Semi-Supervised Self-Taught Deep Learning for Finger Bones Segmentation

Ziyuan Zhao[*1,2], Xiaoman Zhang[*1,2], Cen Chen[1], Wei Li[4], Songyou Peng[1]
Jie Wang[1], Xulei Yang[3], Le Zhang[1] and Zeng Zeng[1]
[1]Institute for Infocomm Research, A*STAR, Singapore
[2]National University of Singapore, Singapore
[3] YITU Tech, Singapore
[4] Huazhong University of Science and Technology, China

Abstract—Segmentation stands at the forefront of many high-level vision tasks. In this study, we focus on segmenting finger bones within a newly introduced semi-supervised self-taught deep learning framework which consists of a student network and a stand-alone teacher module. The whole system is boosted in a life-long learning manner wherein each step the teacher module provides a refinement for the student network to learn with newly unlabeled data. Experimental results demonstrate the superiority of the proposed method over conventional supervised deep learning methods.

I. INTRODUCTION

We study a fundamental problem of finger bones segmentation which has broad applications in clinical practice such as bone age assessment [1], anomaly detection [2] and so on. Compared with other segmentation tasks [3] in the computer vision community, finger bones segmentation is challenging due to to the variation of hand bones and scarceness of large-scale well-annotated dataset.

As other pixel-wise regression tasks in computer vision [4], image segmentation is an important area of research in medical image analysis where many attempts and success have been made so far. In the domain of medical image analysis, on the one hand, thresholding based segmentation methods were widely applied. These methods assume that the background and foreground of an image have distinct intensity range. Pietka *et al.* [5] used the Sobel gradient to segment bones with soft tissue region. Similar techniques like Derivative of Gaussian (DoG) and dynamic thresholding have also been extensively studied for hand bone images [6]–[8]. However, the performance of these methods is not satisfying on the radiographs without bi-model histogram. On the other hand, clustering-based approaches have also been applied on segmentation [9]–[11] and they are shown to perform well on images with a low-intensity range. However, it is difficult to generate relatively stable masks for different bones due to significant variations in the maturity of bone and image quality [12].

Collecting a large scale dataset containing all challenging scenarios of fingers, as a common practice in the vision community, may partially alleviate the mentioned challenges.

However, manual labeling is costly, time-consuming, error-prone, and requires massive human intervention for each new task. This is often impractical, especially for clinical practices due to some privacy concerns. This further motivates us to study the following question: *Is it possible to improve the finger bone segmentation performance with the unlabeled dataset?*

In this paper, we propose a novel self-taught deep learning framework which consists of a student network and a stand-alone teacher module. The U-Net is firstly trained on a small subset of well-annotated training images and then boosted in a life-long learning manner. In each step, the teacher module provides a refinement for the student network to learn with newly unlabeled data. Experimental results show that the proposed method outperformed conventional supervised deep learning methods.

II. RELATED WORK

In recent years, the number of success stories of segmentation has seen an all-time rise [13], [14]. The unifying idea behind all of the above is deep learning, the utilization of neural networks [15] with many hidden layers, for the purposes of learning complex feature representations from raw data, rather than relying on handcrafted feature extraction. They have shown consistent improvements over their non-deep counterparts across many tasks beyond segmentation [16], [17]. Those approaches usually adopt an "encoder-decoder" structure which could gradually decrease the resolution of the input with the depth of the network in the encoding stage, and then up-sampling and skip connections are applied to recover the resolution of the input in the decoding stage. So far, much of the recent research on segmentation using this kind of network has been made [18]–[21]. In particular, U-Net [19] is the most widely used in biomedical image segmentation. It concatenates multi-scale feature maps in the encoding stage to upsample feature maps in the decoding stage. This design helps to generate richer feature hierarchies and achieve outstanding performance under the condition of very few annotated images. Although much progress has been made, the results from deep networks are not ideal from a practical point of view because the boundary information

---

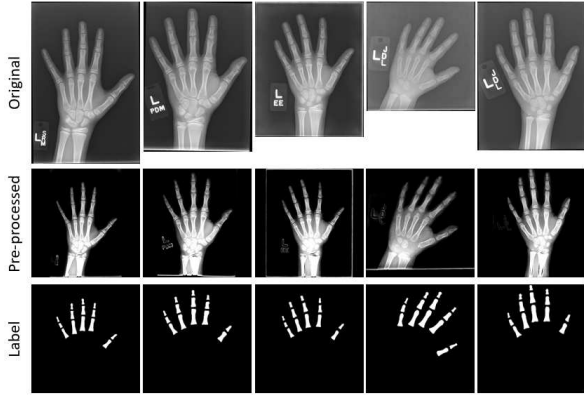* The first two authors contribute equally in this work.

Fig. 1. Examples of preprocessing:(first row) original images; (second row) preprocessed images; (third row) binary finger bone masks

of the resulting segmentation maps may be lost for some complex background. To address this, the predicted results are usually further refined with an additional graphical model, for instance, a fully connected CRF, in which, pixels treated as nodes are pairwise connected with each other [22].

III. METHODOLOGY

In this section, we explore the semi-supervised self-taught deep learning pipeline proposed for pseudo-labeling and finer segmentation.

A. Preprocessing and Augmentation

Intensity normalization was first applied by subtracting the mean and then dividing by the standard deviation to improve the brightness and augment the edges of bones. After that, radiographs were cropped and scaled to a fixed size with 600 x 600 pixels. Finally, the ground truth masks of these 209 radiographs selected from Section IV-A were annotated manually for the following experiments, in which, phalanges(Distal, Middle, Proximal) were annotated. Some examples are shown in Fig 1. To enhance the data, the images were randomly rotated by ± 20 degrees, shifted in width and height by 0.05, zoomed by 0.05, sheared by 0.05 and flipped horizontally.

B. Self-taught Learning Pipeline

As shown in Fig 2, the proposed system consists of a student module, which is realized by a Deep U-Net, and a teacher module embodied by a dense CRF. First, the U-Net is trained firstly on a small set of the well-annotated training set and makes predictions on the unlabeled. After that, a stand-alone dense CRF module is utilized to make refinements and provide pseudo-label for the U-Net.

More specifically, every pixel i, which is regarded as a node, has a label x_i and an observation value y_i, and the relationship among pixels are regarded as edges. The labels behind pixels can be inferred by observations y_i, and the dense CRF I is characterized by a Gibbs distribution,

$$P\left(X = x | I\right) = \frac{1}{Z(I)} exp(-E(x|I)) \tag{1}$$

where $E(x|I)$ is the Gibbs energy of a label x, which is formulated as follows,

$$E(x) = \sum_i \Psi_u\left(x_i\right) + \sum_{i<j} \Psi_p(x_i, x_j) \tag{2}$$

among which, the unary potential function $\Psi_u\left(x_i\right)$ is donated by the output of U-Net, and the pairwise potentials in our model is given by

$$\Psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{M} w^{(m)} k_G^{(m)}(f_i, f_j) \tag{3}$$

where each $k_G^{(m)}$ is a Gaussian kernel $k_m(f_i, f_j)$, the vectors f_i and f_j are feature vectors for pixels i and j, $w^{(m)}$ are linear combination weights, and μ is a label compatibility function.

Our system iterates in a "curriculum Learning" manner in which easier samples are firstly chosen to improve the U-Net and difficult samples are gradually included. In each iteration, we choose N easiest samples and use the results from dense CRF as the pseudo-label for U-Net. More specifically, we calculate the Dice's Coefficient, as defined in Eq 4 where X and Y are the cardinalities of the two sets, between the input and output of dense CRF. Dice's Coefficient serves as a proxy of the difficulty for each sample. A larger Dice's Coefficient indicates that U-Net performs relatively well because no significant refinements are given by dense CRF. In this way, more reliable information is first utilized in the system, which provides better audiences for the U-Net to learn.

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{4}$$

IV. EXPERIMENTS

A. Dataset and Evaluation Protocol

The dataset used in this paper is from the 2017 Pediatric Bone Age Prediction Challenge [23] organized by the Radiological Society of North America (RSNA). It consists of 12611 radiographs of the hand containing 6833 images of male and 5778 images of the female. The overall age distribution of this dataset is severely imbalanced, as shown in Fig 3. And Fig 4 shows some examples of radiographs in the dataset, and hand bones differ significantly from size, brightness, orientation and contrast across the samples. Besides, some artifacts are shown on the radiographs, such as watches and plaster casts.

Due to the size of the dataset and time consuming, small batch of the dataset is applied to our experiments to validate the effectiveness of the method proposed. To have a suitable generalization of the segmentation, the dataset was grouped in 19 year-based age groups (0-1, 1-2, 2-3, ...,18-19), then 11 cases were selected from each year group to form the small subset of Training Set ($11 \times 19 = 209$) for experiments.Finally, the dataset is split a into a train (139) /validation (20) /test (50) set randomly.

To validate the effectiveness of the pipeline we proposed, experiments were done in 209 images with ground truth masks, which is further randomly split into 4 subsets: training,
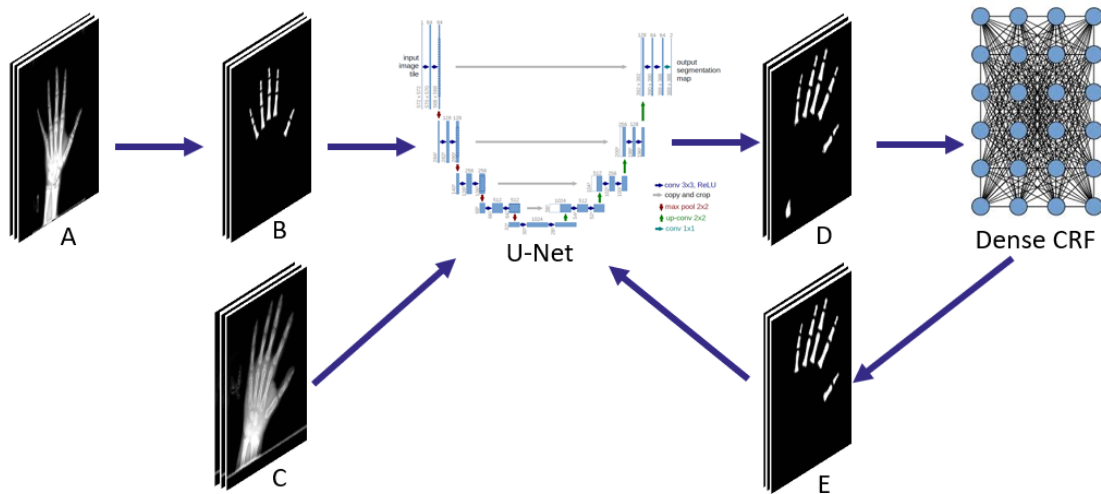
Fig. 2. The iterative procedure of self-taught learning utilizing U-Net and Dense CRF for pseudo-labeling: (A) preprocessed input data; (B) masks manually labeled; (C) new data; (D) raw prediction; (E) refine prediction. The system boosts itself in a Curriculum Learning manner. More specifically, in each iteration, the CRF refines the raw predictions of U-Net and return the results from the easiest samples therein as pseudo-labels for future learning.
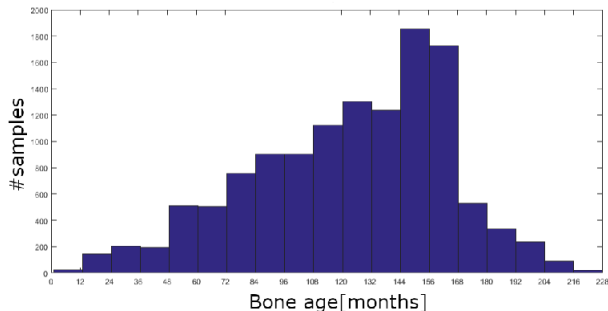

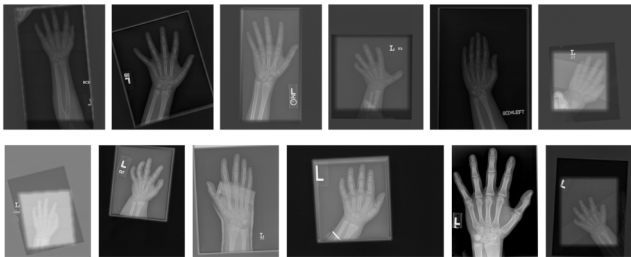Fig. 3. Distribution of estimated bone age.

Fig. 4. Example of radiographs in the dataset in different size, contrast,brightness.And some shows additional artifacts.

validation, test and one 50-size subset prepared to enhance the training data. The U-Net has an input size of 512×512 pixels and is optimized with Dice's Coefficient. Hyper-parameters of dense CRF are tuned on the training set (detailed shown in Table I). In this study, we set $N = 10$ and compare our system with the three following baselines:

1) U-Net-89: U-Net trained with 89 training samples.
2) U-Net-139: U-Net trained with 139 training samples.
3) U-Net-Only: U-Net trained with 89 training samples.

After that, we randomly select 10 results of U-Net as the pseudo labels.

TABLE I
HYPERPARAMETERS OF DENSE CRF

	sdims	schan	compat	step
PairwiseEnergy	10	10	3	-
PairwiseGaussian	-	-	3	-
inference	-	-	-	50

B. Results and Discussions

An Overview of the final results in test data is shown in Table II and Fig 5. The first column shows the results of selecting top 10 Dice's Coefficient masks in each iteration. And for the second column, in which, we just implemented U-Net without post-processing. Last two columns stand for the best U-Net performance with 89 and 139 ground truth masks respectively. Results show that top-DSC has the best performance with 89 ground truth masks and 50 pseudo labels. More interestingly, it even outperforms the results of U-Net trained with all the 139 well-annotated images. We hypothesise that the following two mechanisms contribute this phenomenon:1) the "curriculum Learning" strategy reduces the risk of local optimal, and 2) pseudo-labels provided by dense CRF insert certain level of noises into the learning process. This has been reported to be effective in reducing the risk of overfitting [24]. In addition, we also observe that the proposed method still shows an uptrend by the end, while U-Net-only modeling is almost vibrating.

V. CONCLUSIONS

This paper explores a novel semi-supervised self-taught deep learning method for finger bones segmentation. The proposed method utilizes a deep U-Net as the student module

TABLE II
RESULTS OF EXPERIMENTS

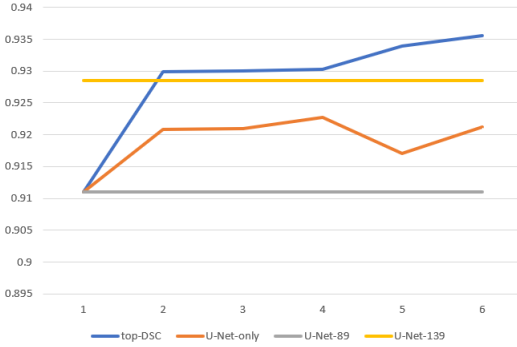| Iteration	top-DSC	U-Net-only	U-Net-89	U-Net-139
1	0.911	0.911	0.911	-
2	0.9299	0.92088	-	-
3	0.93008	0.92093	-	-
4	0.93031	0.92279	-	-
5	0.93393	0.91709	-	-
6	0.93562	0.92123	-	0.92853

Fig. 5. Iterative procedure of self-taught learning utilizing U-Net and Dense CRF for pseudo-labeling: (A) preprocessed input data; (B) masks manually labeled; (C) new data; (D) raw prediction; (E) refine prediction

and a dense CRF as a teacher module. The student module is first initialized on a limited number of training set. Then the system is able to boost itself in a "Curriculum Learning" manner. More specifically, in each iteration, the CRF refines the raw predictions of U-Net and return the results from the easiest samples therein as pseudo-labels for future learning. Experimental results show that the proposed method outperforms conventional supervised learning approaches.

ACKNOWLEDGMENT

REFERENCES

[1] Arkadiusz Gertych, Aifeng Zhang, James Sayre, Sylwia Pospiech-Kurkowska, and HK Huang, "Bone age assessment of children using a digital hand atlas," *Computerized medical imaging and graphics*, vol. 31, no. 4-5, pp. 322–331, 2007.

[2] Alfonso Rojas Dominguez and Asoke K Nandi, "Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection," *Computerized Medical Imaging and Graphics*, vol. 32, no. 4, pp. 304–315, 2008.

[3] Yun Liu, Peng-Tao Jiang, Vahan Petrosyan, Shi-Jie Li, Jiawang Bian, Le Zhang, and Ming-Ming Cheng, "Del: Deep embedding learning for efficient image segmentation.," in *IJCAI*, 2018, vol. 864, p. 870.

[4] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai, "Richer convolutional features for edge detection," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 5872–5881.

[5] Ewa Pietka, Arkadiusz Gertych, Sylwia Pospiech, Fei Cao, HK Huang, and Vicente Gilsanz, "Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal roi extraction," *IEEE transactions on medical imaging*, vol. 20, no. 8, pp. 715–729, 2001.

[6] Ewa Pietka, Lotfi Kaabi, ML Kuo, and HK Huang, "Feature extraction in carpal-bone analysis," *IEEE transactions on medical imaging*, vol. 12, no. 1, pp. 44–49, 1993.

[7] BS Sharif, SA Zaroug, EG Chester, JP Owen, and EJ Lee, "Bone edge detection in hand radiographic images," in *Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*. IEEE, 1994, vol. 1, pp. 514–515.

[8] Daniela Giordano, Rosalia Leonardi, Francesco Maiorana, Giacomo Scarciofalo, and Concetto Spampinato, "Epiphysis and metaphysis extraction and classification by adaptive thresholding and dog filtering for automated skeletal bone age analysis," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. IEEE, 2007, pp. 6551–6556.

[9] Aifeng Zhang, Arkadiusz Gertych, and Brent J Liu, "Automatic bone age assessment for young children from newborn to 7-year-old using carpal bones," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4-5, pp. 299–310, 2007.

[10] Antonio Tristán-Vega and Juan Ignacio Arribas, "A radius and ulna tw3 bone age assessment system," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 5, pp. 1463–1476, 2008.

[11] Daniela Giordano, Concetto Spampinato, Giacomo Scarciofalo, and Rosalia Leonardi, "An automatic system for skeletal bone age measurement by robust processing of carpal and epiphysial/metaphysial bones," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 10, pp. 2539–2553, 2010.

[12] Tom Van Steenkiste, Joeri Ruyssinck, Olivier Janssens, Baptist Vandersmissen, Florian Vandecasteele, Pieter Devolder, Eric Achten, Sofie Van Hoecke, Dirk Deschrijver, and Tom Dhaene, "Automated assessment of bone age using deep learning and gaussian process regression," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 674–677.

[13] Zenglin Shi, Guodong Zeng, Le Zhang, Xiahai Zhuang, Lei Li, Guang Yang, and Guoyan Zheng, "Bayesian voxdrn: A probabilistic deep voxelwise dilated residual network for whole heart segmentation from 3d mr images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 569–577.

[14] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng, "Crowd counting with deep negative correlation learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5382–5390.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[16] Le Zhang, Songyou Peng, and Stefan Winkler, "PersEmoN: A deep network for joint analysis of apparent personality, emotion and their relationship," *arXiv preprint arXiv:1811.08657*, 2018.

[17] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[20] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5168–5177.

[21] Zeng Zeng, Nanying Liang, Xulei Yang, and Steven Hoi, "Multi-target deep neural networks: Theoretical analysis and implementation," *Neurocomputing*, vol. 273, pp. 634–642, 2018.

[22] Philipp Krähenbühl and Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[23] David B Larson, Matthew C Chen, Matthew P Lungren, Safwan S Halabi, Nicholas V Stence, and Curtis P Langlotz, "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, vol. 287, no. 1, pp. 313–322, 2017.

[24] Ye Ren, Le Zhang, and Ponnuthurai N Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE Comp. Int. Mag.*, vol. 11, no. 1, pp. 41–53, 2016.