Fine-Grained Hand Bone Segmentation via Adaptive Multi-dimensional Convolutional Network and Anatomy-Constraint Loss

Bolun Zeng[1], Li Chen[2], Yuanyi Zheng[2], Ron Kikinis[3], and Xiaojun Chen[1(✉)]

[1] Institute of Biomedical Manufacturing and Life Quality Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China
xiaojunchen@sjtu.edu.cn

[2] Department of Ultrasound in Medicine, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3] The Surgical Planning Laboratory, Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, USA

Abstract. Ultrasound imaging is a promising tool for clinical hand examination due to its radiation-free and cost-effective nature. To mitigate the impact of ultrasonic imaging defects on accurate clinical diagnosis, automatic fine-grained hand bone segmentation is highly desired. However, existing ultrasound image segmentation methods face difficulties in performing this task due to the presence of numerous categories and insignificant inter-class differences. To address these challenges, we propose a novel Adaptive Multi-dimensional Convolutional Network (AMCNet) for fine-grained hand bone segmentation. It is capable of dynamically adjusting the weights of 2D and 3D convolutional features at different levels via an adaptive multi-dimensional feature fusion mechanism. We also design an anatomy-constraint loss to encourage the model to learn anatomical relationships and effectively mine hard samples. Experiments demonstrate that our method outperforms other comparison methods and effectively addresses the task of fine-grained hand bone segmentation in ultrasound volume. We have developed a user-friendly and extensible module on the 3D Slicer platform based on the proposed method and will release it globally to promote greater value in clinical applications. The source code is available at https://github.com/BL-Zeng/AMCNet.

Keywords: Hand bone segmentation · Adaptive convolution · Anatomical constraint · 3D Slicer · Ultrasound images

1 Introduction

Hand imaging examination is a standard clinical procedure commonly utilized for various medical purposes such as predicting biological bone age [16] and

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_38.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 395–404, 2023.
https://doi.org/10.1007/978-3-031-43901-8_38

diagnosing finger bone and joint diseases [4,8]. Ultrasound (US) is a promising alternative imaging modality for clinical examinations due to its radiation-free and cost-effective nature, especially the three-dimensional (3D) US volume, which is increasingly preferred for its intuitive visualization and comprehensive clinical information. However, the current US imaging technology is limited by low signal-to-noise ratio and inherent imaging artifacts, making the examination of hand with complex and delicate anatomical structure highly dependent on high-level expertise and experience.

To address this challenge, deep learning-based US image segmentation methods have been explored. For instance, Liu et al. [10] propose an attention-based network to segment seven key structures in the neonatal hip bone. Rahman et al. [14] present a graph convolutional network with orientation-guided supervision to segment bone surfaces. Studies such as [2,11] use the convolutional-based network for efficient bone surface segmentation. Additionally, some studies have focused on the automatic identification and segmentation of soft tissues, such as finger tendons and synovial sheaths [9,12]. Although these methods are effective in segmenting specific objects, they lack fine-grained analysis. Some studies have revealed that each hand bone has clinical analysis value [1], thus making fine-grained segmentation clinically significant. However, this is a challenging task. The hand comprises numerous structures, with a closely related anatomical relationship between the phalanges, metacarpal bones, and epiphysis. Moreover, different categories exhibit similar imaging features, with the epiphysis being particularly indistinguishable from the phalanges and metacarpal bones.

Fine-grained segmentation demands a model to maintain the inter-slice anatomical relationships while extracting intra-slice detailed features. The 2D convolution excels at capturing dense information but lacks inter-slice information, while the 3D convolution is complementary [6]. This motivates us to develop a proper adaptive fusion method. Previous studies used 2D/3D layouts to address data anisotropy problems. For example, Wang et al. [17] propose a 2.5D UNet incorporating 2D and 3D convolutions to improve the accuracy of MR image segmentation. Dong et al. [6] present a mesh network fusing multi-level features for better anisotropic feature extraction. However, the effectiveness of these methods in capturing fine-grained feature representations is limited due to their relatively fixed convolution distributions and feature fusion approaches. Moreover, the lack of supervision on complex anatomical relationships makes them inevitably suffer from anatomical errors such as missing or confusing categories.

To overcome the deficiencies of existing methods, this study proposes a novel Adaptive Multi-dimensional Convolutional Network (AMCNet) with an anatomy-constraint loss for fine-grained hand bone segmentation. Our contribution is three-fold. 1) First, to the best of our knowledge, this is the first work to address the challenge of automatic fine-grained hand bone segmentation in 3D US volume. We propose a novel multi-dimensional network to tackle the issue of multiple categories and insignificant feature differences. 2) Second, we propose an adaptive multi-dimensional feature fusion mechanism to dynamically adjust the weights of 2D and 3D convolutional feature layers according to different objectives, thus improving the fine-grained feature representation of the model.

3) Finally, we propose an anatomy-constraint loss that minimizes the anatomical error and mines hard samples, further improving the performance of the model.

2 Methods

2.1 Network Design

As shown in Fig. 1, the architecture of AMCNet consists of four down-sampling layers, four up-sampling layers, and four skip-connections. Each layer contains an adaptive 2D/3D convolutional module (ACM) which is proposed to dynamically balance inter-layer and intra-layer feature weight through adaptive 2D and 3D convolutions for better representations. In the encoder, each ACM block is followed by a max-pooling layer to compress features. In the decoder, the trilinear interpolate is used to up-sample features. The number of channels across each layer is empirically set to 64, 128, 256, and 512. The output layer uses the 1×1×1 convolutional layer to obtain the segmentation map.

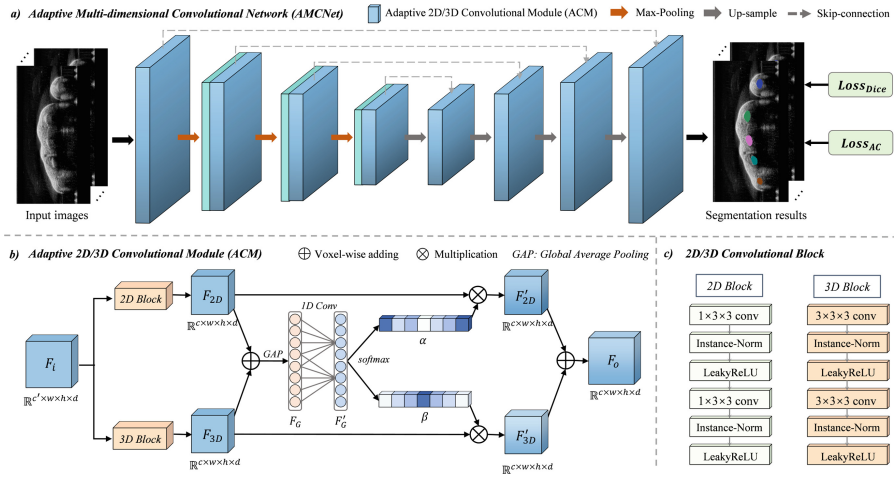

Fig. 1. An overview of the proposed AMCNet. The network consists of four down-sampling layers and four up-sampling layers. Each layer is composed of an ACM to fuse 2D and 3D convolutional features at different levels. $Loss_{Dice}$ denotes the Dice loss and $Loss_{AC}$ denotes the proposed anatomy-constraint loss.

2.2 Adaptive 2D/3D Convolutional Module (ACM)

To enable the model to capture the inter-layer anatomical connection and intra-layer dense semantic feature, the ACM is proposed to adaptively fuse the 2D and 3D convolution at different levels. Figure 1(b) illustrates the architecture of the ACM. Firstly, the feature map $F_i \in \mathbb{R}^{c' \times w \times h \times d}$ passes through 2D and 3D convolutional block respectively to obtain the 2D convolutional feature $F_{2D} \in \mathbb{R}^{c \times w \times h \times d}$ and 3D convolutional feature $F_{3D} \in \mathbb{R}^{c \times w \times h \times d}$. Figure 1(c) shows the details of the 2D and 3D convolutional block, which includes two $1 \times 3 \times 3$ or

$3 \times 3 \times 3$ convolution, instance normalization (Instance-Norm), and LeakyRuLU operations. The use of Instance-Norm considers the limitation of batch size in 3D medical image segmentation. Then, the F_{2D} and F_{3D} are performed the voxel-wise adding and the global average pooling (GAP) to generate channel-wise statistics $F_G \in \mathbb{R}^{c \times 1 \times 1 \times 1}$, which can be expressed as:

$$F_G = GAP(F_{2D} + F_{3D}) = \frac{1}{w \times h \times d} \sum_{i=1}^{w} \sum_{j=1}^{h} \sum_{k=1}^{d} (F_{2D} + F_{3D}) \tag{1}$$

where w, h, and d are the width, height, and depth of the input feature map, respectively.

Further, a local cross-channel information interaction attention mechanism is applied for the fusion of multi-dimensional convolutional features. Specifically, the feature map F_G is squeezed to a one-dimension tensor of length c, which is the number of channels, and then a one-dimensional (1D) convolution with a kernel size of K is applied for information interaction between channels. The obtained feature layer is re-expanded into a 3D feature map F'_G, which can be expressed as:

$$F'_G = G_U(C1D_K(G_S(F_G))) \tag{2}$$

where $C1D_K$ denotes the 1D convolution with the kernel size of K, G_S and G_U denote the operation of squeezing and re-expanding respectively.

To adaptively select the feature information from different convolutions, the softmax operation is performed channel-wise to compute the weight vectors α and β corresponding to F_{2D} and F_{3D} respectively, which can be expressed as:

$$\alpha_i = \frac{e^{A_i F'_G}}{e^{A_i F'_G} + e^{B_i F'_G}}, \beta_i = \frac{e^{B_i F'_G}}{e^{A_i F'_G} + e^{B_i F'_G}} \tag{3}$$

where $A, B \in \mathbb{R}^{c \times c}$ denote the learnable parameters, A_i and B_i denote to the i-th row of A and B respectively, α_i and β_i denote to i-th element of α and β respectively.

$$F_o = \alpha \cdot F_{2D} + \beta \cdot F_{3D} \tag{4}$$

where F_o denotes the output feature map of the ACM.

2.3 Anatomy-Constraint Loss

Fine-grained hand bone segmentation places stringent demands on the anatomical relationships between categories, but the lack of supervision during model training renders it the primary source of segmentation error. For instance, the epiphysis is highly prone to neglect due to its small and imbalanced occupation, while the index and ring phalanx bones can easily be mistaken due to their symmetrical similarities. To address this issue, we propose the anatomy-constraint loss to facilitate the model's learning of anatomical relations.

Anatomical errors occur when pixels significantly deviate from their expected anatomical locations. Thus, we utilize the loss to compute and penalize these

deviating pixels. Assume that the Y and P are the label and segmentation map respectively. First, the map representing anatomical errors is generated, where only pixels in the segmentation map that do not correspond to the anatomical relationship are activated. To mitigate subjective labeling errors caused by the unclear boundaries in US images, we perform morphological dilation on the segmentation map P. This operation expands the map and establishes an error tolerance, allowing us to disregard minor random errors and promote training stability. To make it differentiable, we implement this operation with a kernel=3 and stride=1 max-pooling operation. Subsequently, the anatomical error map F_E is computed by pixel-wise subtracting the Y and the expanded segmentation map P. The resulting difference map is then activated by ReLU, which ensures that only errors within the label region and beyond the anatomically acceptable range are penalized. The process can be expressed as:

$$F_E = \sigma_R(Y_{C_i} - G_{mp}(P_{C_i})) \tag{5}$$

where Y_{C_i} and P_{C_i} denote the i-th category maps of the label Y and segmentation map P respectively, $G_{mp}(\cdot)$ denotes the max-pooling operation, and $\sigma_R(\cdot)$ denotes the ReLU activation operation.

Next, we intersect F_E with the segmentation map P and label Y, respectively, based on which the cross entropy is computed, which is used to constrain the anatomical error:

$$Loss_{AC} = Loss_{CE}(P \odot F_E,\ Y \odot F_E) \tag{6}$$

where $Loss_{AC}(\cdot)$ denotes the proposed anatomy-constraint loss, $Loss_{CE}(\cdot)$ denotes the cross-entropy loss, and $\odot$ denotes the intersection operation.

To reduce the impact of class imbalance on model training and improve the stability of segmentation, we use a combination of Dice loss and anatomy-constraint loss function:

$$L = Loss_{Dice} + \gamma Loss_{AC} \tag{7}$$

where L is the overall loss, $Loss_{Dice}$ denotes the Dice loss, and γ denotes the weight-controlling parameter of the anatomy-constraint loss.

3 Experiments and Results

3.1 Dataset and Implementation

Our method is validated on an in-house dataset, which consists of 103 3D ultrasound volumes collected using device IBUS BE3 with a 12MHz linear transducer from pediatric hand examinations. The mean voxel resolution is $0.088\times0.130\times0.279\,\mathrm{mm}^3$ and the mean image size is $512\times1023\times609$. Two expert ultrasonographers manually annotated the data based on ITK-snap [18]. Each phalanx, metacarpal, and epiphysis were labeled with different categories, and there are a total of 39 categories including the background. The dataset was

randomly spitted into 75% training set and 25% test set. All data were resized to 256×512 in the transverse plane and maintained the axial size. We extracted 256×512×16 voxels training patches from the resized images as the training samples.

The training and test phases of the network were implemented by PyTorch on an NVIDIA GeForce RTX 3090 GPU. The network was trained with Adam optimization with momentum of 0.9. The learning rate was set as 10e-3. For the hyperparameter, the kernel size K of 1D convolution in ACM was set as 3 and the weight-controlling parameter γ was set as 0.5. We used the Dice coefficient (DSC), Jaccard Similarity (Jaccard), Recall, F1-score, and Hausdorff Distance (HD95) as evaluation metrics.

3.2 Performance Comparison

We compared our network with recent and outstanding medical image segmentation methods, which contain UNet [15], UNet++ [19], 3D UNet [5], VNet [13], MNet [6], and the transformer baseline SwinUNet [3]. For a fair comparison, we used publicly available hyperparameters for each model. For the 3D network, the data processing method is consistent with ours, while for the 2D network, we slice the images along the axis to convert the 3D data into 2D.

Table 1 lists the results. Note that our method achieved the highest quantitative performance of DSC, Jaccard, Recall and F1-score, with values of 0.900, 0.819, 0.871, and 0.803, respectively. These results improved by 1.3%, 2.1%, 0.8%, and 1.3% compared to the best values of other methods. Note that our method outperformed the MNet that is a state-of-the-art (SOTA) method, which demonstrated the effectiveness of adaptive multi-dimensional feature fusion and anatomy-constraint for enhancing model performance.

Table 1. Quantitative comparison experiments between the proposed method and the outstanding segmentation methods. Dim denotes to dimension.

Dim	Methods	DSC↑	Jaccard↑	Recall↑	F1-score↑	HD95 (mm) ↓
2D	UNet [15]	0.855	0.747	0.837	0.736	1.215
	UNet++ [19]	0.875	0.779	0.763	0.697	3.328
	SwinUNet [3]	0.829	0.709	0.796	0.657	1.278
3D	3D UNet [5]	0.829	0.709	0.584	0.632	0.960
	VNet [13]	0.864	0.761	0.863	0.730	3.380
2D⊕3D	MNet [6]	0.887	0.798	0.830	0.790	**0.695**
	Ours	**0.900**	**0.819**	**0.871**	**0.803**	1.184

Figure 2 shows the visualization results of our method and comparative methods (Due to page limitations, we only presented the comparison results

of our method with baseline and SOTA methods). Compared to other methods, our method has advantages in effectively mining difficult samples, particularly in accurately identifying and classifying clinically important but difficult-to-distinguish epiphysis. Additionally, it can effectively learn the anatomical relationships of different categories, reducing the occurrence of category confusion, particularly in the phalanges of the index finger, middle finger, and ring finger.

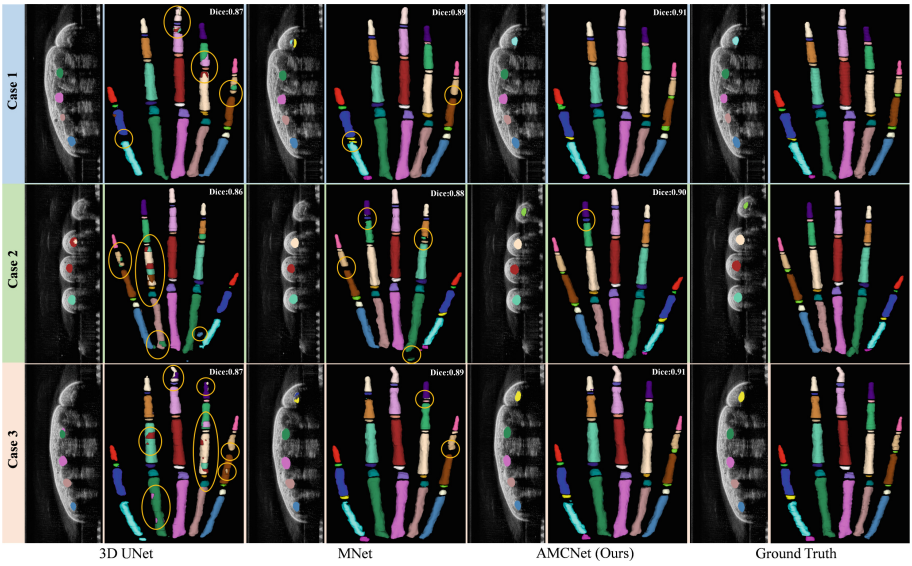

Fig. 2. The visualization results. Orange circles indicate obvious segmentation errors. (Color figure online)

3.3 Ablation Study

To validate the effect of anatomy-constraint loss, we compare the results of training with $Loss_{Dice}$ and the combination of $Loss_{Dice}$ and $Loss_{AC}$ on both 3D UNet and the proposed AMCNet. Table 2 lists the results. Note that compared with only $Loss_{Dice}$, UNet and our method have improved in various metrics after adding $Loss_{AC}$, boosting 0.6% in DSC and 1.1% in Jaccard. The results indicate that enforcing anatomical constraints to encourage the model to learn anatomical relationships improves the model's feature representation, resulting in better performance. Additionally, to verify the effect of the adaptive multi-dimensional feature fusion mechanism, we modified the ACM module to only 2D and 3D convolutional blocks, respectively. The results are shown in Table 2. Note that the 2D and 3D feature adaptive fusion mechanism improves model performance. Specifically, under $Loss_{Dice}$ and $Loss_{AC}$, it has resulted in an increase of 1.0% and 0.9% in DSC, 1.1% and 1.5% in Jaccard, respectively, compared to using only 2D or 3D convolution.

Table 2. Ablation study on the effect of $Loss_{AC}$ and ACM

Methods	DSC↑	Jaccard↑	Recall↑	F1-score↑	HD95 (mm) ↓
3D UNet [15]	0.829	0.709	0.584	0.632	0.960
3D UNet+$Loss_{AC}$	0.835	0.718	0.692	0.686	0.874
AMCNet 3D	0.887	0.797	0.787	0.648	2.705
AMCNet 2D	0.887	0.798	0.842	0.759	4.979
AMCNet 2D⊕3D	0.894	0.808	0.875	0.800	1.903
AMCNet 3D+$Loss_{AC}$	0.891	0.804	**0.877**	0.787	2.728
AMCNet 2D+$Loss_{AC}$	0.890	0.808	0.719	0.750	**0.748**
AMCNet 2D⊕3D +$Loss_{AC}$ (Ours)	**0.900**	**0.819**	0.871	**0.803**	1.184

3.4 Software Development and Application

Based on the method described above, a user-friendly and extensible module was developed on the 3D Slicer platform [7] to facilitate user access, as shown in Fig. 3. To use the module, users simply select the input and output file formats on the module interface and click the "apply" button. The software will then automatically perform image preprocessing, call the model for inference, and deliver the segmentation result within twenty seconds (see supplementary material 1). This plugin will be released globally to promote the greater value of the proposed method in clinical applications.

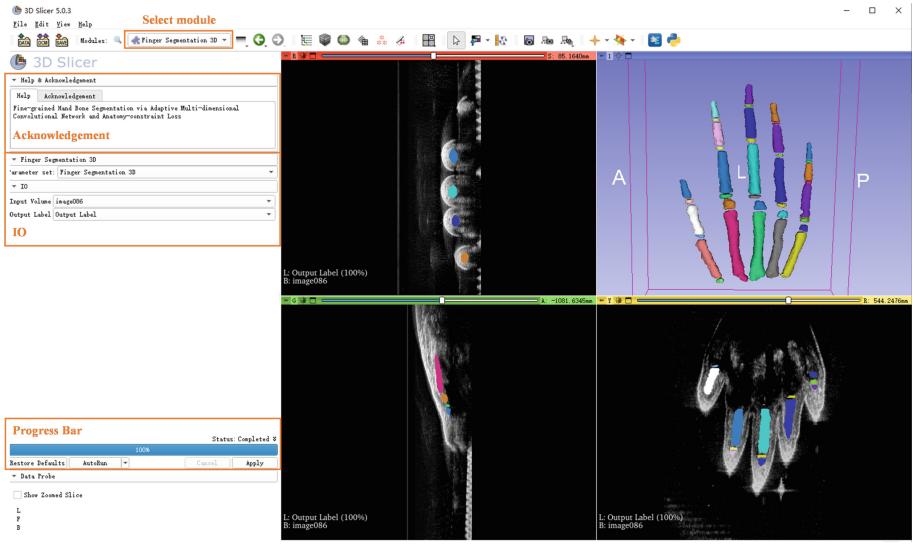

Fig. 3. The extensible module for our method based on the 3D Slicer platform

4 Conclusion

In this work, we have presented an adaptive multi-dimensional convolutional network, called AMCNet, to address the challenge of automatic fine-grained hand bone segmentation in 3D US volume. It adopts an adaptive multi-dimensional feature fusion mechanism to dynamically adjust the weights of 2D and 3D convolutional feature layers according to different objectives. Furthermore, an anatomy-constraint loss is designed to encourage the model to learn anatomical relationships and effectively mine hard samples. Experiments show that our proposed method outperforms other comparison methods and effectively addresses the task of fine-grained hand bone segmentation in ultrasound volume. The proposed method is general and could be applied to more medical segmentation scenarios in the future.

Acknowledgements. This work was supported by grants from the National Natural Science Foundation of China (81971709; M-0019; 82011530141), the Foundation of Science and Technology Commission of Shanghai Municipality (20490740700; 22Y11911700), Shanghai Jiao Tong University Foundation on Medical and Technological Joint Science Research (YG2021ZD21; YG2021QN72; YG2022QN056; YG2023ZD19; YG2023ZD15), the Funding of Xiamen Science and Technology Bureau (3502Z20221012).

References

1. Ahmed, O., Moore, D.D., Stacy, G.S.: Imaging diagnosis of solitary tumors of the phalanges and metacarpals of the hand. Am. J. Roentgenol. **205**(1), 106–115 (2015)
2. Alsinan, A.Z., Patel, V.M., Hacihaliloglu, I.: Automatic segmentation of bone surfaces from ultrasound using a filter-layer-guided CNN. Int. J. Comput. Assist. Radiol. Surg. **14**, 775–783 (2019)
3. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: ECCV Workshops (2021)
4. Cecava, N.D., Kephart, D.A., Bui-Mansfield, L.T.: Bone lesions of the hand and wrist: systematic approach to imaging evaluation. Contemp. Diagn. Radiol. **43**(5), 1–7 (2020)
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
6. Dong, Z., et al.: MNet: rethinking 2D/3D networks for anisotropic medical image segmentation. arXiv preprint arXiv:2205.04846 (2022)
7. Fedorov, A., et al.: 3D slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012)
8. Iagnocco, A., et al.: The reliability of musculoskeletal ultrasound in the detection of cartilage abnormalities at the metacarpo-phalangeal joints. Osteoarthritis Cartilage **20**(10), 1142–1146 (2012)
9. Kuok, C.P., et al.: Segmentation of finger tendon and synovial sheath in ultrasound image using deep convolutional neural network. Biomed. Eng. Online **19**(1), 24 (2020)

10. Liu, R., et al.: NHBS-Net: a feature fusion attention network for ultrasound neonatal hip bone segmentation. IEEE Trans. Med. Imaging **40**(12), 3446–3458 (2021)
11. Luan, K., Li, Z., Li, J.: An efficient end-to-end CNN for segmentation of bone surfaces from ultrasound. Comput. Med. Imaging Graph. **84**, 101766 (2020)
12. Martins, N., Sultan, S., Veiga, D., Ferreira, M., Teixeira, F., Coimbra, M.: A New active contours approach for finger extensor tendon segmentation in ultrasound images using prior knowledge and phase symmetry. IEEE J. Biomed. Health Inf. **22**(4), 1261–1268 (2018)
13. Milletari, F., Navab, N., Ahmadi, S.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision, pp. 565–571 (2016)
14. Rahman, A., Bandara, W.G.C., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Orientation-guided graph convolutional network for bone surface segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science. vol. 13435. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_40
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., Leonardi, R.: Deep learning for automated skeletal bone age assessment in X-ray images. Med. Image Anal. **36**, 41–51 (2017)
17. Wang, G., et al.: Automatic segmentation of vestibular schwannoma from T2-weighted MRI by deep spatial attention with hardness-weighted loss. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 264–272. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_30
18. Yushkevich, P.A., et al.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. NeuroImage **31**(3), 1116–1128 (2006))
19. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N.; Liang, J.:UNet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging **39**(6), 1856–1867 (2020)