# CONVFORMER: COMBINING CNN AND TRANSFORMER FOR MEDICAL IMAGE SEGMENTATION

*Pengfei Gu*[†⋆] *Yejia Zhang*[†⋆] *Chaoli Wang*[†] *Danny Z. Chen*[†]

[†]University of Notre Dame, Department of Computer Science and Engineering, Notre Dame, IN, USA

## ABSTRACT

Convolutional neural network (CNN) based methods have achieved great successes in medical image segmentation, but their capability to learn global representations is still limited due to using small effective receptive fields of convolution operations. Transformer based methods are capable of modelling long-range dependencies of information for capturing global representations, yet their ability to model local context is lacking. Integrating CNN and Transformer to learn both local and global representations while exploring multi-scale features is instrumental in further improving medical image segmentation. In this paper, we propose a hierarchical CNN and Transformer hybrid architecture, called ConvFormer, for medical image segmentation. ConvFormer is based on several simple yet effective designs. (1) A feed forward module of Deformable Transformer (DeTrans) is re-designed to introduce local information, called Enhanced DeTrans. (2) A residual-shaped hybrid stem based on a combination of convolutions and Enhanced DeTrans is developed to capture both local and global representations to enhance representation ability. (3) Our encoder utilizes the residual-shaped hybrid stem in a hierarchical manner to generate feature maps in different scales, and an additional Enhanced DeTrans encoder with residual connections is built to exploit multi-scale features with feature maps of different scales as input. Experiments on several datasets show that our ConvFormer, trained from scratch, outperforms various CNN- or Transformer-based architectures, achieving state-of-the-art performance.

## 1. INTRODUCTION

Image segmentation is a central problem in medical image analysis. Convolutional neural networks (CNNs), especially fully convolutional networks (FCNs), have become predominant approaches for this problem (e.g., U-Net [1], UNet 3+ [2], etc). Despite their successes, CNNs still have yet to address the limitation in learning long-range dependencies (global information) to see a "big picture" due to limited effective receptive fields (ERFs) of convolution (Conv) operations. Attempts were made to enlarge ERFs, such as utilizing atrous Convs with different dilated rates (e.g., [3]), applying pyramid pooling (e.g., [4]), and designing large kernels (e.g., [5]). Although these methods enlarged ERFs to some extent, they still suffered from limited ERFs, yielding sub-optimal image segmentation accuracy.

Recently, Transformers (e.g., vision Transformer [6]) became a de-facto choice for modelling long-range dependencies in computer vision, inspired by their success with self-attention mechanism in natural language processing. Compared to CNN methods, Transformer models have larger receptive fields and excel at learning global information. But, they also have drawbacks, e.g., high computation cost, slow convergence, and short of CNN's inductive biases. Two types of methods attempted to reduce their computation cost. (1) Limiting self-attention to local windows (e.g., [7, 8]). (2) Downsampling the key and value feature maps (e.g., [9]). Though effective in capturing global information, these Transformer-based methods still yielded unsatisfactory performance due to deficiency in learning local information.

In medical image segmentation, efforts were made to combine Transformer and CNN to learn both local and global representations. In [10], Transformer was utilized as a supplement, appended to the last Conv block of the CNN encoder to learn global information. MedT [11] exploited local and global information by employing two branches, where a gated axial Transformer was applied to explore global information and CNN was leveraged to learn local information. CoTr [12] employed Deformable Transformer (DeTrans) as an additional encoder for exploring multi-scale information from multi-scale feature maps for 3D medical image segmentation. MISSFormer [13] introduced a Conv layer to Transformer to enhance its capability to learn local information. UNETR [14] proposed a U-shaped encoder-decoder network, where Transformer was used as the encoder to capture global multi-scale information and a CNN decoder was used to compute final segmentation output. Although achieving promising performances, these methods still incurred several drawbacks. (1) Limited capability to learn both local and global representations due to ineffective integration of CNN and Transformer. For example, CoTr [12] did not apply Conv to DeTrans for effectively learning local information. TransUNet [10] simply appended Transformer to a Conv block in the encoder. UNETR [14] did not leverage Conv in the

⋆ Equal Contribution.

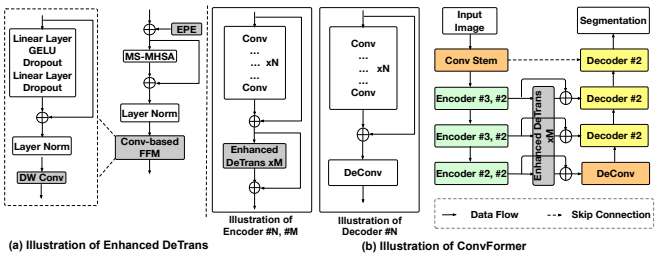


**Fig. 1**. (a) An overview of our Enhanced DeTrans with the proposed enhanced positional encoding (EPE) and Conv-based feed forward module (FFM) (shown in gray color). (b) Our ConvFormer: The two big boxes on the left are for the encoder and decoder, respectively (best viewed in color).

encoder. (2) Deficiency in capturing multi-scale information. For instance, MedT [11] did not explore multi-scale information in its global Transformer branch, and such drawbacks in learning both local and global representations and capturing multi-scale information led to sub-optimal segmentation.

In this paper, we propose a new hierarchical CNN and Transformer hybrid architecture, called ConvFormer, to capture both local and global representations while exploiting multi-scale features for medical image segmentation. Specifically, ConvFormer is based on several key designs. (1) We re-design the feed forward module of DeTrans to introduce local information, called Enhanced DeTrans. (2) We develop a residual-shaped hybrid stem based on a combination of Convs and Enhanced DeTrans to capture both local and global representations to enhance representation ability. (3) The residual-shaped hybrid stem is utilized in the encoder in a hierarchical manner to generate feature maps in different scales, and an additional Enhanced DeTrans encoder with residual connections is built to explore multi-scale features with feature maps of different scales as input. (4) An enhanced positional encoding (EPE) introduces a Conv layer to the fixed sinusoidal encoding method [15] to improve adaptability and flexibility.

Extensive experiments on two public datasets (2017 ISIC Skin Lesion segmentation (2D) [16] and MM-WHS CT (3D) [17]) and one in-house dataset (lymph node segmentation (2D)) show that our ConvFormer, trained from scratch, outperforms various known CNN-based or Transformer-based methods, achieving state-of-the-art performance.

## 2. METHOD

Fig. 1 shows the architecture of our ConvFormer that contains three main components: (1) Enhanced DeTrans that captures global representations with local information, using a Conv-based feed forward module (FFM); (2) the residual-shaped hybrid stem that extracts local and global representations in different scales that are fed to an additional Enhanced DeTrans encoder to explore multi-scale features; (3) enhanced positional encoding (EPE) that improves the adaptability and flexibility of the fixed sinusoidal encoding method [15].

### 2.1. Enhanced DeTrans

FFM is an essential component of Transformer to enhance the representation ability. The FFM proposed in DeTrans [8, 12] consists of a multilayer perceptron (MLP) composing of two linear layers separated by GELU, as follows:

$$\text{FFM}(x) = \text{LN}\left(\text{GELU}(xW_1 + b_1)W_2 + b_2 + x\right), \quad (1)$$

where $x$ is an input feature map, $W_1$ and $W_2$ are weights of the two linear layers respectively, $b_1$ and $b_2$ are bias terms, and LN denotes layer norm. For simplicity, Dropouts are omitted. Though effective, this design does not have good capability to learn local information, which is critical in dense predictions.

To overcome this limitation and enable FFM to introduce local information to the global representations captured by multi-scale multi-head self attention (MS-MHSA), we resort to a Conv layer. Specifically, we insert depth-wise convolution (DW Conv) to the end of FFM. The resulted FFM is called Conv-based FFM. Note that we leverage DW Conv instead of Conv for reducing computation cost. As a result, the proposed Conv-based FFM inherits the merit of both CNN and DeTrans in learning local and global representations.

To adapt DW Conv to FFM, we first reshape the input 1D sequences captured by FFM to 2D/3D feature maps, apply DW Conv to the reshaped feature maps to learn local information, and reshape the result back to 1D sequences as output:

$$\text{Conv-based FFM}(x) = \text{reshape}(\text{DW Conv}(\text{reshape}(\text{FFM}(x)))), \quad (2)$$

where $x$ is an input feature map. Note that our Conv-based FFM is capable of dealing with multi-scale feature map input, which is important for accurate segmentation. DW Conv is shared when processing multi-scale feature maps.

We replace the FFM of DeTrans by our Conv-based FFM, and the resulted DeTrans is called Enhanced DeTrans. Enhanced DeTrans is capable of capturing both local and global information by combining Conv and Transformer.

### 2.2. Residual-shaped Hybrid Stem

Stacking Transformer to the last Conv block as a global feature extractor is a common way to combine CNN/Transformer in encoder to learn local and global representations. For example, in TransUNet [10] and CoTr [12], Transformer was followed by a Conv block to capture global information for local features obtained by Conv blocks. Although such methods might enhance the models' representation ability on learning local and global information to some extent, they did not exploit both local and global features effectively. More importantly, they did not generate and explore multi-scale local and global features, which are highly important to handle datasets with large variations in object size, shape, and texture, which are common characteristics of medical image datasets.

To effectively capture both local and global representations, we propose the residual-shaped hybrid stem, which

consists of two key designs (see the leftmost box in Fig. 1(b)). (1) Residual connections are introduced to the stacked Convs for feature diversity and delivery. (2) Local representations captured by the stacked Convs are combined with global representations learned by Enhanced DeTrans via an add operation. These two designs offer two compelling advantages: (i) The introduced residual connections are critical for avoiding feature collapse (i.e., feature diversity is continuously reduced as the layer depth increases); (ii) the captured local and global representations can be better fused via add operation. The improvement (e.g., in Table 4, F1 improved by $1.0\%$, $p < 0.05$, t-test) shows the importance of the introduced residual connections.

We use the residual-shaped hybrid stem in the encoder in a hierarchical manner, so that it can generate global and local representations in different scales. Given an input image, we can obtain hierarchical global and local representations with different resolutions, which are then fed to an additional Enhanced DeTrans encoder to exploit multi-scale representations. Different from CoTr [12] which utilized DeTrans to exploit multi-scale features, we add residual connections to our Enhanced DeTrans to reuse the local representations captured by the stacked Convs and consolidate the captured local and global representations.

The architecture of our ConvFormer is shown in Fig. 1. Following the architecture of CoTr [12], ConvFormer has four encoder and decoder stages, with a Conv stem and three residual-shaped hybrid stems in encoder, a deconvolution (DeConv) stem and three decoder stems in decoder, and an additional Enhanced DeTrans encoder.

### 2.3. Enhanced Positional Encoding (EPE)

In Transformer, we first flatten input feature maps into 1D sequences. But, this flattening process may cause loss of spatial information that is critical for segmentation. Known methods [12, 15] employed fixed sinusoidal encoding to supplement flattened sequences with position information. In particular, sine and cosine functions of different frequencies were used to compute the positional coordinates of each dimension:

$$\mathrm{PE}_{(pos,2i)} = \sin\left(pos/10000^{2i/C}\right), \tag{3}$$

$$\mathrm{PE}_{(pos,2i+1)} = \cos\left(pos/10000^{2i/C}\right), \tag{4}$$

where $pos$ is for position, $i$ is the dimension, and $C$ is a constant. However, such a fixed sinusoidal encoding method lacks adaptability and flexibility, as the code (e.g., frequencies) is predefined.

To improve adaptability and flexibility, we introduce DW Conv to the fixed sinusoidal encoding [15], since a Conv kernel naturally encodes pixel locality and semantic continuity with adaptability and flexibility. The resulted positional encoding method is called enhanced positional encoding (EPE). Specifically, it uses two branches: one branch applies the fixed sinusoidal encoding [15] to learn position information, and the other uses DW Conv to capture position information. The two branches are combined by an add operation:

$$x' = \text{fixed sinusoidal encoding}(x) + \mathrm{ReLU}(\mathrm{BN}(\text{DW Conv}(x))), \tag{5}$$

where $x$ is an input feature map, $x'$ denotes the output feature map embedded with position information, and BN is batch normalization. The improvement (in Table 4, F1 improved by $0.4\%$, $p < 0.05$, t-test) demonstrates the importance of EPE.

**Table 1**. Quantitative results of various models on the lymph node dataset. The reported values are average±STD for 3 runs with different random seeds. The best results are in **bold**.

| Method | IoU | Precision | Recall | F1 Score |
|---|---|---|---|---|
| U-Net [1] | 0.661 | 0.834 | 0.761 | 0.796 |
| Zhang et al. [20] | 0.810 | 0.901 | 0.889 | 0.895 |
| $k$CBAC-Net [21] | 0.829 | 0.909 | 0.904 | 0.906 |
| ConvFormer (Ours) | **0.845**±0.002 | **0.925**±0.002 | **0.907**±0.002 | **0.916**±0.002 |

## 3. EXPERIMENTS AND RESULTS

**Datasets.** (1) **The lymph node segmentation ultrasound dataset:** This in-house dataset is for segmenting lymph nodes in 2D ultrasound images. It contains 137 training and 100 test images. (2) **The 2017 MM-WHS CT dataset:** This public dataset [17] is for segmenting seven cardiac structures, the left/right ventricle blood cavity (LV/RV), left/right atrium blood cavity (LA/RA), myocardium of the left ventricle (LV-myo), ascending aorta (AO), and pulmonary artery (PA). It contains 20 unpaired 3D CT images, which are randomly split into 16 images and 4 images for training and testing, following [18]. (3) **The 2017 ISIC skin lesion segmentation dataset:** This public dataset [16] is for segmenting lesion boundaries in 2D dermoscopic images. It contains 2000 training, 150 validation, and 600 test images.

**Implementation Details.** Our ConvFormer is implemented with PyTorch, and is trained on an NVIDIA Tesla P100 Graphics Card with 16GB GPU memory using the AdamW [19] optimizer with weight decay $0.005$. We apply the "poly" learning rate policy with an initial learning rate of $2e-4$. The maximum number of iterations is $100$k for the lymph node and 2017 ISIC skin lesion datasets, and $240$k for the 2017 MM-WHS CT dataset (using about $34$, $48$, and $96$ training hours, respectively). To avoid overfitting, standard data augmentation (e.g., random flip, crop, etc) is applied.

**Experimental Results.** Table 1 reports quantitative results on the lymph node dataset, showing that our ConvFormer outperforms the known methods by a clear margin in all the evaluation measures. In particular, ConvFormer outperforms the state-of-the-art (SOTA) method, $k$CBAC-Net [21], by $1.6\%$ and $1.0\%$ in IoU and F1, respectively, achieving new SOTA performances. Table 2 shows quantitative results on the 2017 MM-WHS CT 3D dataset. Our ConvFormer outperforms the best-known CNN-based (HFA-Net [18], $k$CBAC-Net [21]) and Transformer-based (CoTr [12], UNETR [14]) methods, achieving new SOTA performances. Specifically, ConvFormer outperforms HFA-Net [18], $k$CBAC-Net [21],

**Table 2**. Quantitative results of different models on the 2017 MM-WHS CT dataset. "—" means that the results were not reported by the original papers, "Para." means the number of parameters of the model, and "HDF" represents Hausdorff.

| Method | Para. | Metric | LV | RV | LA | RA | LV -myo | AO | PA | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| Payer et al. [23] | — | Dice | 0.918 | 0.909 | 0.929 | 0.888 | 0.881 | 0.933 | 0.840 | 0.900 |
| Dou et al. [24] | — | Dice | 0.888 | — | 0.891 | — | 0.733 | 0.813 | — | — |
| Chen et al. [25] | — | Dice | 0.919 | — | 0.911 | — | 0.877 | 0.927 | — | 0.909 |
| HFA-Net [18] | — | Dice | 0.946 | 0.893 | 0.925 | 0.897 | 0.910 | 0.964 | 0.830 | 0.909 |
| | | IoU | 0.898 | 0.810 | 0.861 | 0.816 | 0.836 | 0.930 | 0.722 | 0.839 |
| | | HDF | 7.148 | 33.128 | 42.173 | 22.903 | 36.954 | 12.075 | 37.845 | 27.461 |
| | | ADB | 0.076 | 0.562 | 0.210 | 0.334 | 0.225 | 0.103 | 1.685 | 0.456 |
| *k*CBAC-Net [21] | 134M | Dice | 0.951 | 0.902 | 0.938 | 0.911 | 0.922 | 0.974 | 0.837 | 0.919 |
| | | IoU | 0.907 | 0.825 | 0.883 | **0.838** | 0.855 | 0.949 | 0.734 | 0.856 |
| | | HDF | 5.500 | 14.940 | 12.403 | 15.081 | 7.337 | 6.848 | 32.499 | 13.516 |
| | | ADB | 0.074 | 0.285 | 0.163 | 0.248 | 0.119 | 0.059 | 1.403 | 0.336 |
| CoTr [12] | 42M | Dice | 0.944 | 0.896 | 0.934 | 0.898 | 0.908 | 0.953 | **0.845** | 0.911 |
| | | IoU | 0.895 | 0.816 | 0.876 | 0.816 | 0.832 | 0.912 | 0.742 | 0.841 |
| | | HDF | 5.905 | 15.819 | 12.746 | 15.767 | 10.346 | 12.330 | **31.539** | 14.922 |
| | | ADB | 0.084 | 0.352 | 0.159 | 0.285 | 0.131 | 0.118 | 1.391 | 0.360 |
| UNETR [14] | 93M | Dice | 0.950 | 0.888 | 0.939 | 0.903 | 0.919 | 0.964 | 0.843 | 0.915 |
| | | IoU | 0.896 | 0.806 | 0.868 | 0.818 | 0.854 | 0.930 | **0.744** | 0.845 |
| | | HDF | 5.869 | 15.698 | 11.769 | 14.080 | 8.816 | 11.864 | 33.392 | 14.498 |
| | | ADB | 0.080 | 0.349 | **0.151** | 0.278 | **0.115** | 0.103 | 1.404 | 0.354 |
| ConvFormer (Ours) | 61M | Dice | **0.955** | **0.910** | **0.943** | **0.914** | **0.926** | **0.980** | 0.844 | **0.925** |
| | | IoU | **0.914** | **0.836** | **0.885** | 0.835 | **0.857** | **0.957** | 0.740 | **0.861** |
| | | HDF | **5.089** | **12.873** | **11.745** | **13.360** | **7.000** | **6.823** | 34.560 | **13.064** |
| | | ADB | **0.069** | **0.264** | 0.175 | **0.242** | 0.120 | **0.030** | **1.359** | **0.323** |

**Table 3**. Quantitative results of different models on the 2017 ISIC skin lesion dataset.

| Method | Jaccard Index | Dice | Sensitivity | Specificity |
|---|---|---|---|---|
| Yuan et al. [22] | 0.765 | 0.849 | 0.825 | 0.975 |
| Li et al. [26] | 0.765 | 0.866 | 0.825 | 0.984 |
| Lei et al. [27] | 0.771 | 0.859 | 0.835 | 0.976 |
| Mirikharaji et al. [28] | 0.773 | 0.857 | 0.855 | 0.973 |
| Xie et al. [29] | 0.788 | 0.868 | **0.884** | 0.957 |
| *k*CBAC-Net [21] | 0.794 | 0.887 | 0.847 | 0.984 |
| ConvFormer (Ours) | **0.797**±0.003 | **0.889**±0.002 | 0.846±0.002 | **0.986**±0.001 |

CoTr [12], and UNETR [14] by 1.6%, 0.6%, 1.4%, and 1.0% in Dice score on average, respectively. Table 3 gives quantitative results on the 2017 ISIC skin lesion dataset. Our ConvFormer attains accuracy gain of 3.2% in Jaccard Index over the 2017 ISIC Challenge winner [22] and slightly outperforms the SOTA method, *k*CBAC-Net [21], achieving new SOTA performances. Note that, (1) our ConvFormer shows excellent generalization on both 2D and 3D datasets, on ultrasound, dermoscopic, and CT images, outperforming previous SOTA methods; (2) ConvFormer achieves competitive performances without pre-training, different from most of the known Transformer-based models. All these suggest that our ConvFormer is a promising and robust method suitable for medical image segmentation tasks on datasets of different imaging characteristics and modalities, and it is capable of learning effective local, global, and multi-scale representations. Fig. 2 shows some qualitative results.

**Ablation Study.** We conduct an ablation study to examine the effects of different components in our ConvFormer using the lymph node dataset. As Table 4 shows, (1) when using the fixed sinusoidal encoding [15] for positional encoding (the resulted ConvFormer is denoted by ConvFormer w/o EPE), the F1 score drops by 0.4%; (2) when removing the additional Enhanced DeTrans encoder that aims to explore multi-scale information from ConvFormer w/o EPE (denoted by ConvFormer w/o Additional Enhanced DeTrans), the F1 score drops by 0.5%; (3) when removing Conv-based FFM from ConvFormer w/o Additional Enhanced DeTrans (denoted by ConvFormer w DeTrans), the F1 score drops by 0.4%; (4) when completely removing DeTrans (the resulted network is

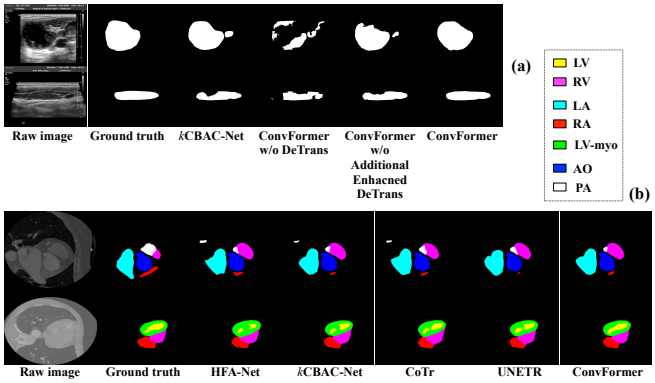


**Fig. 2**. Some visual qualitative results on the lymph node dataset (a) and the 2017 MM-WHS CT dataset (b), demonstrating the capability of our ConvFormer.

**Table 4**. Ablation study of the effects of different components in ConvFormer using the lymph node dataset.

| DeTrans | Conv-based FFM | Addi. E-DeTrans | EPE | Method | IoU | Precision | Recall | F1 |
|---|---|---|---|---|---|---|---|---|
| | | | | ConvFormer w/o DeTrans | 0.673 | 0.833 | 0.782 | 0.807 |
| √ | | | | ConvFormer w DeTrans | 0.825 | 0.906 | 0.900 | 0.903 |
| √ | √ | | | ConvFormer w/o Additional Enhanced DeTrans | 0.830 | 0.908 | 0.906 | 0.907 |
| √ | √ | √ | | ConvFormer w/o EPE | 0.839 | 0.912 | 0.910 | 0.912 |
| √ | √ | √ | √ | ConvFormer | **0.845** | **0.925** | **0.907** | **0.916** |
| ConvFormer w/o residual connections | | | | | 0.829 | 0.910 | 0.902 | 0.906 |

a pure CNN baseline, denoted by ConvFormer w/o DeTrans), the F1 score drops by 9.6%. These effects demonstrate the importance of our proposed components for better capturing local, global, and multi-scale information for accurate image segmentation. Besides, we remove the residual connections of the residual-shaped hybrid stem (denoted by ConvFormer w/o residual connections), and the F1 score drops by 1.0%. This validates the importance of the residual connections.

## 4. CONCLUSIONS

In this paper, we proposed a new hierarchical CNN and Transformer hybrid architecture, ConvFormer, for medical image segmentation. ConvFormer is capable of well learning local, global, and multi-scale representations by introducing Conv-based FFM, residual-shaped hybrid stem, and an additional Enhanced DeTrans encoder with residual connections. Moreover, we presented an enhanced positional encoding to improve the adaptability and flexibility of the fixed sinusoidal encoding method. Experiments on 2D and 3D datasets of different imaging characteristics and modalities demonstrated the effectiveness of ConvFormer.

## 5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by two publicly available datasets [17, 16] and one in-house dataset. Ethical approval was not required as confirmed by the licenses attached with the open access datasets.

## 6. ACKNOWLEDGEMENTS

This research was supported in part by NSF grants CNS-1629914, DUE-1833129, IIS-1955395, IIS-2101696, and OAC-2104158.

## 7. REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[2] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP*, 2020, pp. 1055–1059.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE PAMI*, vol. 40, no. 4, pp. 834–848, 2018.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017, pp. 6230–6239.

[5] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters – improve semantic segmentation by global convolutional network," in *CVPR*, 2017, pp. 1743–1751.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10012–10022.

[8] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *ICLR*, 2021.

[9] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021, pp. 568–578.

[10] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[11] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical Transformer: Gated axial-attention for medical image segmentation," in *MICCAI*, 2021, pp. 36–46.

[12] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging cnn and transformer for 3D medical image segmentation," in *MICCAI*, 2021, pp. 171–180.

[13] X. Huang, Z. Deng, D. Li, and X. Yuan, "MISSFormer: An effective medical image segmentation transformer," *arXiv preprint arXiv:2109.07162*, 2021.

[14] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *WACV*, 2022, pp. 1748–1758.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[16] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *IEEE, ISBI*, 2018, pp. 168–172.

[17] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI," *Medical Image Analysis*, vol. 31, pp. 77–87, 2016.

[18] H. Zheng, L. Yang, J. Han, Y. Zhang, P. Liang, Z. Zhao, C. Wang, and D. Z. Chen, "HFA-Net: 3D cardiovascular image segmentation with asymmetrical pooling and content-aware fusion," in *MICCAI*, 2019, pp. 759–767.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[20] Y. Zhang, M. T. Ying, and D. Z. Chen, "Decompose-and-integrate learning for multi-class segmentation in medical images," in *MICCAI*, 2019, pp. 641–650.

[21] P. Gu, H. Zheng, Y. Zhang, C. Wang, and D. Z. Chen, "$k$CBAC-Net: Deeply supervised complete bipartite networks with asymmetric convolutions for medical image segmentation," in *MICCAI*, 2021, pp. 337–347.

[22] Y. Yuan and Y.-C. Lo, "Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks," *IEEE JBHI*, vol. 23, no. 2, pp. 519–526, 2019.

[23] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Multi-label whole heart segmentation using CNNs and anatomical label configurations," in *International Workshop, STACOM*, 2017, pp. 190–198.

[24] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, "Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss," in *IJCAI*, 2018, pp. 691–697.

[25] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation," *IEEE TMI*, vol. 39, no. 7, pp. 2494–2505, 2020.

[26] H. Li, X. He, F. Zhou, Z. Yu, D. Ni, S. Chen, T. Wang, and B. Lei, "Dense deconvolutional network for skin lesion segmentation," *IEEE JBHI*, vol. 23, no. 2, pp. 527–537, 2019.

[27] B. Lei, Z. Xia, F. Jiang, X. Jiang, Z. Ge, Y. Xu, J. Qin, S. Chen, T. Wang, and S. Wang, "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Medical Image Analysis*, vol. 64, pp. 101716, 2020.

[28] Z. Mirikharaji and G. Hamarneh, "Star shape prior in fully convolutional networks for skin lesion segmentation," in *MICCAI*, 2018, pp. 737–745.

[29] Y. Xie, J. Zhang, H. Lu, C. Shen, and Y. Xia, "SESV: Accurate medical image segmentation by predicting and correcting errors," *IEEE TMI*, vol. 40, no. 1, pp. 286–296, 2021.