observed when applying pediatric substitution. Overall, increasing the amount of adult training data being substituted with pediatric data leads to improvements in the segmentation metrics (Fig. 5B). With only 5 % P-Tr substitution, we only have 0.72 ± 0.22 and 0.75 ± 0.21 Mean IoU for DeepLabv3 + and Inception respectively. However, increasing P-Tr substitution up to 50 % increases these values up to peak Mean IoU values of 0.88 ± 0.08 and 0.9 ± 0.1 for DeepLabv3 + and Inception respectively. A clear pattern is that increasing P-Tr substitution percentage not only increases mean values, but also decreases standard deviations. Nevertheless, what we can note is that even a relatively low percentage of P-Tr substitution (*i.e.,* 20 %) led to a Mean IoU of 0.804 ± 0.157 and 0.819 ± 0.168 for DeepLabv3 + and Inception respectively. These values are markedly better than with a lower proportion of substitution, and potentially better than utilizing a training dataset of 100 % NSM-Tr. To illustrate the use of a minimal proportion of pediatric data further, we selected 20 % P-Tr inclusion for comparison with the zero pediatric data approach and against networks trained solely with 100 % NSM-Tr and P-Tr.

3.2.3. *Approach comparison*

To compare their efficacy, we selected the best performing networks from both approaches across both network architectures and compared them against negative and positive controls (*i.e.,* networks trained with 100 % NSM-Tr and 100 % P-Tr data respectively). For this comparison, the additional metric of mPa was included to enhance the diversity in comparisons for our proposed approaches. For all networks and metrics, networks trained on 100 % NSM-Tr and 100 % P-Tr are statistically significantly different, with the latter performing comparatively better as expected. This confirms our initial claim that networks trained solely on NSM data do not generalize well to pediatric subjects (Fig. 6). We can then note that for both networks, similar patterns are observed across the utilized approaches. Firstly, for both Mean IoU and MC Dice, networks trained on 100 % P-Tr perform statistically significantly better than all of the others as expected, except P-Tr substitution. In fact, P-Tr substitution performs statistically significantly better than 100 % NSM-Tr, but there is no statistically significant difference compared to 100 % P-Tr. This is interesting because for both of these networks, only 20 % of NSM-Tr was substituted with P-Tr, and this led to an up to an impressive 21.1 % increase in Mean IoU as compared to networks trained on 100 % NSM-Tr. In comparison, the scaling augmentation approach yielded minimal improvements over the 100 % NSM-Tr networks. Specifically, for both networks, we only found an up to 9.98 % increase in Mean IoU with a no statistically significant difference compared to 100 % NSM-Tr. In terms of mPa, Inception exhibits similar results for the minimal and zero pediatric data approaches. However for DeepLabv3 + 100 % P-Tr is not statistically significantly different from scaling augmentation. Nevertheless, the similarities in our comparison of metrics, that is, between Mean IoU, MC Dice, and mPa suggest an underlying pattern. That is, our results suggest that the minimal pediatric data approach (*i.e.,* P-Tr substitution) can improve the segmentation performance on pediatric subjects while requiring only 20 % pediatric data. We then re-tested the networks trained with P-Tr substitution on NSM-Te (Fig. 7). We found that the use of P-Tr substitution did not adversely affect their performance on NSM-Te. In fact, for DeepLabv3+, there was not a statistically significant difference between the negative control (*i.e.,* 100 % NSM-Tr) and P-Tr substitution, implying that performances are comparable and P-Tr leads to similar results. There is, however, a statistically significant difference compared to the positive control (*i.e.,* 100 % P-Tr) for all metrics. This demonstrates that P-Tr leads to improved pediatric generalizability without decreasing the performance on NSM subjects. For Inception, we found similar results but with the additional benefit of a statistically significant improvement over the positive control. Thus, this approach improves both generalizability towards pediatric population and the overall performance of the networks.

4. Discussion

An initial finding we had was the marked contrast in the performance of the networks trained and tested for the NSM segmentation task (Figs. 2-3). The under-performance of FCN-32 can be attributed to its simplistic architecture, wherein the drastic up-sampling led to inaccurate and indistinct segmentation masks being learned. However, an unexpected result is that U-Net under-performed, with statistically similar performance metrics to that of FCN-8. The characteristic skip-connections and encoder- decoder architecture of U-Net usually leads to better performance, especially as compared to simpler FCN-type networks [50]. This under-performance could be due to the relative complexity of this particular segmentation task. Wherein, the varied shapes and sizes of the assorted phalanges cannot be effectively captured solely by skip-connections and an encoder-decoder style architecture. A marked difference in our results is between the networks with multi-scale filters (Inception and DeepLabv3+) and without (FCN-32, FCN-8, and U-Net). From our results, we can surmise that multi-scale filters enhance the ability of segmentation networks to learn segmentation maps of assorted anatomical structures of interest with varying shapes and sizes. Not only do the average metrics of these multi-scale filter networks outperform the others, a distinct aspect is their relatively low standard deviations. Especially in the context of our *K*-folds validation, this demonstrates that these networks not only perform better, but are also more consistent and can more easily generalize as compared to the other architectures.

In comparison to other studies, Ding *et al.* achieved an IoU and Dice score of 0.929 ± 0.023 for the segmentation of paediatric hand bones from radiographs using a multi-scale block in an encoder-decoder configuration, relatively similar to our implementation of Inception U-Net [60]. In comparison, we achieved a Mean IoU and MC Dice of 0.887 ± 0.062 and 0.937 ± 0.04, respectively, using Inception U-Net, demonstrating similar performance. However, in their study, the segmentation task was slightly different, as they sought to obtain binary masks of all the hand bones, including the carpals, radius, and ulna, while we attempted to obtain semantic masks of the phalanges exclusively. Furthermore, they explicitly delineated unfused epiphyses of the phalanges as two separate structures while our masks considered them as parts of the same structure. Differences notwithstanding, the similarly improved performance when using multi-scale filters demonstrates the superiority of these types of modules for segmentation tasks involving structures of varying sizes, especially hand bones. Ono *et al.* utilized vanilla DeepLabv3 + and developed a variant for a similar task of semantically segmenting proximal and medial phalanges of adult hand bones (excluding the thumb) from radiographs [36]. Both configurations of DeepLabv3 + achieved a Mean IoU of 0.949, similar to us achieving a Mean IoU of 0.899 ± 0.035. Our comparatively lower performance is attributed to the difference in task, wherein ours can be

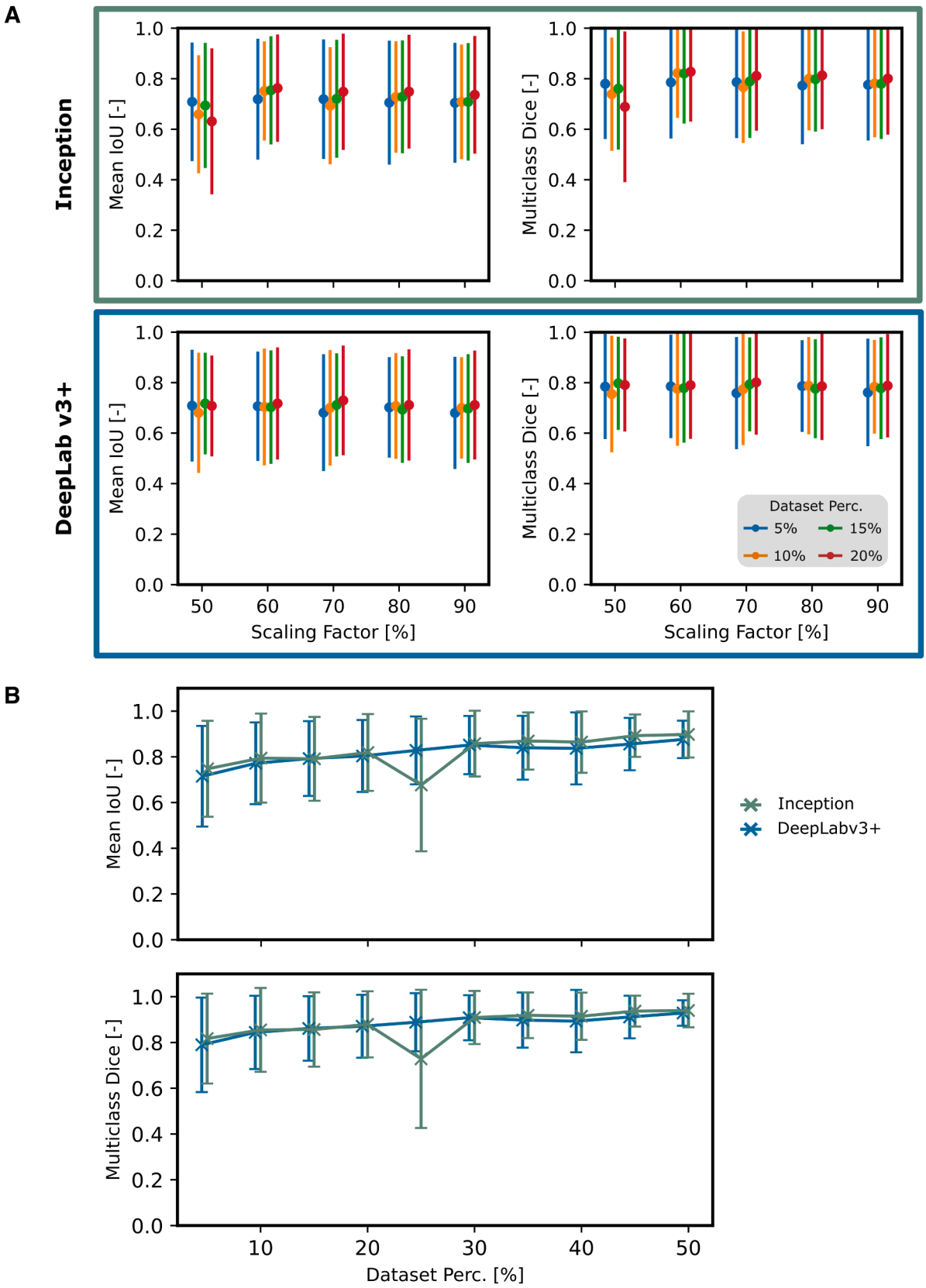

Fig. 5. A) Results from the parametric augmentation study, split into Inception U-Net (top) and DeepLabv3+ (bottom). B) Results from the P-Tr substitution study. Note that vertical lines represent standard deviation.

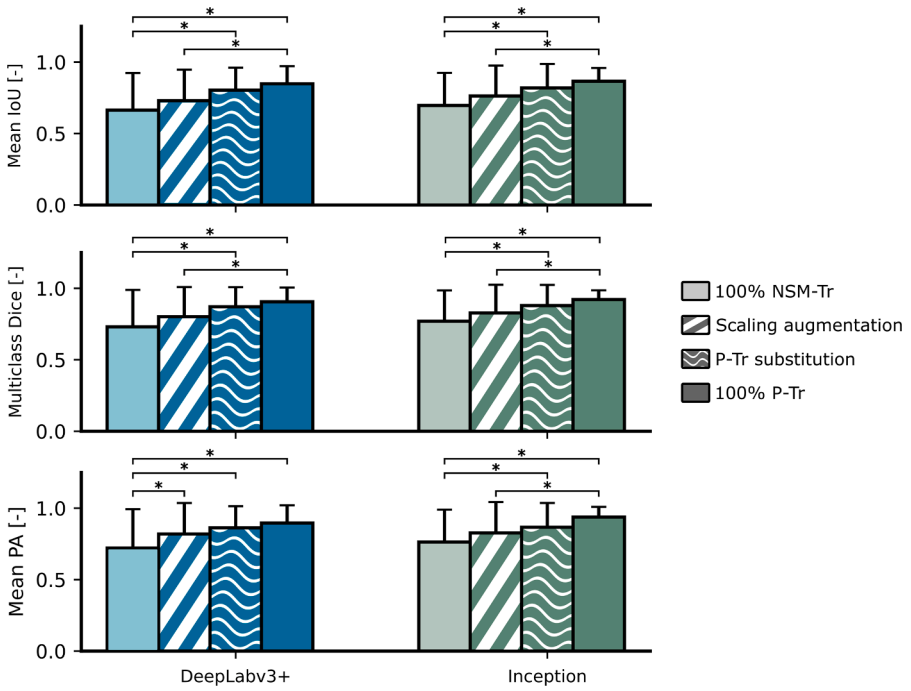

Fig. 6. Comparison of our alternative generalization approaches, with both metrics calculated by testing on P-Te. Metrics reported are mean intersection-over-union, multiclass Dice, and mean pixel accuracy (top to bottom). 100 % NSM-Tr refers to a training dataset of exclusively NSM-Tr, while 100 % P-Tr refers to a training dataset of exclusively P-Tr. Significance is marked with *, corresponding to $p < 0.0167$. The significance level was determined by adjusting the standard $p < 0.05$ value with a Bonferroni correction. Note that vertical lines represent standard deviation.

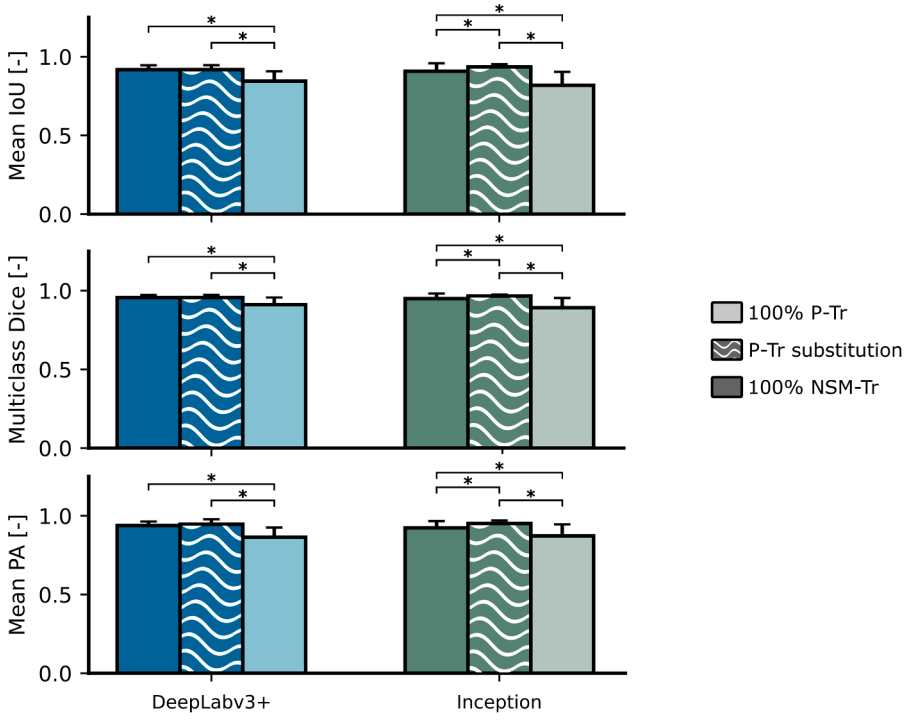

Fig. 7. Results from testing trained networks on NSM-Te to validate if P-Tr substitution affected performance on NSM subjects. Metrics reported are mean intersection-over-union, multiclass Dice, and mean pixel accuracy (top to bottom). 100 % NSM-Tr refers to a training dataset of exclusively NSM-Tr, while 100 % P-Tr refers to a training dataset of exclusively P-Tr. Significance is marked with *, corresponding to $p < 0.0167$. The significance level was determined by adjusting the standard $p < 0.05$ value with a Bonferroni correction. Note that vertical lines represent standard deviation.

considered more complex due to the additional category of phalange we segmented in our networks. Nevertheless, the comparable performance could indicate, again, that networks with multi-scale filters are more suitable for segmentation tasks involving structures of varying sizes. Hatano *et al.* utilized U-Net to segment binary masks of the medial and proximal phalanges from NSM radiographs, and achieved an impressive

Mean IoU of 0.914 [34]. In contrast, in our study, U-Net achieved a Mean IoU of only 0.332 ± 0.335. The superior performance of U-Net in their study could be due to their relatively simpler segmentation task, as they sought to only generate binary masks. Furthermore, their segmentation masks consisted of only two anatomical structures of relatively similar sizes. In comparison, our inclusion of an additional anatomical structure in the distal phalanges may have made the feature maps too complicated for U-Net.

In terms of our investigation into our diverging approaches to improve age-domain generalizability, we find that the minimal pediatric data approach (scaling-based augmentations) had little to no effect (Fig. 5A). For both networks, all combinations of augmentation parameters seemingly does not provide additional useful information to the networks, leading to an no observable effect on segmentation outcomes. A potential reason for this could be related to the notably high standard deviations. These high standard deviations found in both approaches (Fig. 5A-B) could be due to the granularity of our analysis and grouping of pediatric data. Specifically, the age range we labeled as pediatric (stages A-D, corresponding to 0–7 years of age), could be too broad. Thus, our augmentations may be performing better on some ages and not others leading to consistently high standard deviations. As this age range constitutes rapid growth and changes of anatomy, a more comprehensive analysis using more dedicated approaches may be required. This potentially warrants investigating the parametric effects at each TW3 stage or bone age, either on a yearly or monthly graduation. This, however, requires significant additional resources. In practical terms, sufficient numbers of segmentation masks for each TW3 or bone age would need to be annotated and a nontrivial number of DL networks have to be trained for validation, warranting significant additional computational resources.

Comparing the zero pediatric data and the minimal pediatric data approach (P-Tr substitution), we found that the latter performs better. Using real pediatric data as constituent training data led to the networks robustly learning the features of pediatric data. For example, the nonlinear nature of anatomical growth of hand phalanges (*i.e.,* the emergence and fusing of the epiphysis) cannot be captured by our naive scaling strategy. Thus, ensuring that training data is as age-diverse as possible is a sound strategy to enhance age domain generalizability. As we demonstrated, increasing P-Tr percentage definitely leads to improvements in age-domain generalizability (Fig. 5B). Nonetheless, this approach is still hampered by the aforementioned underlying issue of pediatric data paucity. Nevertheless, we demonstrate that utilizing what little pediatric data is available in conjunction with existing, widely available data from older subjects leads to increased generalizability as opposed to solely NSM-Tr. Our results align with the findings of Kumar *et al.,* whose study concluded that increasing age-diversity in training data improves age generalizability [39]. However, as previously mentioned, their study did not conduct a parametric investigation into the inclusion of varying proportions of pediatric training data. In contrast, our parametric analysis demonstrates that even a relatively small percentage of pediatric training data significantly improves segmentation outcomes. This quantitative insight provides better guidelines to practitioners to ensure sufficiently age-domain generalizable networks.

In principle, our study indicates that DL networks for segmentation trained exclusively on adult radiographs are not effective in segmenting pediatric radiographs. While our strategies of scaling-based augmentation do not lead to any significant improvements, we found that ensuring only 20 % of training data are from pediatric subjects leads to marked improvements in pediatric segmentation outcome without an adverse effect on NSM subjects. In fact, the minimal pediatric data approach led to performances reaching that of the positive controls. Simply put, our findings demonstrate increased age-domain generalizability in this particular semantic segmentation task. A potential means to improve this study could be to ground the scaling-based augmentations in real anatomical morphometrics capturing shape changes during growth, which we could not find in the existing literature. With some population based morphometrical data, we could define scaling parameters *a posteriori* as opposed to our current parametric approach. This could potentially lead to more realistic scaling augmentations and, thus, a better segmentation performance. Finally, one may investigate the use of deep generative models [61]. Implementing generative networks would enable the generation of pediatric-like image-mask pairs and developing these generative networks constitutes further works. Nevertheless, our study demonstrates the effectiveness of utilizing multi-scale filters for semantic segmentation tasks involving structures of interest with varying sizes. Our results, however, are restricted to this specific task of phalange segmentation from X-rays, and further work could seek to explore pediatric data generalizability on alternative anatomical structures and imaging modalities in both 2D and 3D (*e.g.*, MRI, CT, ultrasound).

5. Conclusions

To surmise, our study investigated the task of semantic segmentation of phalanges from radiographs using DL and sought to improve age-domain generalizability to pediatric patients. We found that, due to the varying shapes and sizes of each phalange, network architectures with multi-scale filters (*i.e.*, Inception U-Net and DeepLabv3+) led to better and more robust segmentation outcomes. Specifically, we found that standard U-Net only achieved a Mean IoU of 0.332 ± 0.335 whilst Inception U-Net and DeepLabv3 + achieved scores of 0.887 ± 0.062 and 0.899 ± 0.035 respectively. In improving the generalizability of DL networks to semantically segment pediatric data, we found that scaling-based augmentations have slight but not statistically significant improvements. Nevertheless, we found that substituting just 20 % of NSM training data with pediatric data leads to a 21.1 % improvement in pediatric segmentation outcomes, without a statistically significant effect on NSM segmentation outcomes. These findings provide quantitative justification and evidence of the necessity of age-diverse training data for DL models to ensure efficacy when utilized on pediatric subjects. While, pediatric data remains nevertheless elusive, ensuring a minimal inclusion of pediatric training data would lead to marked improvements in age-domain generalizability.

CRediT authorship contribution statement

Edwin Tay: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Amir A. Zadpoor:** Writing – review & editing, Supervision, Conceptualization. **Nazli Tümer:** Writing – review & editing, Supervision, 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.

Acknowledgments

The authors would like to thank the Radiological Society of North America for publishing the RSNA Pediatric Bone Age Challenge 2017 image dataset as open source, which we used in this study. This dataset is available at the following link (https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/RSNA-Pediatric-Bone-Age-Challenge-2017).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2024.107338.

Data availability

Data will be made available on request.

References

[1] P. Celard, E.L. Iglesias, J.M. Sorribes-Fdez, R. Romero, A.S. Vieira, L. Borrajo, A survey on deep learning applied to medical images: from simple artificial neural networks to generative models, Neural Comput. Appl. 35 (2023) 2291–2323, https://doi.org/10.1007/s00521-022-07953-4.

[2] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A.W. M. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, https://doi.org/10.1016/j.media.2017.07.005.

[3] M.H. Hesamian, W. Jia, X. He, P. Kennedy, Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges, J. Digit. Imaging 32 (2019) 582–596, https://doi.org/10.1007/s10278-019-00227-x.

[4] D. Shen, G. Wu, H.-I. Suk, Deep Learning in Medical Image Analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248, https://doi.org/10.1146/annurev-bioeng-071516-044442.

[5] M.I. Razzak, S. Naz, A. Zaib, Deep Learning for Medical Image Processing: Overview, Challenges and Future (2017), https://doi.org/10.48550/ARXIV.1704.06825.

[6] H. Greenspan, B. Van Ginneken, R.M. Summers, Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique, IEEE Trans. Med. Imaging 35 (2016) 1153–1159, https://doi.org/10.1109/TMI.2016.2553401.

[7] M. Hauptmann, G. Byrnes, E. Cardis, M.-O. Bernier, M. Blettner, J. Dabin, H. Engels, T.S. Istad, C. Johansen, M. Kaijser, K. Kjaerheim, N. Journy, J. M. Meulepas, M. Moissonnier, C. Ronckers, I. Thierry-Chef, L. Le Cornet, A. Jahnen, R. Pokora, M. Bosch De Basea, J. Figuerola, C. Maccia, A. Nordenskjold, R. W. Harbron, C. Lee, S.L. Simon, A. Berrington De Gonzalez, J. Schüz, A. Kesminiene, Brain cancer after radiation exposure from CT examinations of children and young adults: results from the EPI-CT cohort study, Lancet Oncol. 24 (2023) 45–53, https://doi.org/10.1016/S1470-2045(22)00655-6.

[8] S.M. Jung, Drug selection for sedation and general anesthesia in children undergoing ambulatory magnetic resonance imaging, Yeungnam Univ, J. Med. 37 (2020) 159–168, https://doi.org/10.12701/yujm.2020.00171.

[9] B.B. Thukral, Problems and preferences in pediatric imaging, Indian J. Radiol, Imaging 25 (2015) 359–364, https://doi.org/10.4103/0971-3026.169466.

[10] J. Downie, M. Schmidt, N. Kenny, R. D'Arcy, M. Hadskis, J. Marshall, Paediatric MRI Research Ethics: The Priority Issues, J. Bioethical Inq. 4 (2007) 85–91, https://doi.org/10.1007/s11673-007-9046-5.

[11] F.F. Alqahtani, F. Messina, A.C. Offiah, Are semi-automated software program designed for adults accurate for the identification of vertebral fractures in children? Eur. Radiol. 29 (2019) 6780–6789, https://doi.org/10.1007/s00330-019-06250-4.

[12] M. Drai, B. Testud, G. Brun, J.-F. Hak, D. Scavarda, N. Girard, J.-P. Stellmann, Borrowing strength from adults: Transferability of AI algorithms for paediatric brain and tumour segmentation, Eur. J. Radiol. 151 (2022) 110291, https://doi.org/10.1016/j.ejrad.2022.110291.

[13] R.C. Hardie, A.T. Trout, J.R. Dillman, B.N. Narayanan, A.A. Tanimoto, Performance Analysis in Children of Traditional and Deep Learning CT Lung Nodule Computer-Aided Detection Systems Trained on Adults, Am. J. Roentgenol. 222 (2024) e2330345.

[14] B.J. Nelson, P. Kc, A. Badal, L. Jiang, S.C. Masters, R. Zeng, Pediatric evaluations for deep learning CT denoising, Med. Phys. 51 (2024) 978–990, https://doi.org/10.1002/mp.16901.

[15] J. Lee, C. Park, M. Cho, Y.H. Choi, J.H. Kim, Age-dependent generalizability of lumbar spine detection and segmentation models: a comparative study in pediatric populations, in: O. Colliot, J. Mitra (Eds.), Med. Imaging 2024 Image Process., SPIE, San Diego, United States, 2024: p. 74. Doi: 10.1117/12.3006168.

[16] Z.Y. Hamd, E.G. Osman, A.I. Alorainy, A.F. Alqahtani, N.R. Alshammari, O. Bajamal, S.H. Alruwaili, S.S. Almohsen, R.I. Almusallam, M.U. Khandaker, The role of machine learning in detecting primary brain tumors in Saudi pediatric patients through MRI images, J. Radiat. Res. Appl. Sci. 17 (2024) 100956, https://doi.org/10.1016/j.jrras.2024.100956.

[17] S.C. Shelmerdine, R.D. White, H. Liu, O.J. Arthurs, N.J. Sebire, Artificial intelligence for radiological paediatric fracture assessment: a systematic review, Insights Imaging 13 (2022) 94, https://doi.org/10.1186/s13244-022-01234-3.

[18] J. Chan, S.C. Raju, E. Topol, Towards a tricorder for diagnosing paediatric conditions, The Lancet 394 (2019) 907, https://doi.org/10.1016/S0140-6736(19)32087-2.

[19] C. Pringle, J.-P. Kilday, I. Kamaly-Asl, S.M. Stivaros, The role of artificial intelligence in paediatric neuroradiology, Pediatr. Radiol. 52 (2022) 2159–2172, https://doi.org/10.1007/s00247-022-05322-w.

[20] N. Davendralingam, N.J. Sebire, O.J. Arthurs, S.C. Shelmerdine, Artificial intelligence in paediatric radiology: Future opportunities, Br. J. Radiol. 94 (2021) 20200975, https://doi.org/10.1259/bjr.20200975.

[21] K. Iyer, A. Morris, B. Zenger, K. Karanth, N. Khan, B.A. Orkild, O. Korshak, S. Elhabian, Statistical shape modeling of multi-organ anatomies with shared boundaries, Front. Bioeng. Biotechnol. 10 (2023) 1078800, https://doi.org/10.3389/fbioe.2022.1078800.

[22] A. Saito, S. Nawano, A. Shimizu, Joint optimization of segmentation and shape prior from level-set-based statistical shape model, and its application to the automated segmentation of abdominal organs, Med. Image Anal. 28 (2016) 46–65, https://doi.org/10.1016/j.media.2015.11.003.

[23] J.J. Cerrolaza, M.L. Picazo, L. Humbert, Y. Sato, D. Rueckert, M.Á.G. Ballester, M. G. Linguraru, Computational anatomy for multi-organ analysis in medical imaging: A review, Med. Image Anal. 56 (2019) 44–67, https://doi.org/10.1016/j.media.2019.04.002.

[24] Y. Fu, Y. Lei, T. Wang, W.J. Curran, T. Liu, X. Yang, A review of deep learning based methods for medical image multi-organ segmentation, Phys. Med. 85 (2021) 107–122, https://doi.org/10.1016/j.ejmp.2021.05.003.

[25] S. Asgari Taghanaki, K. Abhishek, J.P. Cohen, J. Cohen-Adad, G. Hamarneh, Deep semantic segmentation of natural and medical images: a review, Artif. Intell. Rev. 54 (2021) 137–178, https://doi.org/10.1007/s10462-020-09854-1.

[26] D.-S.-R. Maru, R. Schwarz, J. Andrews, S. Basu, A. Sharma, C. Moore, Turning a blind eye: the mobilization of radiology services in resource-poor regions, Glob. Health 6 (2010) 18, https://doi.org/10.1186/1744-8603-6-18.

[27] G. Frija, I. Blažić, D.P. Frush, M. Hierath, M. Kawooya, L. Donoso-Bach, B. Brkljačić, How to improve access to medical imaging in low- and middle-income countries ?, eClinicalMedicine 38 (2021) 101034. Doi: 10.1016/j.eclinm.2021.101034.

[28] S.M. Ryu, K. Shin, S.W. Shin, S. Lee, N. Kim, Enhancement of evaluating flatfoot on a weight-bearing lateral radiograph of the foot with U-Net based semantic segmentation on the long axis of tarsal and metatarsal bones in an active learning manner, Comput. Biol. Med. 145 (2022) 105400, https://doi.org/10.1016/j.compbiomed.2022.105400.

[29] C. Wang, Segmentation of Multiple Structures in Chest Radiographs Using Multi-task Fully Convolutional Networks, in: P. Sharma, F.M. Bianchi (Eds.), Image Anal, Springer International Publishing, Cham, 2017, pp. 282–289, https://doi.org/10.1007/978-3-319-59129-2_24.

[30] G. Holste, R.P. Sullivan, M. Bindschadler, N. Nagy, A. Alessio, Multi-class semantic segmentation of pediatric chest radiographs, in: B.A. Landman, I. Išgum (Eds.), Med. Imaging 2020 Image Process., SPIE, Houston, United States, 2020: p. 49. Doi: 10.1117/12.2544426.

[31] Y. Liu, X. Zhang, G. Cai, Y. Chen, Z. Yun, Q. Feng, W. Yang, Automatic delineation of ribs and clavicles in chest radiographs using fully convolutional DenseNets, Comput. Methods Programs Biomed. 180 (2019) 105014, https://doi.org/10.1016/j.cmpb.2019.105014.

[32] Y. Lv, J. Wang, W. Wu, Y. Pan, Performance comparison of deep learning methods on hand bone segmentation and bone age assessment, in, Int. Conf. Cult.-Oriented Sci, Technol. Cost, IEEE, Lanzhou, China 2022 (2022) 375–380, https://doi.org/10.1109/CoST57098.2022.00083.

[33] E.L. Siegel, What Can We Learn from the RSNA Pediatric Bone Age Machine Learning Challenge? Radiology 290 (2019) 504–505, https://doi.org/10.1148/radiol.2018182657.

[34] K. Hatano, S. Murakami, H. Lu, J.K. Tan, H. Kim, T. Aoki, Detection of Phalange Region Based on U-Net, in 18th Int, Conf. Control Autom. Syst. ICCAS 2018 (2018) 1338–1342.

[35] K. Kawagoe, K. Hatano, S. Murakami, H. Lu, H. Kim, T. Aoki, Automatic Segmentation Method of Phalange Regions Based on Residual U-Net and MSGVF Snakes in 19th Int, Conf. Control Autom. Syst. ICCAS 2019 (2019) 1046–1049, https://doi.org/10.23919/ICCAS47443.2019.8971740.

[36] H. Ono, S. Murakami, T. Kamiya, T. Aoki, Automatic Segmentation of Finger Bone Regions from CR Images Using Improved DeepLabv3+ in 21st Int. Conf. Control Autom, Syst. ICCAS, IEEE, Jeju, Korea, Republic of 2021 (2021) 1788–1791, https://doi.org/10.23919/ICCAS52745.2021.9649864.

[37] A. Boutillon, P.-H. Conze, C. Pons, V. Burdin, B. Borotikar, 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, https://doi.org/10.1016/j.media.2022.102556.

[38] S. Rajaraman, F. Yang, G. Zamzmi, Z. Xue, S. Antani, Can deep adult lung segmentation models generalize to the pediatric population? Expert Syst. Appl. 229 (2023) 120531 https://doi.org/10.1016/j.eswa.2023.120531.

[39] K. Kumar, A.U. Yeo, L. McIntosh, T. Kron, G. Wheeler, R.D. Franich, Deep Learning Auto-Segmentation Network for Pediatric Computed Tomography Data Sets: Can We Extrapolate From Adults? Int. J. Radiat. Oncol. 119 (2024) 1297–1306, https://doi.org/10.1016/j.ijrobp.2024.01.201.

[40] E. Somasundaram, Z. Taylor, V.V. Alves, L. Qiu, B.L. Fortson, N. Mahalingam, J. A. Dudley, H. Li, S.L. Brady, A.T. Trout, J.R. Dillman, Deep Learning Models for Abdominal CT Organ Segmentation in Children: Development and Validation in Internal and Heterogeneous Public Datasets, Am. J. Roentgenol. 223 (2024) e2430931.

[41] C. Shorten, T.M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, J. Big Data 6 (2019) 60, https://doi.org/10.1186/s40537-019-0197-0.

[42] L. Perez, J. Wang, The Effectiveness of Data Augmentation in Image Classification using Deep, Learning (2017), https://doi.org/10.48550/ARXIV.1712.04621.

[43] T. Stern, R. Aviram, C. Rot, T. Galili, A. Sharir, N. Kalish Achrai, Y. Keller, R. Shahar, E. Zelzer, Isometric Scaling in Developing Long Bones Is Achieved by an Optimal Epiphyseal Growth Balance, PLOS Biol. 13 (2015) e1002212.

[44] S.S. Halabi, L.M. Prevedello, J. Kalpathy-Cramer, A.B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L.A. Pereira, R.T. Sousa, N. Abdala, F.C. Kitamura, H. H. Thodberg, L. Chen, G. Shih, K. Andriole, M.D. Kohli, B.J. Erickson, A. E. Flanders, The RSNA Pediatric Bone Age Machine Learning Challenge, Radiology 290 (2019) 498–503, https://doi.org/10.1148/radiol.2018180736.

[45] J.M. Tanner, N. Cameron, Assessment of skeletal maturity and prediction of adult height (TW3 method), 3. ed, Saunders, London, 2001.

[46] Y. Yoshimi, Y. Mine, S. Ito, S. Takeda, S. Okazaki, T. Nakamoto, T. Nagasaki, N. Kakimoto, T. Murayama, K. Tanimoto, Image preprocessing with contrast-limited adaptive histogram equalization improves the segmentation performance of deep learning for the articular disk of the temporomandibular joint on magnetic resonance images, Oral Surg, Oral Med. Oral Pathol. Oral Radiol. 138 (2024) 128–141, https://doi.org/10.1016/j.oooo.2023.01.016.
[47] K. Wada, Mpitid, M. Buijs, N. Zhang Ch., なるみ, Bc. Martin Kubovčík, A. Myczko, Latentix, Lingjie Zhu, N. Yamaguchi, S. Fujii, Iamgd67, IlyaOvodov, Akshar Patel, C. Clauss, Eisoku Kuroiwa, R. Iyengar, S. Shilin, T. Malygina, K. Kawaharazuka, J. Engelberts, A. J, AlexMa, Changwoo Song, Charlie, D. Rose, D. Livingstone, Doug, Erik, H. Toft, wkentaro/labelme: v4.6.0, (2021). Doi: 10.5281/ZENODO.5711226.
[48] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation in IEEE Conf. Comput. Vis. Pattern Recognit, CVPR, IEEE, Boston, MA, USA 2015 (2015) 3431–3440, https://doi.org/10.1109/CVPR.2015.7298965.
[49] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Med. Image Comput. Comput.-Assist. Interv. – MICCAI 2015, Springer International Publishing, Cham, 2015: pp. 234–241. Doi: 10.1007/978-3-319-24574-4_28.
[50] N. Siddique, S. Paheding, C.P. Elkin, V. Devabhaktuni, U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications, IEEE Access 9 (2021) 82031–82057, https://doi.org/10.1109/ACCESS.2021.3086020.
[51] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Comput. Vis. – ECCV 2018, Springer International Publishing, Cham, 2018: pp. 833–851. Doi: 10.1007/978-3-030-01234-2_49.
[52] C. Szegedy, W. Liu, P. Yangqing Jia, S. Sermanet, D. Reed, D. Anguelov, V. Erhan, A.R. Vanhoucke, Going deeper with convolutions in IEEE Conf. Comput. Vis. Pattern Recognit, CVPR, IEEE, Boston, MA, USA 2015 (2015) 1–9, https://doi.org/10.1109/CVPR.2015.7298594.
[53] I. Delibasoglu, M. Cetin, Improved U-Nets with inception blocks for building detection, J. Appl. Remote Sens. 14 (2020), https://doi.org/10.1117/1.JRS.14.044512.
[54] D.E. Cahall, G. Rasool, N.C. Bouaynaya, H.M. Fathallah-Shaykh, Inception Modules Enhance Brain Tumor Segmentation, Front. Comput. Neurosci. 13 (2019) 44, https://doi.org/10.3389/fncom.2019.00044.
[55] S.R. Ravichandran, B. Nataraj, S. Huang, Z. Qin, Z. Lu, A. Katsuki, W. Huang, Z. Zeng, 3D Inception U-Net for Aorta Segmentation using Computed Tomography Cardiac Angiography in IEEE EMBS Int. Conf. Biomed. Health Inform, BHI, IEEE, Chicago, IL, USA 2019 (2019) 1–4, https://doi.org/10.1109/BHI.2019.8834582.
[56] Y. Ho, S. Wookey, The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling, IEEE Access 8 (2020) 4806–4813, https://doi.org/10.1109/ACCESS.2019.2962617.
[57] E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss, J.P. Cunningham, Uses and Abuses of the Cross-Entropy Loss: Case Studies in Modern Deep, Learning (2020), https://doi.org/10.48550/ARXIV.2011.05231.
[58] D.H.P.C. Centre (DHPC), DelftBlue Supercomputer (Phase 2), (2024). https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2.
[59] T. Ganokratanaa, S. Aramvith, Generative adversarial network for video anomaly detection, in: Gener. Advers. Netw. Image–Image Transl., Elsevier, 2021: pp. 377–420. Doi: 10.1016/B978-0-12-823519-5.00011-7.
[60] L. Ding, K. Zhao, X. Zhang, X. Wang, J. Zhang, A Lightweight U-Net Architecture Multi-Scale Convolutional Network for Pediatric Hand Bone Segmentation in X-Ray Image, IEEE Access 7 (2019) 68436–68445, https://doi.org/10.1109/ACCESS.2019.2918205.
[61] J. Xu, H. Li, S. Zhou, An Overview of Deep Generative Models, IETE Tech. Rev. 32 (2015) 131–139, https://doi.org/10.1080/02564602.2014.987328.