DC-Net: Dual Context Network for 2D Medical Image Segmentation

Rongtao Xu[1,3], Changwei Wang[1,3], Shibiao Xu[2(✉)], Weiliang Meng[1,3], and Xiaopeng Zhang[1,3]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
shibiaoxu@bupt.edu.cn
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

Abstract. Medical image segmentation is essential for disease diagnosis analysis. There are many variants of U-Net that are based on attention mechanism and dense connections have made progress. However, CNN-based U-Net lacks the ability to capture the global context, and the context information of different scales is not effectively integrated. These limitations lead to the loss of potential context information. In this work, we propose a Dual Context Network (DC-Net) to aggregate global context and fuse multi-scale context for 2D medical image segmentation. In order to aggregate the global context, we present the Global Context Transformer Encoder (GCTE), which reshapes the original image and the multi-scale feature maps into a sequence of image patches, and combines the advantages of Transformer Encoder on global context aggregation to improve the performance of encoder. For the fusion of multi-scale context, we propose the Adaptive Context Fusion Module (ACFM) to adaptively fuse context information by learning Adaptive Spatial Weights and Adaptive Channel Weights to improve the performance of decoder. We apply our DC-Net with GCTE and ACFM to skin lesion segmentation and cell contour segmentation tasks, experimental results show that our method can outperform other advanced methods and get state-of-the-art performance.

Keywords: Medical image segmentation · Visual transformer · Multi-scale context fusion.

1 Introduction

In recent years, deep learning has been successfully applied to medical image segmentation tasks such as skin lesion segmentation, vascular wind, lung segmentation, and cell segmentation. In particular, U-Net [16] and its variants [5,8,

R. Xu and C. Wang—Contributed equally

© Springer Nature Switzerland AG 2021
M. de Bruijne et al. (Eds.): MICCAI 2021, LNCS 12901, pp. 503–513, 2021.
https://doi.org/10.1007/978-3-030-87193-2_48

14,15,23] that are based on fully convolutional neural network (FCN) [13] have achieved great success in many medical image segmentation tasks. Basically, U-Net consists of an encoder, a decoder and a skip connection between them. The skip connection can add details to the decoder that the encoder loses. However, the convolutional layer of the above methods is only a local operation and lacks the integration of context information, while medical image segmentation usually benefits from a wide range of context information. Additionally, U-Net can obtain limited context aggregation ability by continuously stacking downsampling encoders, which is far from the optimal solution.

To improve the CNN's awareness of context, non-local Neural Networks [19] uses self-attention mechanisms to model long-distance dependencies between pixels to obtain global context information. On the other hand, DeeplabV3 [3] and CE-Net [8] integrate multi-scale context information, these context information from different receptive fields can further enhance the expressive ability of features. Inspired by the above works, we try to combine these two ways of context augmented to make U-Net gain stronger context awareness.

In this work, we present DC-Net, powered by Global Context Transformer Encoder (GCTE) and Adaptive Context Fusion Module (ACFM) to mitigate the above issues, the contributions are given as below: First, we designed the GCTE based on the VIT [21] inspiration, and aggregated global context information by modeling the long-range dependence between pixels. Second, we designed the ACFM to adaptively fuse context information of different scales to further take advantage of U-Net's inherent feature hierarchy. Our proposed method has enhanced U-Net's ability to perceive context information from the two aspects of aggregating global context and fusing multi-scale context, thereby greatly improving the accuracy of segmentation. Finally, we verified the effectiveness of our method on the ISIC 2018 and ISBI 2012 datasets. Experiments show that our method surpasses the existing state-of-the-art methods on both datasets.

2 Related Works

Medical Image Segmentation Based on Deep Learning. Fully Convolutional Network (FCN) [13] frameworks based U-Net [16] and DeepLab [3,4] are widely used for image segmentation. There are also some variants of U-Net with good performance for medical image segmentation, such as Attention U-Net [15], U-Net++ [23], Inf-Net [6]. Although these methods usually use the attention mechanism and dense connections to improve the feature representation ability of U-Net, they do not integrate the global context or multi-scale context, which are crucial for medical image segmentation.

Global Context Augmented. The self-attention mechanism [19] is proposed based on capturing the dependencies between long-range features, which has been used for image classification [19], image generation [22] and image segmentation [20], etc. No-local U-Net proposes a global aggregation block based on the self-attention mechanism to gather global information to obtain more

accurate segmentation results. Transformer [17], which has been successfully applied in NLP tasks recently, was introduced into vision tasks by Visual Transformer [21]. Visual Transformer provides a purely self-attention-based pipeline by converting images into patch sequences to further enhance the network's ability to model the global context. In this work, we propose a Global Context Transformer Encoder based on Visual Transformer to augment U-Net's perception of the global context.

Cross-Scale Context Aggregation. Integrating multi-scale context information is a common method to improve the expressive ability of features in many visual tasks [3,8,12,18]. DeepLabV3 [3] designed an Atrous Spatial Pyramid Pooling module to integrate the context information of multiple receptive fields. CE-Net [8] proposed dense atrous convolution block and residual multi-kernel pooling to capture multi-scale context information. In this work, we try to fuse the multi-scale context information output by the U-Net decoding layer through an adaptive weighting method to obtain a more powerful feature representation.

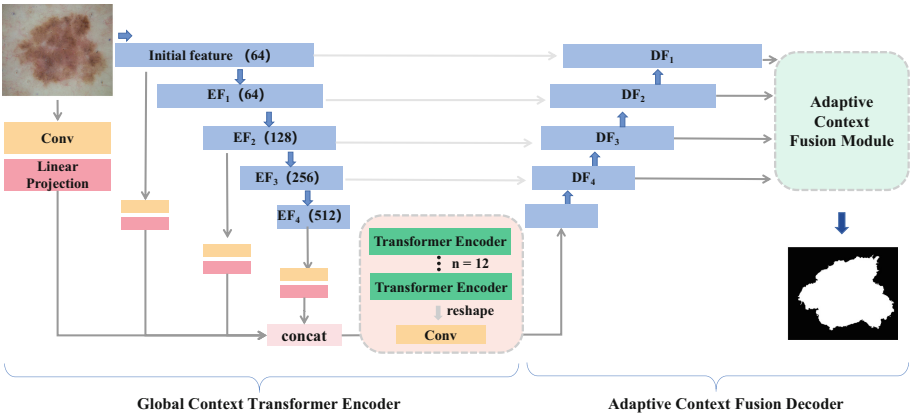

Fig. 1. The structure of the proposed DC-Net with Global Context Transformer Encoder (GCTE) and Adaptive Context Fusion Module (ACFM).

3 Our Proposed Method

We propose a Dual-context network (DC-Net) for medical image segmentation tasks. As shown in Fig. 1, the proposed DC-Net consists of two main parts: Global Context Transformer Encoder and Adaptive Context Fusion Module. In Sect. 3.1, our transformer is combined to encode the feature representation of the image patches decomposed by the multi-scale feature maps. In Sect. 3.2, different proportions of context information are adaptively fused for decoding.

3.1 Global Context Transformer Encoder

Multi-scale Feature Serialization. Different from Vision Transformer [21], our Global Context Transformer Encoder first performs multi-scale feature Serialization. As shown in Fig. 1, the initial feature maps is obtained once the original image undergoes the first pooling. Then the Initial feature maps are processed by four feature extraction blocks of ResNet-34 [9] in turn to obtain the first Encode Feature maps (EF_1), the second Encode Feature maps (EF_2), the third Encode Feature maps (EF_3), and the fourth Encode Feature maps (EF_4). To aggregate global context information, we perform linear projection and patch embeddings to the original image, initial feature maps, EF_2 and EF_4, respectively. Specifically, in order to process 2D images with a resolution of (H, W) and feature maps of various scales, we reshape the input $x \in \mathbb{R}^{H\times W\times C}$ into a series of flattened 2D patches $x_p^i \in \mathbb{R}^{P^2\times C}$ where P is the size of each patch, the value of i is an integer ranging from 1 to N. The length of the input sequence is obtained by $N = \frac{H\times W}{P^2}$.

Patch Embedding. We map the flattened image and multi-scale feature maps to the D dimension using a trainable linear projection. And we learn specific position embeddings and add them to the patch embedding to encode the patch context information:

$$Z_0^i = \left[x_p^1 E^i; x_p^2 E^i; \cdots x_p^N E^i\right] + E_{pos}^i \qquad i = 1, 2, 3, 4 \tag{1}$$

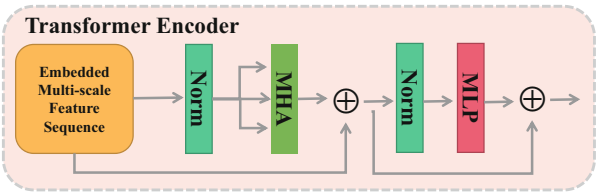

Fig. 2. The structure of the Transformer Encoder, including Multi-Head Attention (MHA) and Multi-Layer Perceptron (MLP) blocks.

Global Context Transformer as Encoder. We use the encoded image patches and the multi-scale feature maps patches as the input of Transformer Encoder [21]. The structure of the Transformer Encoder is shown in Fig. 2. The Transformer based on the self-attention mechanism can solve the limitation that CNN cannot model long-range dependence. For specific details, we suggest that readers review [21]. As an advantage, CNN has inherent hierarchical feature maps, and feature maps at different levels contain different information. The high-level feature maps contain more semantic information and the low-level feature maps contain more detailed information. In order to combine the advantages

of CNN and transformer, we fuse multi-level information to obtain better image representation, which is fed into the transformer. Better feature representation feeds can further stimulate the global context modeling ability of transformer. Compared with simply applying transformer directly, our Global Context Transformer Encoder can achieve a higher segmentation score. We will discuss this in the ablation experiment.

3.2 Decoder with Adaptive Context Fusion Module

Effectively recovering the high-level semantic feature maps extracted from the encoder is very important to the decoder, which can affect the quality of the segmentation results significantly. Due to different receptive fields, feature maps of different scales contain different levels of context information. However, the inconsistency between different feature scales limits the fusion of context information of each scale. The lack of cross-scale context leads to inflexible in processing lesions with complex boundaries because it is difficult to adapt to changes in the scale and pathological environment of the lesion or cell. In order to solve the above problems, we propose the Adaptive Context Feature Module (ACFM), which can make full use of the context and semantic information of high-level feature maps and the detailed and local information of low-level feature maps. The learnable weight parameters allow ACFM to adaptively fuse cross-scale context information with low computing cost.

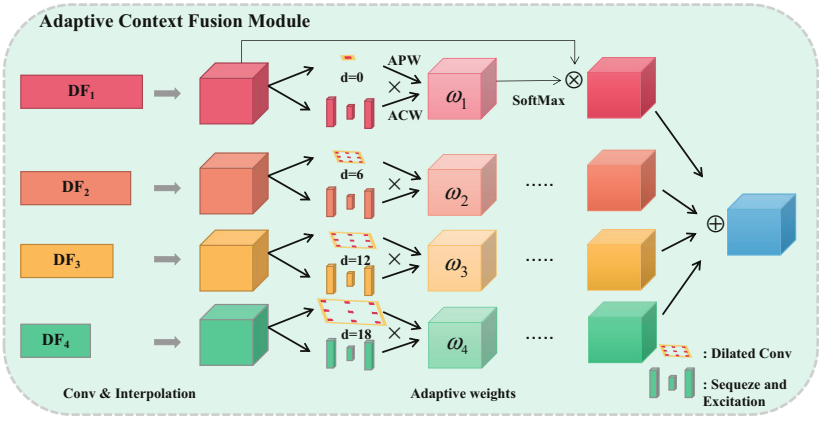

Fig. 3. The details of the Adaptive Context Fusion Module.

The detailed structure of ACFM is shown in Fig. 3. We cascade four feature decoding blocks, each feature decoding block contains a 1×1 convolution, a 3×3 deconvolution, and a 1×1 convolution. Then we obtain the Decode Feature maps DF_1, DF_2, DF_3, and DF_4 after the skip connection and the feature decoding blocks. For adaptive learning context fusion, we decouple adaptive weights (ω_i)

into Adaptive Spatial Weights (APW) and Adaptive Channel Weights (ACW). We learn the APW and ACW of different scales through carefully designed modules, and multiply them with the feature maps of the corresponding resolutions and then concatenate the results. Specifically, in order to obtain context information of different receptive fields, we use a 1×1 convolution, and three dilated convolutions with dilation rate of 6, 12, and 18 respectively to process the DF_1, DF_2, DF_3, and DF_4. As shown in Fig. 3, then the obtained initial spatial weights are softmax normalized to obtain the APW. Similarly, we learn the corresponding ACW by applying the Sequeze and Excitation[10] to DF_1, DF_2, DF_3, and DF_4. ω_i is obtained by multiplying APW and ACW. This softmax process is shown in the following formula:

$$\omega_i = \frac{e^{\omega_i}}{\sum_i^4 e^{\omega_i}} \tag{2}$$

The four feature maps obtained by multiplication are added to obtain adaptive context aggregation feature maps, and the final segmentation results are obtained after two convolution layers processing.

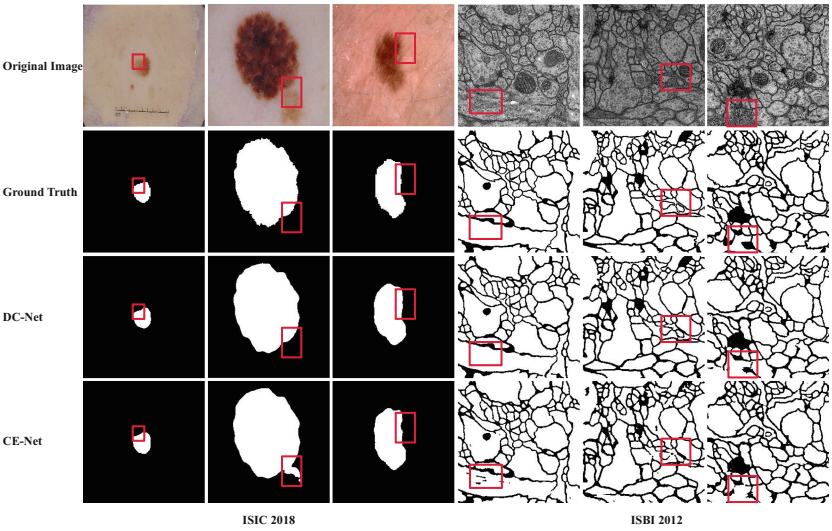

Fig. 4. Visual comparison of DC-Net and CE-Net on ISIC 2018 dataset and ISBI 2012 dataset. From top to bottom are the original image, the ground truth, the segmentation result of our DC-Net, and the segmentation result of CE-Net.

4 Experiments and Discussion

4.1 Experimental Settings

Dataset and Evaluation. The ISIC 2018 skin lesion segmentation dataset [2] is annotated by experienced dermatologists, and the goal is to automatically segment melanoma from dermoscopic images. This data set contains 2594 skin lesion images and their corresponding ground truth. We randomly divide the data set into training set, validation set and test set by 70%, 10%, and 20%. For all experiments, we apply simple data augmented methods such as flipping and random rotation.

The ISBI 2012 [1] is a cell segmentation dataset. This dataset with ground truth contains 30 images (512 × 512 pixels). We augmented all 30 images of the ISBI training set to obtain 300 images. We use 240 of them as the training set and 60 as the testing set. For all experiments, we use the same data set settings. DICE and IoU are used to evaluate our proposed method.

Loss Function and Implementation Details. We use the Dice loss function, which is widely used in the field of medical image segmentation.

$$L_{dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{3}$$

Where $|X|$ and $|Y|$ represent the number of ground truth and predicted image elements respectively. We use ViT [21]'s Transformer Encoder with 12 layers and hidden layer $D = 768$. The model is trained with adam optimizer with learning rate of 0.0002. For the original image, initial feature maps, EF_2 and EF_4, patch sizes P are set to 32, 8, 4, 1 respectively to make the length of the input sequence consistent. All experiments are run on a single NVIDIA TITAN V.

4.2 Evaluation on ISIC 2018 and ISBI 2012

We first apply our DC-Net to skin lesion segmentation. Due to the low contrast of skin lesions and the huge variation of melanoma, the task of skin lesion segmentation is extremely challenging. In order to verify the superiority of our proposed DC-Net, we compared the state-of-the-arts on Table 1 under the same experimental environment and data set settings. Compared with CE-Net, the Dice

Table 1. Evaluation on ISIC 2018 and ISBI 2012.

ISIC 2018						
Method	U-Net [16]	Deeplabv3+ [3]	Attention UNet [15]	CE-Net [8]	CA-Net [7]	DC-Net
Dice	0.885	0.893	0.898	0.922	0.921	**0.943**
IoU	0.778	0.790	0.801	0.873	0.871	**0.896**

ISBI 2012							
Method	U-Net [16]	Deeplabv3+ [3]	[15]	U-Net++ [23]	CE-Net [8]	Inf-Net [6]	DC-Net
Dice	0.935	0.873	0.933	0.939	0.939	0.945	**0.963**
IoU	0.878	0.775	0.874	0.884	0.886	0.896	**0.930**

score of our DC-Net increased from 0.921 to 0.943, and the IoU score increased from 0.871 to 0.896 on Table 1. This proves the effectiveness of our proposed method for the skin lesion segmentation task (Fig. 4).

In addition to the segmentation at the skin level, the segmentation at the cell level is also very important for evaluating the performance of the model. We apply our Dual Context Network to the task of cell contour segmentation and compare DC-Net with a series of state-of-the-arts. From Table 1, our DC-Net got the highest Dice and IoU, reaching 0.963 and 0.930 respectively. Compared with CE-Net, the segmentation performance of DC-Net has been greatly improved, and IoU has increased by 4.96%. In addition, the average hausdorff distance [11] of our DC-Net is more competitive than that of advanced Inf-Net and CE-Net on ISIC 2018 (27.85mm vs 32.34mm vs 37.82mm). This further proves the advantages of our DC-Net with GCTE and ACFM over other advanced methods. Some visualization examples are shown in Fig. 5. Due to the global context enhancement of the GCTE and the effective integration of cross-scale information by ACFM, our DC-Net suppresses conflicting information, and the obtained results have more accurate boundaries and less noise.

4.3 Ablation Study

In this section, we prove through experiments on ISIC 2018 and ISBI 2012 that our proposed GCTE and ACFM can improve the performance of encoding and decoding, respectively, and our DC-Net can greatly improve the performance of medical image segmentation.

Table 2. Ablation study on ISIC 2018 and ISBI 2012.

Method	ISIC 2018			ISBI 2012		
	Dice	IoU	ACC	Dice	IoU	ACC
Backbone (Baseline)	0.9146	0.8515	0.8992	0.9382	0.8814	0.8398
Backbone + DAC & RMP (CE-Net cite[8])	0.9224	0.8735	0.9226	0.9391	0.8860	0.8475
Backbone + VIT	0.9253	0.8770	0.9253	0.9531	0.9103	0.8619
Backbone + GCTE	0.9298	0.8831	0.9302	0.9618	0.9269	0.8740
Backbone + ACFM without APW	0.9291	0.8826	0.9291	0.9622	0.9277	0.8745
Backbone + ACFM without ACW	0.9392	0.8918	0.9354	0.9624	0.9282	0.8752
Backbone + ACFM	0.9418	0.8945	0.9403	0.9628	0.9288	0.8743
Our DC-Net	**0.9429**	**0.8960**	**0.9437**	**0.9634**	**0.9302**	**0.8761**

The first four feature extraction blocks of ResNet-34 [9] are used to replace the U-Net encoder to derive our Backbone. As observed from Table 2, whether it is using GCTE alone or ACFM alone, the performance of the two datasets has

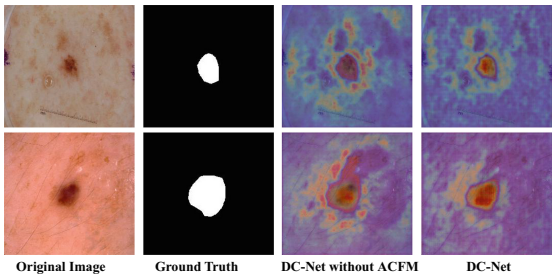

Fig. 5. Comparison of visual attention weight maps of DF_2 between DC-Net without ACFM and DC-Net.

been greatly improved. Backbone + VIT [21] refers to the last feature layer of the encode of backbone as the input of VIT. The result proves that the performance of our proposed GCTE is better than Backbone + VIT. Backbone + ACFM without APW means to add ACFM without Adaptive Spatial Weights on the basis of the backbone. Similarly, Backbone + ACFM without ACW means that the added ACFM without Adaptive Channel Weights. Both ACW and APW can improve segmentation performance significantly, and the APW contributes a lot to the improvement of network performance. The ablation experiment verified the importance of effective global context aggregation and adaptive feature fusion, indicating that our DC-Net with GCTE and ACFM significantly improves the performance of the baseline model.

5 Conclusion

In this work, we propose the DC-Net with Global Context Transformer Encoder and Adaptive Context Fusion Module for medical image segmentation. As pointed out, CNN-based U-Net lacks a summary of the global context, and the integration of multi-scale contexts is also worthy of further improvement. We carefully designed Global Context Transformer Encoder to capture the global context with transformer pipeline and fusion inherent multi-level feature maps of encoder. We propose Adaptive Context Fusion Module to adaptively fuse multi-scale context information to obtain better feature representation. Experiments on multiple medical image segmentation tasks show that our DC-Net significantly exceeds the previous methods.

Acknowledgement. This work is supported by National Key R&D Program of China (No. 2020YFC2008500, No. 2020YFC2008503), and the Open Research Fund of Key Laboratory of Space Utilization, Chinese Academy of Sciences (No. LSU-KFJJ-2020-04), National Natural Science Foundation of China (61971418, 91646207).

References

1. Cardona, A., et al.: An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. PLoS Biol. **8**(10), e1000502 (2010)
2. Challenge, I.: Isic challenge. https://challenge2018.isic-archive.com/
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: The European Conference on Computer Vision (ECCV) (September 2018)
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. Fan, D.P., et al.: Inf-net: automatic covid-19 lung infection segmentation from CT images. IEEE Trans. Med. Imaging **39**, 2626–2637 (2020)
7. Gu, R., et al.: Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging **40**(2), 699–711 (2020)
8. Gu, Z., et al.: Ce-net: context encoder network for 2d medical image segmentation. IEEE Trans. Med. Imaging **38**(10), 2281–2292 (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
11. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. **15**(9), 850–863 (1993)
12. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
14. Milletari, F., Navab, N., Ahmadi, S.: V-net: fully convolutional neural networks for volumetric medical image segmentation. CoRR abs/1606.04797 (2016)
15. Oktay, O., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
16. 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
17. Vaswani, A., et al.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
18. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Scaled-yolov4: Scaling cross stage partial network. arXiv preprint arXiv:2011.08036 (2020)

19. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
20. Wang, Z., Zou, N., Shen, D., Ji, S.: Non-local u-nets for biomedical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6315–6322 (2020)
21. Wu, B., et al.: Visual transformers: Token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677 (2020)
22. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning, pp. 7354–7363. PMLR (2019)
23. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. CoRR abs/1807.10165 (2018)