2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822692

BAformer: Boundary Adaptive Transformer for Medical Image Segmentation

Junlong Cheng
Sichuan University & Shanghai AI Lab
School of Computer Science
Chengdu, China
cjl951015@163.com

Jilong Chen
Sichuan University
School of Computer Science
Chengdu, China
chenjilong617@gmail.com

Lei Jiang
Sichuan University
School of Computer Science
Chengdu, China
2022223045177@stu.scu.edu.cn

Chenrui Gao
Sichuan University
School of Computer Science
Chengdu, China
cr@stu.scu.edu.cn

Junren Chen
Sichuan University
School of Computer Science
Chengdu, China
ichenjunren@outlook.com

Ruoyu Li
Sichuan University
School of Computer Science
Chengdu, China
ruoyuli@ieee.org

Min Zhu*
Sichuan University
School of Computer Science
Chengdu, China
zhumin@scu.edu.cn

***Abstract*—Accurate automatic segmentation of medical images has long faced challenges such as significant lesion scale variations and blurred boundaries. This study proposes a Boundary-Adaptive Transformer (BAformer) specifically designed for 2D medical image segmentation. BAformer employs a hierarchical architecture that focuses on boundary-related features across varying resolutions. Additionally, we introduce a Dense Feed-forward Network (DenseFFN) to reuse features, enabling each layer to accumulate information from previous layers. Finally, we extend the benefits of the dense network to the decoder, balancing parameter efficiency and performance. Extensive experiments demonstrate that BAformer achieves state-of-the-art performance on several challenging medical segmentation tasks. Furthermore, BAformer can seamlessly integrate with existing segmentation networks, demonstrating its versatility and effectiveness.**

***Index Terms*—Curvilinear structure segmentation, Skeleton prompt, Large-scale CSS dataset, Foundational model**

I. Introduction

Accurately extracting ”regions of interest” from medical images is crucial [1], [2]. However, pathological regions often show significant scale and color changes. Some pathological regions have low contrast with normal tissues, leading to blurred boundaries. So it is important to develop a segmentation network that can effectively handle blurred boundaries.

End-to-end automatic segmentation methods have shown great potential in medical image analysis [1]–[5]. U-Net [1] captures high-level features through the encoder, the decoder recovers spatial details, and uses jump connections to fuse information of different scales to improve segmentation accuracy. Its variants mainly extract multi-scale features through complex architectures [3], [4] or introduce attention modules to enhance feature representation [2], [5], but are limited by the local receptive field and are difficult to capture long-distance dependencies and global information. In addition, stride and pooling operations reduce computational complexity, but also lead to image resolution loss, increasing the difficulty of fine segmentation. Transformers [6] effectively address this limitation, and the self-attention mechanism can capture long-distance dependencies. ViTs [7] demonstrates the potential of Transformers in image processing, and TransU-Net [8] enhances CNN features on this basis. SwinU-Net [9] introduces a layered Swin Transformer to restore visual resolution through decoder upsampling. Chen et al. [10] combined CNN with probabilistic graphical models; Lee et al. [11] proposed a structure-preserving segmentation framework; BAT [12] utilized boundary prior knowledge; and Wang et al. [13] introduced an edge attention preservation module.

We identify two key aspects: (i) local and global information extraction; (ii) boundary perception. Therefore, we propose BAformer, whose advantages include: (1) using Transformer architecture to capture global features and focus on boundary features at different resolutions; (2) introducing dense connections, balancing parameter efficiency and performance through DenseFFN and decoder networks; (3) verifying the effectiveness on four challenging tasks. Importantly, it can be combined with other segmentation methods to continuously improve segmentation performance.

II. Boundary Adaptive Transformer

A. Overall Architechture

As shown in Fig. 1(a) , BAformer is a layered Transformer model inspired by convolutional neural networks [14]. The input data is downsampled by the Conv3x3 layer in the stem stage (S0). The main body of the network consists of three key stages (S1-S3). Each stage first reduces the resolution of the input features by half through the patch embedding layer, and then the BAformer module captures global dependencies and enhances boundary feature representation. We combine BAformer with a concise dense decoder architecture, which includes convolutional layers that reduce the dimension of feature channels by 4 times, upsampling layers, and a series of dense convolution operations. These dense convolutions expand the feature channels by 2 times and fuse them with the

979-8-3503-8622-6/24/$31.00 ©2024 IEEE

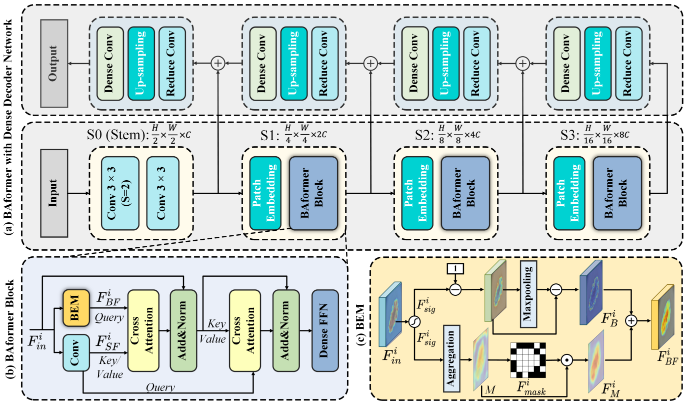

Fig. 1. Overview of BAformer. (a) The BAformer combined with a dense decoder forms a U-shaped segmentation network; (b) Details of the BAformer block; (c) The boundary extraction module (parameter-free).

features of the corresponding stages of the encoder to achieve effective integration and transmission of feature information.

B. BAformer Block

As shown in Fig. 1(b), BAformer consists of three components: boundary extraction module (BEM), cross-attention and dense feedforward network (DenseFFN). First, the input feature F_{in}^i is adaptively extracted from the boundary features by BEM, and the foreground features are enhanced to help the model better understand and locate the boundary area. Then, the cross-attention mechanism establishes the connection between boundary features and global features. Finally, DenseFFN fuses multi-scale feature information, performs nonlinear transformation and fusion on the features, further enhances the feature expression capability.

1) Boundary extraction module: BEM is a parameter-free module designed to capture task-specific boundary feature representations and effectively suppress background noise. As illustrated in Fig. 2(c), for the input features F_{in}^i of the i-th stage, we first process them through a sigmoid function to generate $F_{sig}^i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times 2^{i+1}C}$. Subsequently, we employ a max pooling operator to obtain boundary features, which can be expressed using the formula:

$$F_B^i = Maxpooling(1 - F_{sig}^i, K) - (1 - F_{sig}^i) \qquad (1)$$

In the above formula, K represents the size of the sliding window, which we set to 3, with a stride and padding of 1.

Inspired by [15], we eliminate background noise and fuse it with the extracted boundary features through the following steps: (1) First, we calculate the aggregated map of F_{sig}^i by summing the feature weights across each channel (from 0 to C-1) of F_{sig}^i, resulting in regions with higher target weights. This is expressed as $M = \sum_{c=0}^{C-1} F_{sig}^i$; (2) Then, We take the average value $\bar{M}$ of all positions (i,j) in M as the threshold, and based on this threshold $\bar{M}$, we determine the foreground and background positions in M to generate $F_{mask}^i \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times 1}$:

$$F_{mask}^i = \begin{cases} 1 & \text{if} \quad M^{(i,j)} > \bar{M} \\ 0 & \quad \text{other} \end{cases} \qquad (2)$$

where 1 in F_{mask}^i represents the foreground and 0 represents the background; (3) Finally, we perform an element-wise Hadamard product operation between F_{mask}^i and M to obtain F_M^i, after eliminating background noise. Through the above process, we obtain the boundary feature map F_B^i and the noise-removed feature map F_M^i. We then obtain the final boundary feature map $F_{BF}^i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times 2^{i+1}C}$ by element-wise adding these two feature maps.

2) Cross-Attention: Unlike the traditional Transformer architecture [6], [7], the BAformer module only performs two attention calculations, thereby achieving efficient feature fusion. At a specific stage i, given the input feature F_{in}^i, we first generate a feature map F_{SF}^i through a 3×3 convolutional layer to capture local details and textures. In the first cross-attention calculation, F_{BF}^i (boundary feature) is used as the query, and F_{SF}^i is used as the key and value to effectively fuse the boundary and image space features. Subsequently, in the second calculation, the output of the first time is used as the new key and value, and F_{SF}^i is used as the query to model the global contextual dependency of the image space features. After each calculation, residual connections and layer normalization operations are introduced. The above process can be expressed as:

$$F_{Att1}^i = CorssAtt(F_{BF}^i, F_{SF}^i) + F_{in}^i \qquad (3)$$

$$F_{Att2}^i = CorssAtt(F_{SF}^i, Norm(F_{Att1}^i)) + Norm(F_{Att1}^i) \qquad (4)$$

3) Vanilla FFN vs. DenseFFN: Each layer of the BAformer module consists of a densely connected feedforward network. Different from the traditional Transformer architecture [7], [14], [16], we apply densely connected linear layers in the channel direction to optimize the information flow between channels. Specifically, assuming that the feature map of stage i is $F_0^i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C}$, we first reshape it into $F_0^i \in \mathbb{R}^{N\times C}$, where $N = HW$. In DenseFFN, each layer is connected to all previous layers, that is, $F_0^i, F_1^i, ..., F_{k-1}^i$ as the input of the next layer. This design makes the information transmission and fusion between channels smoother, thereby improving the model's feature extraction and representation capabilities:

$$\begin{aligned} &F_1^i = FFN(FFN(F_0^i, 4*G), G) \copyright F_0^i, \\ &F_2^i = FFN(FFN(F_1^i, 4*G), G) \copyright F_1^i, \\ &......, \\ &F_k^i = FFN(FFN(F_{k-1}^i, 4*G), G) \copyright F_{k-1}^i \end{aligned} \qquad (5)$$

Herein, G signifies the growth rate of the channel, which corresponds to the hidden layer of the FFN layer (we set $G = 16$). "©" symbolizes concatenation. Please note that the Normalization layer and activation layer succeeding the MLP layer have been omitted in the aforementioned equation for simplicity.

DenseFFN is different from vanilla FFN. It adds G channels to the output features of the previous layer in sequence, and generates an output feature map $F_{k-1} \in \mathbb{R}^{N\times 2C}$ after k-1 dense connections (the number of output channels is increased to twice the input to prevent feature information loss). Finally, we reshape F_{k-1} into output $F_{k-1} \in \mathbb{R}^{H\times W\times 2C}$ and reuse features through channel-level connections. The advantages of this method include: (1) easier training; (2) alleviating the gradient vanishing problem; (3) giving the network multi-scale capabilities; (4) achieving a better balance between efficiency and accuracy.

C. Dense Decoder Network

As shown in Fig. 1, we combine BAformer with a dense decoder network. Each decoder block consists of a 1x1 convolution, an upsampling layer, and a series of dense convolutions. To balance efficiency and reduce redundant features, we use Reduce convolution to reduce the number of input feature channels by a quarter, and then double the resolution through bilinear interpolation to align the spatial resolution of the corresponding encoder stage. Finally, the number of channels is increased by 2 times through dense convolution to match the number of channels in the encoder stage, with a growth rate of G 16. Our method increases the network depth without increasing the computational burden and promotes the fusion of features of different scales. In addition, we implement skip connections to recover details from the encoder and decoder to compensate for information loss.

III. Experiments and Discussion

A. Datasets

The **ISIC2017** dataset comprises 2600 dermoscopic images, of which 2000 images are allocated for training and 600 for testing. The **TN-SCUI** dataset encompasses 3644 thyroid nodule ultrasound images, which we randomly partitioned in a 6:2:2 ratio for training, validation, and testing, respectively. The **GLaS** dataset consists of 165 microscopic images of H&E stained slides. We utilized 85 images for training and the remaining 80 for testing. The **COVID-19** dataset (https://medicalsegmentation.com/) includes 100 axial CT images from over 40 COVID-19 patients, we conducted experiments using five-fold cross-validation (i.e., training on 80 images and validating on 20 images each time).

B. Implementation Details

We use a Keras-based approach to train on NVIDIA RTX3090 GPU (24G). The Adam optimizer is adopted, the learning rate is fixed to 1e-4, and the mini-batch size is 16. Training is terminated via an early stopping mechanism when the validation loss stabilizes without significant fluctuations over 30 epochs. All comparative experiments used the same training and test sets.

C. Comparison with State-of-the-arts

We compare the performance of BAformer, the baseline model U-Net and eight other SOTA methods in Table I. The results show that Transformer-based methods and hybrid architectures generally outperform pure CNN methods. In particular, BAformer achieves leading IoU and Dice due to its focus on task-related boundary information, while achieving a significant balance of performance and parameters with only 16M parameters and 30G FLOPs. Figure 2 (a) and (b) show the qualitative comparison of different models on each data set. Each dataset has unique characteristics and varies in lesion size. Experimental results show that BAformer performs well in qualitative analysis and is closer to ground truth.

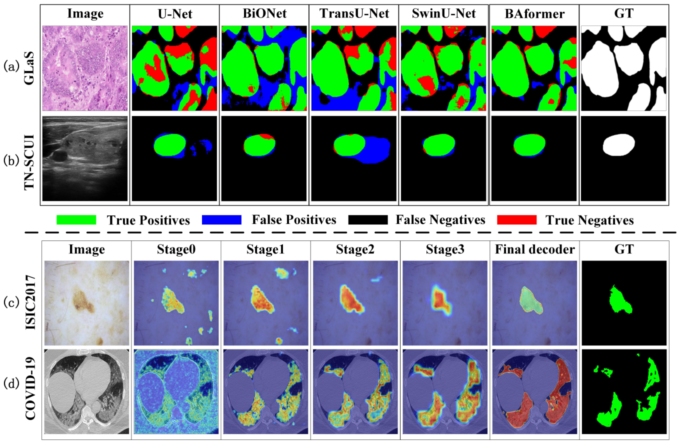

Fig. 2. (a) and (b) Qualitative experimental results of BAformer and other state-of-the-art. (c) and (d) Visualization of the output of BAformer at different stages (S0-S3) and the final decoder layer.

TABLE I
Quantitative results on four datasets. Models with the first, second, and third-best performances are highlighted in red, green, and blue respectively. "#Params" denotes number of model parameters.

Network	ISIC2017				TN-SCUI				#Params	FLOPs
	IoU	Dice	Sens	Spec	IoU	Dice	Sens	Spec		
U-Net [1]	0.746	0.834	0.823	0.974	0.711	0.799	0.828	0.989	30 M	42 G
BiONet [3]	0.774	0.854	0.854	0.968	0.713	0.803	0.839	0.987	57 M	71 G
ResGANet [5]	0.764	0.861	0.842	0.950	0.726	0.818	0.830	0.990	39 M	65 G
TransU-Net [8]	0.785	0.865	0.865	0.965	0.737	0.826	0.844	0.990	100 M	25 G
SwinU-Net [9]	0.775	0.856	0.861	0.964	0.736	0.825	0.841	0.991	26 M	6 G
FATNet [17]	0.772	0.854	0.873	0.949	0.742	0.830	0.856	0.989	27 M	43 G
U-Netv2 [18]	0.751	0.843	0.847	0.970	0.731	0.824	0.855	0.989	25 M	5 G
Convformer [19]	0.773	0.854	0.847	0.967	0.739	0.823	0.851	0.989	30 M	82 G
SegNetr [20]	0.775	0.856	0.863	0.967	0.767	0.850	0.847	0.988	12 M	10 G
BAformer	0.790	0.869	0.880	0.966	0.771	0.855	0.852	0.991	16 M	30 G
Network	**GLaS**				**COVID-19**				**#Params**	**FLOPs**
	IoU	**Dice**	**Sens**	**Spec**	**IoU**	**Dice**	**Sens**	**Spec**		
U-Net [1]	0.773	0.864	0.863	0.873	0.620	0.749	0.775	0.983	30 M	42 G
BiONet [3]	0.758	0.852	0.827	0.903	0.592	0.731	0.742	0.981	57 M	71 G
ResGANet [5]	0.792	0.878	0.878	0.883	0.638	0.768	0.770	0.985	39 M	65 G
TransU-Net [8]	0.801	0.884	0.882	0.898	0.655	0.783	0.795	0.986	100 M	25 G
SwinU-Net [9]	0.797	0.880	0.871	0.895	0.611	0.744	0.764	0.983	26 M	6 G
FATNet [17]	0.800	0.884	0.893	0.871	0.648	0.778	0.773	0.986	27 M	43 G
U-Netv2 [18]	0.785	0.873	0.873	0.879	0.590	0.728	0.775	0.979	25 M	5 G
Convformer [19]	0.807	0.886	0.913	0.850	0.643	0.771	0.790	0.983	30 M	82 G
SegNetr [20]	0.813	0.893	0.916	0.853	0.652	0.780	0.791	0.986	12 M	10 G
BAformer	0.826	0.901	0.897	0.900	0.661	0.787	0.8031	0.986	16 M	30 G

TABLE II
Ablation study of BAformer, where we replace the original backbones of U-Net, FATNet, and TransU-Net with BAformer.

Model	Backbone	Datasets (Dice)	
		ISIC2017	GLaS
U-Net [1]	Original	0.834	0.864
	BAformer	0.855	0.875
FAT-Net [12]	Original	0.854	0.884
	BAformer	0.867	0.890
TransU-Net [8]	Original	0.865	0.884
	BAformer	0.870	0.898

D. BAformer intermediate layer analysis

To qualitatively assess the efficacy of BAformer in focusing on boundary information and establishing long-range dependencies, we visualize the output feature maps of the four encoder blocks and the final decoder block of BAformer. As depicted in Fig. 2 (c) and (d), the shallow S0 stage primarily focuses on texture and grayscale information. As the network depth increases, the network gradually locates the lesion area, and the background noise demonstrates a stage-by-stage reduction trend. By observing the visualization results of the decoder, we find that the weight of the boundaries is greater than the surrounding pixels. This indicates that BAformer facilitates the enhancement of boundary representation.

E. Ablation Study

The ablation study results of BAformer in Table II show that as a backbone network, BAformer outperforms the three mainstream image segmentation models in both the CNN-based architecture and the hybrid architecture combined with Transformer, verifying its effectiveness and wide applicability. This section conducts a detailed ablation study on the boundary extraction module (BEM). Table III shows that removing the BEM module will lead to a decrease in IoU and Dice coefficients, indicating that it is crucial to improving model performance; replacing it with a simple volume product operation can improve performance, but it increases parameter overhead, further confirming the effectiveness and unique advantages of BEM. Finally, Figure 3 shows an ablation study of the growth rate of DenseFFN, and it is found that when the growth rate G is 16, BAformer performs best.

TABLE III
ABLATION STUDIES OF THE BEM ON THE ISIC2017 DATASET.

Model	IoU	Dice	#Params	FLOPs
Remove BEM	0.777	0.858	16 M	30
Conv replaces BEM	0.781	0.861	17 M	30
With BEM	0.790	0.869	16 M	30

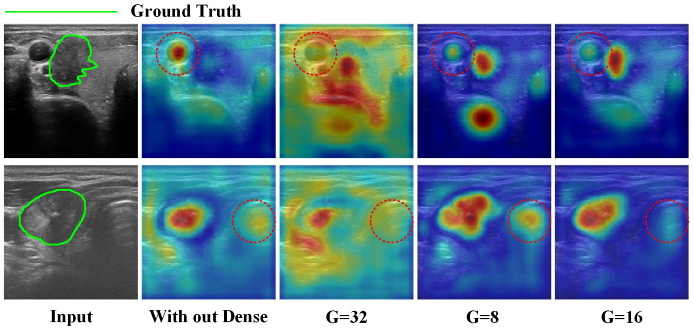

Fig. 3. Example of ablation results for different growth rates. We use heatmaps to visualize the output feature maps of the final stage of BAformer.

CONCLUSION

This study proposes a boundary adaptive Transformer (BAformer) for medical image segmentation, which can adaptively focus on boundary features relevant to the current task and preserve the long-range dependencies of the Transformer. Comprehensive evaluation shows that BAformer can achieve state-of-the-art segmentation performance while maintaining low parameters and GFLOPs when combined with a dense decoder network. Notably, BAformer shows excellent versatility and flexibility, and can be easily extended to other segmentation methods to enhance segmentation results.

REFERENCES

[1] 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)

[2] Cheng, J., Tian, S., Yu, L., Lu, H., Lv, X.: Fully convolutional attention network for biomedical image segmentation. Artif Intell Med. 107, 101899. (2020)

[3] Xiang, T., Zhang, C., Liu, D., Song, Y., Huang, H., Cai, W.: BiO-Net: learning recurrent bi-directional connections for encoder-decoder architecture. In: A. L. Martel et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 74–84. Springer, Cham (2020)

[4] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: D. Stoyanov et al. (Eds.) DLMIA 2018. LNCS, vol. 11045, pp. 3–11. Springer International Publishing. (2018)

[5] Cheng, J., et al.: ResGANet: Residual group attention network for medical image classification and segmentation. Med. Image Anal. 76, 102313. (2022)

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I.: In: Attention is all you need. Advances in neural information processing systems, 30. (2017)

[7] Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: ICLR (2021)

[8] Chen, J., et al.: TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)

[9] Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) ECCV 2022. Lecture Notes in Computer Science, vol. 13803, pp. 205–218. Springer, Cham (2022)

[10] Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPMI. 40(4), 834-848. (2017)

[11] Lee, H. J., Kim, J. U., Lee, S., Kim, H. G., Ro, Y. M.: Structure boundary preserving segmentation for medical image with ambiguous boundary. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4817-4826. (2020)

[12] Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J.: Boundary-aware transformers for skin lesion segmentation. In: M. de Bruijne et al. (Eds.): MICCAI 2021. LNCS, vol. 12901, pp. 206–216. Springer, Cham (2021)

[13] Wang, K., Zhang, X., Zhang, X., Lu, Y., Huang, S., Yang, D.: EANet: Iterative edge attention network for medical image segmentation. Pattern Recognition. 127, 108636. (2022)

[14] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012-10022. (2021)

[15] Wei, X. S., Luo, J. H., Wu, J., Zhou, Z. H.: Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing. 26(6), 2868-2881. (2017)

[16] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10819-10829. (2022)

[17] Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z.: FAT-Net: feature adaptive transformers for automated skin lesion segmentation. Med. Image Anal. 76, 102327 (2022)

[18] Peng Y, Sonka M, Chen D Z. U-Net v2: Rethinking the Skip Connections of U-Net for Medical Image Segmentation[J]. arXiv preprint arXiv:2311.17791, 2023.

[19] Lin, X., Yan, Z., Deng, X., Zheng, C., Yu, L.: ConvFormer: Plug-and-Play CNN-Style Transformers for Improving Medical Image Segmentation. In: H. Greenspan et al. (Eds.): MICCAI 2023. LNCS, vol. 14223, pp. 642–651. Springer, Cham (2023).

[20] Cheng, J., Gao, C., Wang, F., Zhu, M.: Segnetr: Rethinking the local-global interactions and skip connections in u-shaped networks. In: H. Greenspan et al. (Eds.): MICCAI 2023. LNCS, vol. 14225, pp. 64–74. Springer, Cham (2023).