2E-Net: Unet-like Double Encoder for Medical Image Segmentation

Jiahao Li[1], Yanbing Liang[2,*]
[1]Hebei Key Laboratory of Data Science and Application, North China University of Science and Technology, Tangshan, Hebei, China
[2]College of Science, North China University of Science and Technology, Tangshan, Hebei, China
**Corresponding Author*

Abstract: As an important branch in the field of medical image segmentation, abdominal multi-organ segmentation plays a crucial role in disease diagnosis and treatment supervision. Through long-term research in the field of medical image segmentation, a series of medical image segmentation models represented by U-Net have made significant progress. There are many types of abdominal organs with different sizes and blurred boundaries, so the existing models still fail to achieve satisfactory segmentation results. Most existing abdominal multi organ segmentation methods use pure CNN, pure Transformer, or a combination of both to construct the entire network structure, which cannot better leverage the advantages of both. In order to solve these problems, we propose a double encoder medical image segmentation method called 2E-Net, which improves the MISSFormer method based on Transformer by introducing ResNet50. 2E-Net introduces an Encoder Fusion Module (EFM) to merge the two encoders for feature extraction. Extensive experiments on the Synapse dataset show that our proposed method outperforms the MISSFormer method, achieving a Dice Similarity Coefficient (DSC) score of 82.13%.

Keywords: Transformer; Attention Mechanism; Medical Image Segmentation; CNN

1. Introduction

With the continuous improvement in computing power, the development of deep learning is accelerating, promoting the widespread application of deep learning-based computer vision methods in the field of medical image segmentation. Abdominal multi-organ segmentation plays a crucial role in medical diagnosis and decision-making. However, the large variety of organs in the human abdomen, their overlapping nature, and the differences in size pose significant challenges for medical image segmentation.

To address these issues, a large number of deep learning methods based primarily on traditional convolutional neural networks (CNNs) [1–6] have been applied to abdominal multi-organ segmentation. Most of these methods are inspired by the U-Net [7] model, evolving into U-shaped architectures. Since these methods are mostly built on CNN architectures using 3×3 convolution kernels, they tend to focus more on the local features of the input images. With the development of the natural language processing (NLP) field, the Transformer model [8], capable of modeling from a broader global perspective, was introduced and widely adopted in NLP. Following this, in the field of computer vision (CV), Vision Transformer [9] and its various adaptations [10] were proposed and successfully applied to abdominal multi-organ medical image segmentation.

Compared to traditional CNN methods, Transformer-based approaches can model information more effectively on a global scale. However, they are weaker in extracting local spatial information, and during the compression of feature maps, the loss of local details is more pronounced. Due to the varying sizes of abdominal organs and their indistinct boundaries, the ability of a model to extract and preserve local information is crucial for achieving effective segmentation results.

For existing algorithms, there is still room for further optimization in feature extraction methods. Therefore, we propose the 2E-Net model, which combines two different

approaches, ResNet50 and Transformer, into a double encoder system to provide the decoder with more comprehensive spatial information compensation.

2. Related Work

Deep learning-based methods possess strong expressive capabilities, leading to their widespread application in medical image segmentation. Particularly since 2015, the U-Net [7] architecture, featuring a U-shaped encoder-decoder structure, has demonstrated exceptional performance in the field of medical image segmentation, resulting in extensive research and the proposal of numerous new methods based on this architecture [1,2]. With the emergence of attention mechanisms, an increasing number of researchers have begun to explore how to integrate them with CNN methods. For instance, Oktay et al. proposed the Attention U-Net [3], which incorporates attention mechanisms into the U-Net [7] framework, allowing the model to better extract task-relevant information on a global scale during both encoding and decoding. Additionally, some researchers introduced attention mechanisms into the UNet++ [1] framework, proposing the Attention UNet++ [4] method. The U-Net architecture utilizes skip connection mechanisms to obtain more semantic information from the encoder for input into the decoder. However, this mechanism extracts information directly from the unprocessed encoder, resulting in a relatively simple design. To address this, Feng et al. proposed a CNN-based multi-scale feature fusion method [6], which enriches the decoder with more diverse multi-scale semantic information.

In addition, the Transformer [8] architecture was first applied in the NLP field, achieving great success in machine translation, and has since been successfully utilized in various domains. Its powerful capability for modeling long-sequence features can compensate for the shortcomings of CNNs in global modeling. More researchers are beginning to focus on its applications in the field of computer vision [9]. Chen et al. proposed TransUNet [11], which was the first to apply Vision Transformer (ViT) [9] to medical image segmentation. This method employs a feature extraction approach that combines a ResNet50[12] model pre-trained on ImageNet and ViT in a sequential manner. Although using ResNet50 as the embedding layer can provide good encoding for ViT, this sequential approach does not effectively leverage ResNet50's ability to extract local features. Although attention mechanisms can enhance model performance in many cases, their time complexity of requires substantial computational resources. To address this, a method based on a sliding window attention mechanism, Swin Transformer [10], was introduced, which can still consider global feature extraction while reducing computational costs. Cao et al. proposed Swin-UNet [13], a pure Transformer model based on Swin Transformer that surpasses TransUNet. Both of these medical image segmentation methods employ the skip connection mechanism from the early U-Net model, which limits their ability to effectively transfer information from the encoder to the decoder. Huang et al. proposed the MISSFormer [14] method, whose basic unit is based on the Transformer module. Its skip connections utilize attention mechanisms for multi-scale feature fusion, allowing the decoder to better acquire valuable pattern information for segmentation tasks. However, MISSFormer uses a pure Transformer architecture during the encoding phase, resulting in a weaker capability for extracting local features.

3. Methods

3.1 Architecture

The proposed U-shaped double encoder medical image segmentation model, 2E-Net, has an overall architecture as shown in Figure 1. The 2E-Net model comprises an Encoder, Decoder, Encoder Fusion Module, and Enhanced Transformer Context Bridge [14]. The model's input is an image with channel depth, height, and width of 3, 224, and 224, respectively.

In the Encoder, 2E-Net employs two independent encoders consisting of ResNet50 and Enhanced Transformer [14]. The first encoder is based on the Transformer architecture. In this encoder, the Overlap Patch Merging [14] module performs downsampling, while the Enhanced Transformer module extracts features. This encoder uses a 4×4 patch size and encodes it through the Overlap Patch Embedding [14] module before inputting

it into the Enhanced Transformer module. A total of three downsampling operations and three feature extraction operations are conducted. After each passage through the Enhanced Transformer module, the extracted features are input into the corresponding Encoder Fusion Module.

The second encoder is constructed based on ResNet50. In ResNet50, there are a total of four stages. Starting from STAGE 1 of ResNet50, after each stage, the features from that stage are input into the corresponding Encoder Fusion Module. The features processed by the Encoder Fusion Module are then further input into the Enhanced Transformer Context Bridge module for fusion of features across different scales.

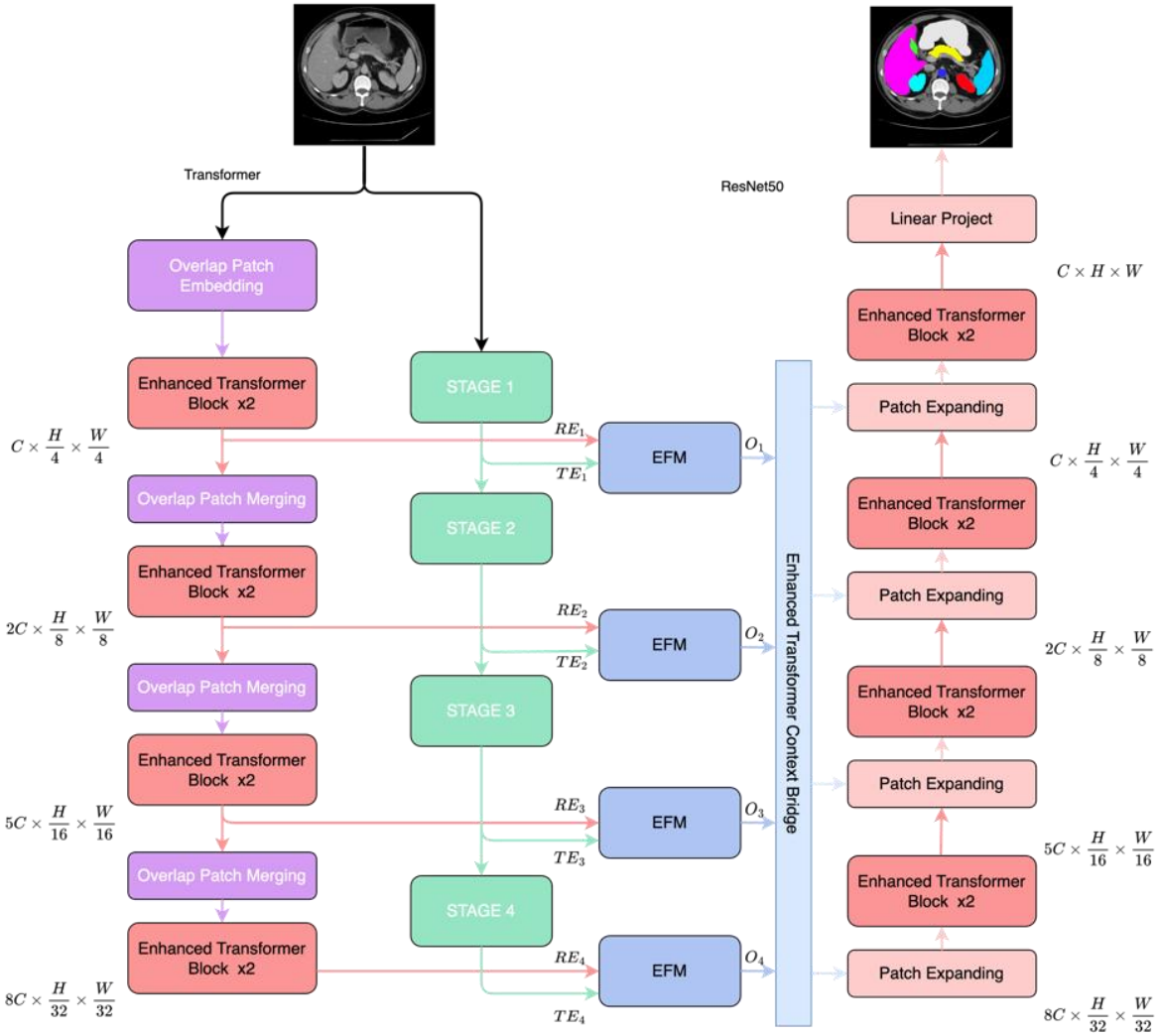

Figure 1. 2E-Net Architecture

In the Decoder, the first three pairs of Patch Expanding and Enhanced Transformer modules perform three instances of two times upsampling. Finally, a four times upsampling is conducted using a Patch Expanding module, followed by a Linear Projection to obtain the segmentation map.

3.2 Encoder Fusion Module

The proposed method, 2E-Net, adopts a double encoder structure, which necessitated the introduction of the Encoder Fusion Module, as illustrated in **Figure 2**.

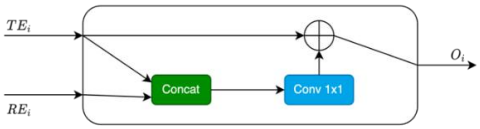

Figure 2. Encoder Fusion Module

The Encoder Fusion Module has two input ports and one output port. The input ports are the feature map produced by the Transformer encoder, $TE \in \mathbb{R}^{C_t \times H_t \times W_t}$, and the feature map produced by the ResNet50 encoder, $RE \in \mathbb{R}^{C_r \times H_r \times W_r}$. The output port is the fused feature map $O \in \mathbb{R}^{C_t \times H_t \times W_t}$. Here, C, H, W represents the channel depth, height, and width of the

feature maps, respectively.

In the Encoder Fusion Module, the feature maps $TE \in \mathbb{R}^{C_t \times H_t \times W_t}$ from the Transformer encoder and $RE \in \mathbb{R}^{C_r \times H_r \times W_r}$ from the ResNet50 encoder are first merged along the channel dimension $Conv_{1\times1}$ using a Concat operation. This combined feature map is then input into a TE module to compress the channel depth and is subsequently added to the value from the input TE, resulting in the fused feature map $O \in \mathbb{R}^{C_t \times H_t \times W_t}$ The computational logic of this module is represented as shown in Formula (1).

$$merge\_encoder = Concatenate(TE, RE, dim = 1)$$
$$output = Conv_{1\times1}(merge\_encoder) + TE \quad (1)$$

On one hand, the encoder is composed of the Transformer and ResNet50 in a sequential manner. Existing methods typically use ResNet50 as an embedding layer, while the backbone of the encoder is predominantly based on the Transformer, as seen in the TransUNet method. Because the semantic information in the decoder derives from the backbone of the encoder, this architectural pattern struggles to leverage the local feature extraction capabilities of ResNet50. On the other hand, current medical image segmentation methods often only fuse semantic information from feature maps during the downsampling process, which limits their ability to integrate diverse semantic information. The Encoder Fusion Module combines the global and local information provided by both the Transformer and ResNet50, accommodating the different scales of various organs, thereby enabling the extraction of richer semantic information.

4. Experiment

4.1 Dataset

The Synapse multi-organ segmentation dataset is used. The Synapse dataset comprises CT images from 30 cases, totaling 3,779 axial abdominal CT images. Among these, 18 cases are utilized as the training dataset, while the remaining 12 cases serve as the testing dataset. After organization, the dataset contains 8 types of abdominal organs, which are the aorta, gallbladder, left kidney, right kidney, liver, pancreas, spleen, and stomach.

4.2 Experiment Result

This paper uses the Dice Similarity Coefficient (DSC) evaluation metric to assess the proposed method, as shown in Formula (2).

$$DSC = \frac{2TP}{2TP+FN+FP} \quad (2)$$

The comparison results of the proposed method, 2E-Net, with existing advanced algorithms on the Synapse abdominal multi-organ dataset are shown in Table 1, where the items marked in bold represent the optimal values in each column.

Table 1. The Comparative Structure of the 2E-Net Method and Previous State-of-the-Art Methods on the Synapse Dataset.

Methods	$DSC \uparrow$	Aorta	Gallbaladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
TransU-Net	77.48	87.23	63.16	81.87	77.02	94.08	55.86	85.08	75.62
Swin U-Net	79.13	85.47	**69.53**	83.95	79.78	93.95	61.02	90.66	76.60
MISSFormer	81.96	86.99	68.65	**85.21**	82.00	94.41	65.67	**91.92**	80.81
2E-Net	**82.13**	**87.88**	68.14	84.89	**82.95**	**94.69**	**65.69**	91.62	**81.14**

The Transformer encoder is a deep learning network model built on the attention mechanism. While the attention mechanism excels in global feature extraction and fusion, its performance is not as effective for the various scales of multiple abdominal organs. The ResNet50 encoder, on the other hand, is a deep learning network model based on CNN modules. The size of the convolutional kernels and the dimensions of the input feature maps determine the receptive field of the CNN module. Although the convolutional kernel sizes for feature extraction in ResNet50 are all 3×3 ,, the input feature map size decreases progressively at each STAGE. Therefore, ResNet50 can obtain features with varying receptive fields at different stages.

The encoder of 2E-Net utilizes parallel encoding with modules based on both Transformer and CNN. After feature fusion through the Encoder Fusion Module, 2E-Net effectively balances global and local features. Experimental results indicate that incorporating CNN-based modules better captures multi-scale information, achieving an improved segmentation accuracy of 82.13%($DSC \uparrow$).

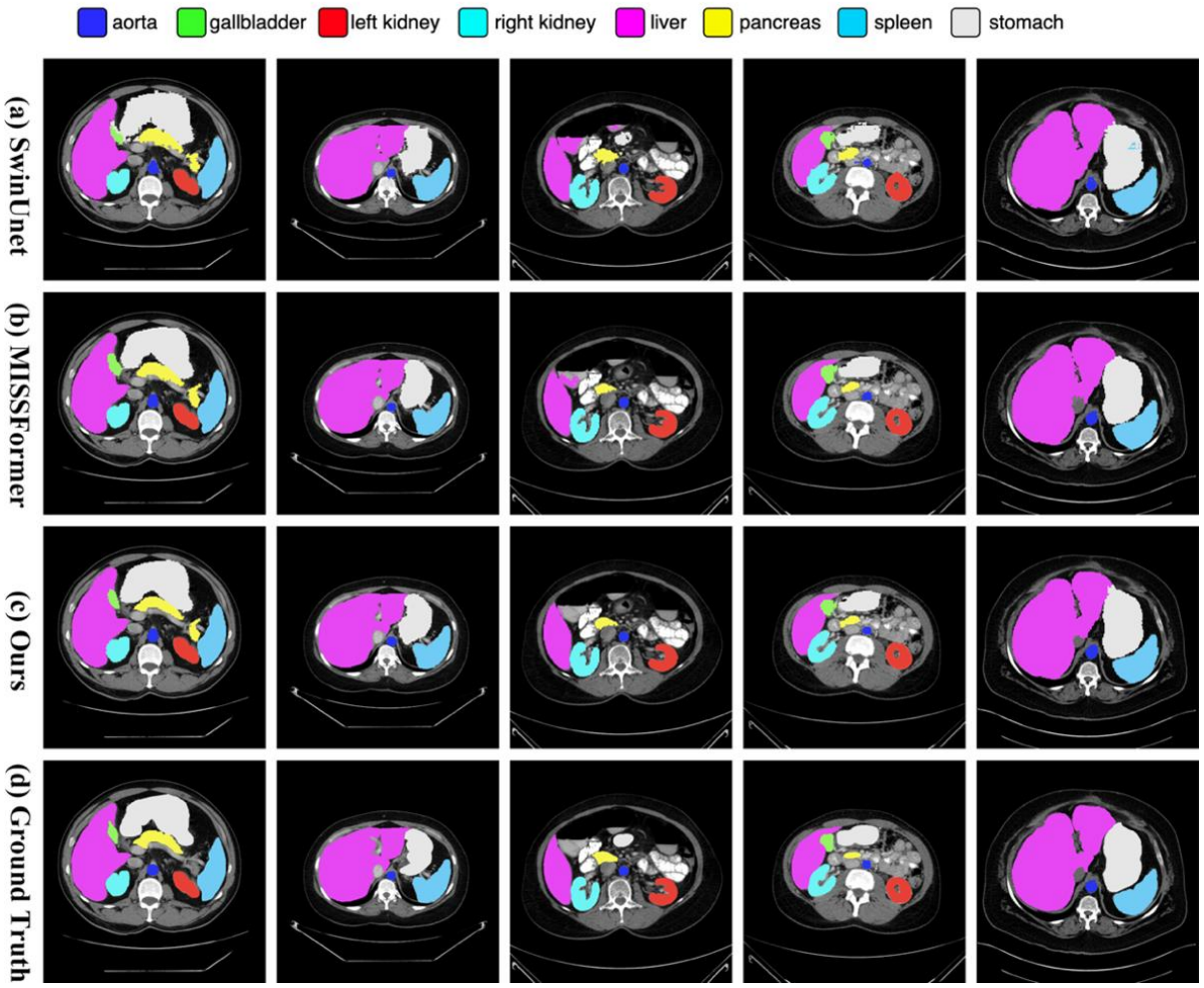

Figure 3. The Ground Truth Segmentation Map and the Model Segmentation Map

In order to clearly demonstrate the effectiveness of the proposed method, this paper visualizes the segmentation results of the model on several test cases. As shown in Figure 3, which includes four groups of images. The first three groups display the segmentation results of (a) SwinUNet, (b) MISSFormer, and (c) our proposed method (Ours) on different test cases. The last group (d) Ground Truth shows the actual segmentation images of these test cases. It can be observed that the proposed method achieves clearer segmentation of organ contours and better preservation of organ integrity compared to the other methods in abdominal multi-organ segmentation.

5. Conclusion

To enable the model to simultaneously leverage local and global features, we proposed the Encoder Fusion Module, which integrates ResNet50 into the MISSFormer framework. Extensive experiments demonstrated that combining ResNet50 with Transformer-based models significantly enhances the model's feature extraction capabilities, resulting in improved segmentation accuracy as evidenced by the Dice Similarity Coefficient (DSC) evaluation metric. This integration allows for better representation of both fine-grained details and broader contextual information, ultimately leading to more accurate segmentation results in medical imaging tasks.

References

[1] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018: pp. 3–11.

[2] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Barcelona, Spain, 2020: pp. 1055–1059. https://doi.org/10.1109/ICASSP40776.2020.9053405.

[3] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, Attention u-net: Learning where to look for the pancreas, arXiv Preprint arXiv:1804.03999 (2018).

[4] C. Li, Y. Tan, W. Chen, X. Luo, Y. Gao, X. Jia, Z. Wang, Attention Unet++: A Nested Attention-Aware U-Net for Liver CT Image Segmentation, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, Abu Dhabi, United Arab Emirates, 2020: pp. 345–349. https://doi.org/10.1109/ICIP40778.2020.9190761.

[5] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, CE-Net: Context Encoder Network for 2D Medical Image Segmentation, IEEE Trans. Med. Imaging 38 (2019) 2281–2292. https://doi.org/10.1109/TMI.2019.2903562.

[6] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context Pyramid Fusion Network for Medical Image Segmentation, IEEE Trans. Med. Imaging 39 (2020) 3008–3018. https://doi.org/10.1109/TMI.2020.2983721.

[7] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015: pp. 234–241.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, \Lukasz Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017). https://proceedings.neurips.cc/paper/7181-attention-is-all (accessed November 13, 2023).

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv Preprint arXiv:2010.11929 (2020).

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, 2021: pp. 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986.

[11] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, arXiv Preprint arXiv:2102.04306 (2021).

[12] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016: pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[13] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, in: European Conference on Computer Vision, Springer, 2022: pp. 205–218.

[14] X. Huang, Z. Deng, D. Li, X. Yuan, Missformer: An effective medical image segmentation transformer, arXiv Preprint arXiv:2109.07162 (2021).

