Channel-Interwined Attention Block in Transformer for Medical Image Segmentation*

Jaejung Lee, Jongbin Ryu

Ajou University, Suwon, Korea

Abstract. In this paper, we introduce the group self-attention based vision transformer architecture for the purpose of medical image segmentation tasks. We tackle the existing transformer blocks in that they only encode the spatial relation in a self-attention operation, whereas neural networks need to learn the channel relation also. This is critical for the medical image segmentation task. It's because the segmentation boundary of the medical images is not clear compared to the real-world images, so deep features must contain between-channel information in order for self-attention to take into account both pixel and channel relations. In addition, in order to utilize deep learning in the field of medical image segmentation, the accuracy of the diagnosis is more important than the speed of the diagnosis. To achieve that, even with a little consideration of the computing cost, the performance should be improved by allowing the model to learn both global and local information better in the image. Therefore, we propose Channel-Interwined Attention block(CIA block), in which the channel is divided into groups through the pixel-shuffle and pixel-unshuffle functions that reshape the feature map, and each of the created three different feature maps is processed by self-attention. This proposed group self-attention in CIA block improves the performance of the medical image segmentation task, such as multi-organ segmentation, and cardiac segmentation. Further, this CIA block is a generalized method that can be applied to any current transformer-based network to improve the medical image segmentation task. We expect this method to solve various segmentation problems in other areas, such as 3D medical imaging and real-world imaging. Codes are available at https://github.com/jaejungscene/Medical-Image-Segmentation.

Keywords: Transformer · Medical Image Segmentation.

1 Introduction

Image segmentation task has advanced quickly for recognizing the medical data. In the early days of medical image segmentation studies, fully convolution network [2] employing convolution neural network performed well due to its convolution-like FC layers. Subsequently, U-Net based models [4, 8, 9, 10] comprised of encoders and decoders architecture have emerged. Although these U-Net-based

* Supported by organization x.

models have produced good results, there was a limitation to focusing on local features while ignoring global information via convolution operation. There are many informative global features in medical images where some organs are located in a whole region of the image. In this situation, if global information is not available, the model's ability to learn the segmentation task is severely limited.

The vision transformer (ViT) model[1], on the other hand, successfully learns global information through its self-attention (SA) operation. ViT separates images into patches and learns global patch-level information by the SA operation. In this manner, ViT efficiently learns global information, but subsequent studies further enhance the ViT model by combining the local features. This approach that takes into account both the local and global level spatial features includes: 1) a hybrid model that employs both convolution and self-attention [5, 6, 7], 2) swin-patch based ViT models[11][12], and ResNet-style 4-stage ViT models[13][14][15].

However, despite the development of learning spatial information, studies on learning channel information in the ViT model have not been performed. The attention operation of ViT doesn't learn pixel relation of different channels, only learning pixel relation of the same channel. This means attention is a channel-wise operation and does not perform a point-wise operation. This is related to the group convolution operations of CNNs that diversify the channel information flow. In contrast, group self-attention has not yet been employed since it is not feasible to implement it because changing the shape of the feature map can increase computational complexity. So, in this paper, we introduce a new method, which is like group self-attention, for learning both local and global features using only one transformer block of ViT using some functions.

Previous ViT based models perform self-attention by a single feature map in the encoder block. Self-attention is generally performed using spatial dimension information. The spatial dimension's patches capture global relationships through self-attention. Feature maps contain not only spatial dimension information but also channel dimension information. Channel dimension tokens contain abstract representations of the entire image. So channel dimension can learn a variety of information from spatial dimension information when calculating attention scores between channels. [3] Furthermore, if blocks that have different sizes can be used together in training, we can obtain more information from a single block.

We propose a Channel-Interwined Attention (CIA) block that integrates the spatial dimension and channel dimension information of ViT. CIA block has three different sub-blocks. Each of these sub-blocks makes the training model get three different local and global information. First, one block, called the original block, just uses the feature map of the original ViT model. Second another block reduces the channel dimension and moves it to the pixel level of the feature map, creating a larger resolution block. Therefore, The model can get more broad features than the original block due to the input of this block. Lastly, another block reduces resolution and increases the channel dimension, allowing the model

to focus on more local features. We implement adjusting the feature map using shuffle and unshuffle functions in the super-resolution field. And, three different feature maps that consist of two adjusted feature maps and one original feature map, go to each specific block as the input. At all three blocks, each feature map is processed by multi head self attention of the traditional ViT model in the same way. After that, the feature maps from the three encoder blocks are readjusted to the original feature map size and we reduce computation cost by concatenating three feature maps and cutting down this channel size through MLP Layer.

We conduct experiments using Synapse and ACDC datasets, which are commonly used for medical image datasets. We compare the performance between the two models, with models known to perform well in medical image segmentation, such as TransUNet, SwinUNet, TransCASCADE, and PVTCASCADE, and CIA blocks attached to these models. Empirical results demonstrate that CIA block improves model learning both of global and local features and makes the model get better segmentation performance.

2 Method

At first, we introduce ViT-based models that we use in our experiment and propose a Channel-Interwined Attention(CIA) block which replaces the encoder block in the existing ViT model. We explain how the CIA block can replace the encoder block for each model.

2.1 ViT models that we use in our experiments

To obtain a variety of and also hierarchical information, we use Swin Transformer [11], TransUNet [6], PVT-v2 [6], which are different architectures and use different methods. However, all ViT model has the same overall shape, like U-shape of U-Net model which have good performance in medical image segmentation by optimizing the model to get global and local information. Therefore, based on these three ViT U-shape models, We compare the performance before and after attaching CIA block that substitutes the encoder block of each of these three models.

2.2 CIA block

The existing ViT model performs one self-attention on a feature map with a single resolution in one transformer block. The proposed model modifies the block for performing three self-attention on three different resolutions of feature maps in one transformer block. Among the three sub-blocks in a transformer block, the block (block-b) whose dimension is 784 is the same as the original transformer block of the pure ViT model. Other two new blocks (block-s, block-l) are added for CIA block.

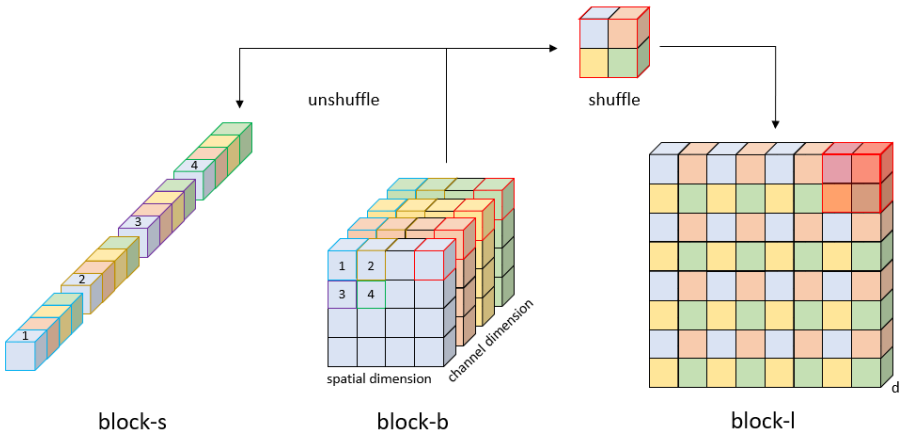

Fig. 1. Basis mechanism of CIA block. Channel dimension information of the feature map of block-l is moved to the spatial dimension through pixel shuffle function. Spatial dimension information of the feature map of block-s is moved to the channel dimension through pixel unshuffle.

As shown in Figure 1, the size of the block feature map is determined by how much information in the channel dimension area we will use for the spatial dimension area of the block or opposite. To do this, there is a scale factor that can manage this scale in both of pixel shuffle and unshuffle function. In the shuffle function, the width and height of the feature map increase by multiplying the size by this scale factor, and the channel dimension decreases by the size divided by this scale factor squared. In the Unshuffle function, this method occurs in an inverted manner. the method equation of these two functions is under Eq.(1),(2). We set this scale factor F as 2.

$$\mathbf{Out} = \mathit{Shuffle}(\mathbf{In},\, F)\,, \quad \mathbf{In} \in R^{H\times W\times D}\ \ \mathbf{Out} \in R^{FH\times FW\times \frac{1}{F^2}D} \tag{1}$$

$$\mathbf{Out} = \mathit{Unshuffle}(\mathbf{In},\, F)\,, \quad \mathbf{In} \in R^{H\times W\times D}\ \ \mathbf{Out} \in R^{\frac{1}{F}H\times \frac{1}{F}W\times\ F^2D} \tag{2}$$

Two shuffled and unshuffled feature maps and one original feature map enter the input of each of their Transformer Encoders(TE), which consists of Multihead Self-Attention(MSA) and Multi-Layer Perceptron(MLP), Layer Normalization(LN) like Eq.(3)(4)(5).

$$\mathbf{Out}^{'} = MSA(LN(\mathbf{In})) + \mathbf{In}, \quad \mathbf{In}, \mathbf{Out}^{'} \in R^{\frac{HW}{P^2}\times D} \tag{3}$$

$$\mathbf{Out} = MLP(LN(\mathbf{Out}^{'})) + \mathbf{Out}^{'}, \quad \mathbf{Out}^{'}, \mathbf{Out} \in R^{\frac{HW}{P^2}\times D} \tag{4}$$

$$\mathbf{Out} = TE(\mathbf{In}), \quad \mathbf{In}, \mathbf{Out} \in R^{\frac{HW}{P^2}\times D} \tag{5}$$

Then two changed feature maps are restored to the original shape through shuffle and unshuffle function as Figure 1. After that, all three feature maps

are connected, and they enter the FC-Layer made of MLP and LN to reduce the channel dimensions for use as input for the next CIA block. Therefore, the output of the l-th layer can be written as follows:

$$\mathbf{Z}_{l-1}^{1} = Unshuffle(TE(Shuffle(\mathbf{Z}_{l-1}))), \tag{6}$$

$$\mathbf{Z}_{l-1}^{2} = TE(\mathbf{Z}_{l-1}), \tag{7}$$

$$\mathbf{Z}_{l-1}^{3} = Shuffle(TE(Unshuffle(\mathbf{Z}_{l-1}))), \tag{8}$$

$$\mathbf{Z}_{l}^{'} = [\mathbf{Z}_{l-1}^{1};\ \mathbf{Z}_{l-1}^{2};\ \mathbf{Z}_{l-1}^{3}], \tag{9}$$

$$\mathbf{Z}_{l} = MLP(LN(\mathbf{Z}_{l}^{'})), \tag{10}$$

$$\mathbf{Z}_{l} = CIA(\mathbf{Z}_{l-1}) \tag{11}$$

where $\mathbf{Z} \in R^{\frac{HW}{P^2} \times D}$ and $\mathbf{Z}_{l}^{'} \in R^{\frac{HW}{P^2} \times 3D}$. The structure of CIA block is illustrated in Figure 2.

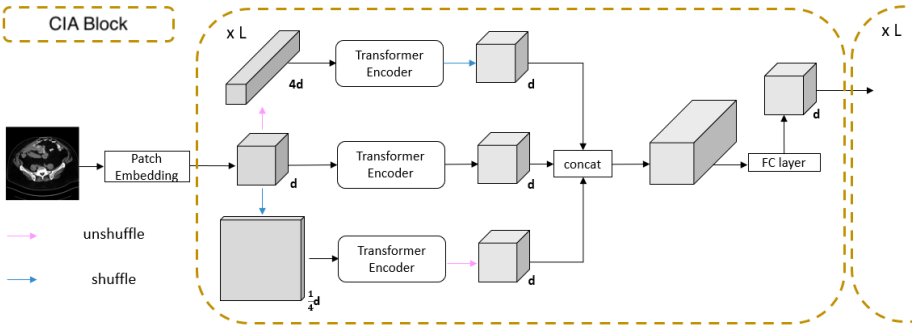

Fig. 2. Overall and CIA block architecture when the original Transformer Layer of ViT that consists of MSA and MLP, LN, is totally replaced by CIA block.

2.3 Applying the CIA block to ViT-based Segmentation Model

We replace the transformer encoder block of TransUnet, TransCASCADE, and PVT-CASCADE with the CIA block. Therefore, in TransUnet and TransCASCADE, the transformer encoder of the original ViT is replaced by the CIA block, and in the case of PVT-CASCADE, the transformer encoder with spatial reduction attachment (SRA) is replaced by the CIA block. Additionally, In PVT-CASCADE, the feature map size continues to decrease as the feature goes deeper into the network, so even when replacing it with a CIA block, we make the CIA block receive this reduced feature map size as input.

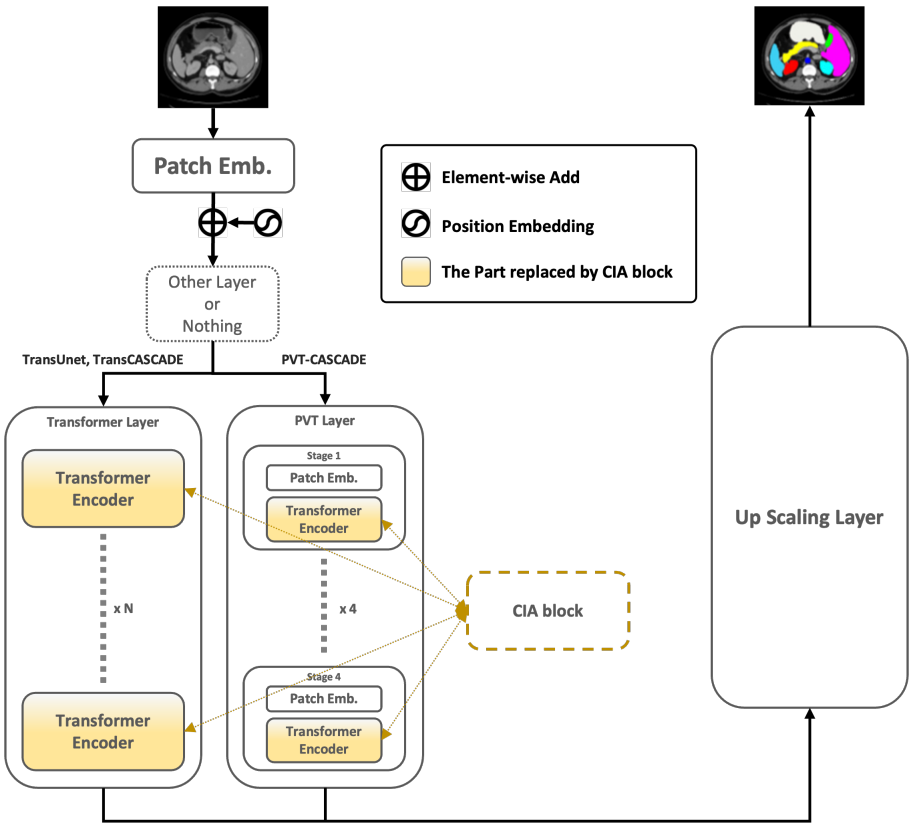

Fig. 3. Entire Architecture when replacing the Transformer Encoder with our CIA block. When using TransUnet and TransCASCADE, Transformer Encoder of Transformer Layer is replaced. When using PVT-CASCADE, Transformer Encoder of PVT Layer is replaced.

3 Experiments

We verify the superiority of our proposed CIA block by comparing the performance between three Vision Transformer models with and without CIA blocks. Also, to verify the effectiveness of our CIA block on the medical image field, We conduct experiments on two major medical image datasets. Ans we compare our feature map's visualized images so that verify our block can see multiple parts.

3.1 Datasets

We use two datasets that have been commonly used in medical image segmentation field and related major studies [5, 6, 7]. The first dataset is **Synapse multi-organ dataset**. The Synapse dataset is a multi-organ image dataset that consists of 30 abdominal CT sans with 3779 abdominal CT images. We divide this into 18 cases(2212 axial slices) for training and 12 cases for testing. We evaluate the performance through the average DICE score of all segmentation results of a total of 8 abdominal organs. They include the aorta(AT), gallbladder (GB), left kidney (KL), right kidney (KR), liver(LV), pancreas (PC), spleen (SP), and stomach (SM).

The second dataset is **ACDC dataset**, which consists of 100 cardiac MRI images with a slice thickness of 5 to 8 mm. The short-axis in-plane spatial resolution goes from 0.83 to 1.75 mm2/pixel. We divide 70 cases(1930 axial slices) for training, 10 cases for validation, and the last 20 cases for testing. The evaluation metric is also the average DICE score of all segmentation results of three organs: the right ventricle(RV), left ventricle(LV), and myocardium(Myo).

3.2 Implement details

We replace the Encoder block of existing models with the CIA block while keeping the entire structure. in the cases of TransUnet and TransCASCADE, we implemented all CIA blocks to receive $\frac{224}{16^2}$ x $\frac{224}{16^2}$ features as input. So, the image size is 224 x 224, and the patch size is 16. in the cases of PVT-CASCADE, we implemented inputs of every CIA block is different due to pyramid structure of PVT-CASCADE. first CIA block receives $\frac{H}{2^2}$ x $\frac{W}{2^2}$ and second $\frac{H}{2^3}$ x $\frac{W}{2^3}$, third $\frac{H}{2^4}$ x $\frac{W}{2^4}$, fourth $\frac{H}{2^5}$ x $\frac{W}{2^5}$, in which H and W is first input image. For a consistent comparison, we didn't use pretrained weight for all architecture, which is TransUnet, TransCASCADE, and PVT-CASCADE. And the hyperparameter values of deep learning, such as scheduler, learning rate, optimizer, etc., were the same as those from each model paper [5, 6]. All experiments were conducted using the Nvidia RTX3090 GPU.

3.3 Experimental Results

As shown in Table 1, With an increase of average dice score(DSC) of about 1-2% points, the model with CIA block is performing better than the baseline model

Table 1. Results of different architectures on Synapse multi-organ segmentation. DICE is the average dice score of all organs (average Dice score %)

Architectures	CIA	DICE	AT	GB	KL	KR	LV	PC	SP	SM
R50+UNet*	X	74.68	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50+AttnUNet*	X	75.57	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
SwinUNet*	X	77.58	81.76	65.95	82.32	79.22	93.73	53.81	88.04	75.79
TransUnet	X	77.39	86.49	64.92	83.70	79.87	92.54	58.15	81.42	72.06
PVT-CASCADE	X	71.38	62.31	60.32	79.81	71.67	92.61	49.69	83.98	70.67
TransCASCADE	X	77.78	83.47	68.56	84.69	78.83	93.62	56.30	85.77	71.01
TransUnet	O	79.10	86.98	67.32	84.81	80.67	93.61	59.69	86.98	72.67
PVT-CASCADE	O	71.52	62.91	60.35	79.81	71.69	92.58	50.08	83.99	70.77
TransCASCADE	O	79.24	84.44	70.71	85.01	80.52	94.41	56.20	87.74	74.77

Table 2. Comparsion results of models with and without CIA block on Synapse multi-organ dataset.

Architectures	CIA	DICE[%]	Params[M]	Flops[G]	Throughput[MB/s]
TransUnet	X	77.39	105	49.32	35.418
PVT-CASCADE	X	71.38	35	12.56	30.106
TransCASCADE	X	77.78	123	80.52	25.088
TransUnet	O	**79.10**	123	35.56	20.411
PVT-CASCADE	O	**71.52**	36	12.48	24.084
TransCASCADE	O	**79.24**	131	59.54	18.246

Table 3. Segmentation performance of different architectures on ACDC dataset. DICE is the average dice score of all organs (average Dice score %)

Architectures	CIA	DICE	RV	Myo	LV
R50+UNet*	X	87.55	87.10	80.63	94.92
R50+AttnUNet*	X	86.75	87.58	79.20	93.48
SwinUNet*	X	88.07	85.77	84.42	94.03
TransUnet	X	89.00	86.01	86.22	94.78
PVT-CASCADE	X	87.52	85.91	83.12	93.54
TransCASCADE	X	89.05	86.48	86.01	94.67
TransUnet	O	89.32	87.09	86.11	94.77
PVT-CASCADE	O	88.07	86.01	84.29	93.92
TransCASCADE	O	90.90	89.24	87.93	95.53

Table 4. Comparsion results of models with and without CIA block on ACDC dataset.

Architectures	CIA	DICE[%]	Params[M]	Flops[G]	Throughput[MB/s]
TransUnet	X	89.00	105	49.24	18.816
PVT-CASCADE	X	87.52	35	12.50	17.709
TransCASCADE	X	89.05	123	80.52	14.336
TransUnet	O	**89.32**	123	35.48	10.752
PVT-CASCADE	O	**88.07**	36	12.48	14.686
TransCASCADE	O	**90.90**	131	59.50	8.855

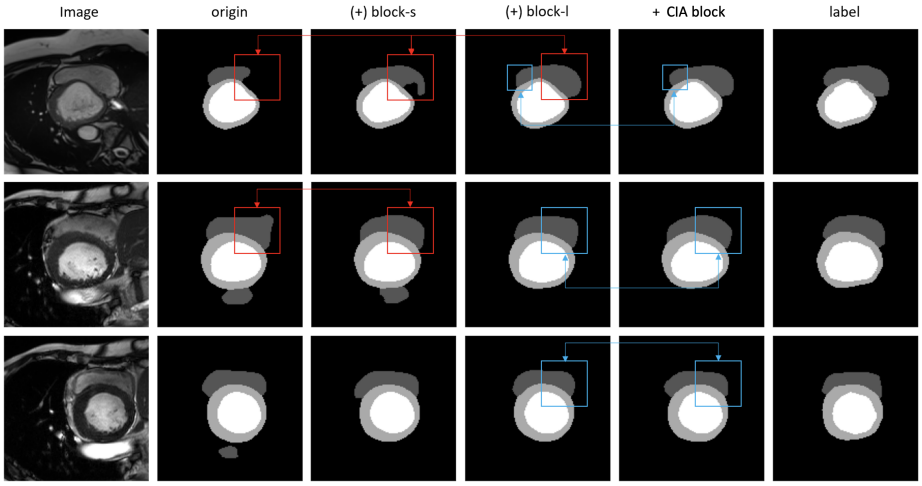

Fig. 4. Comparsion between origin and adding only one block-s or block-l and CIA block through a label.

on Synpase multi-organ dataset. That isn't a critical problem that the CIA block attached model parameter is larger, because the speed problem of inference can be offset by lower flops and throughput, even only to need slightly more memory.

This tendency of results is also followed by ACDC dataset, as shown in Table 2. However, an increase in average dice score(DSC) is lower than Synapse multi-organ dataset.

3.4 Ablation Study

The Number of Ways We examine the impact of each sub-block of CIA block that has different resolutions due to different feature map sizes. Therefore, we compare the differences in performance when using only block-s and block-b of CIA block, block-l and block-b of CIA block, and all three blocks together.

Certainly, the performance is the highest when all three sub-blocks are used. In line with our expectations, block-l for learning local features and block-s for learning global features allow the model to learn global and local features to increase performance by obtaining more detailed and diverse information on medical images. On the other hand, when one of block-l or block-s is excluded, the performance varies depending on the dataset, so if only one sub-block is used, the model learns only one of the global or local biased information and does not seem to increase performance compared to using both. Nevertheless, when only one of the sub-blocks is attached, there is a slight improvement in performance compared to the model not using any block, which indicates it's worth using any sub-block for learning. In addition, when using block-s, there is a problem that the number of parameters increases rapidly due to increased channel dimension, so it seems desirable to use block-l when only one of the two sub-blocks should be used.

Table 5. Result of TransCASCADE on Synapse multi-organ dataset using different sub-block of CIA block

block-b	block-l	block-s	DICE[%]	Params[M]	Flops[G]	Throughput[MB/s]
∨	∨	No	77.30	45	51.10	21.504
∨	No	∨	77.45	126	53.64	24.576
∨	∨	∨	**79.24**	131	59.54	18.246

4 Conclusion

A Transformer encoder block of existing ViT models can better learn global information than CNNs models. To combine these advantages of Transformer and CNNs, and to simplify the model, We make the CIA block that captures not only more global information than the previous transformer encoder block,

Table 6. Result of TransCASCADE on ACDC dataset using different sub-block of CIA block

block-b	block-l	block-s	DICE[%]	Params[M]	Flops[G]	Throughput[MB/s]
∨	∨	No	89.86	45	51.10	10.381
∨	No	∨	89.60	126	53.62	14.003
∨	∨	∨	**90.90**	131	59.50	8.855

but also more local information through learning channel information. Our CIA block attached model achieves state-of-the-art performance on the medical image segmentation field, and can easily replace the original encoder block. So we expect this CIA block to spread and be applied more to other models and more to studies such as 3D medical imaging areas.

Bibliography

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[2] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[3] Ding, Mingyu, et al. "Davit: Dual attention vision transformers." Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV. Cham: Springer Nature Switzerland, 2022.

[4] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

[5] 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," CoRR, vol. abs/2102.04306, 2021.

[6] Rahman, Md Mostafijur, and Radu Marculescu. "Medical Image Segmentation via Cascaded Attention Decoding." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

[7] Cao, Hu, et al. "Swin-unet: Unet-like pure transformer for medical image segmentation." arXiv preprint arXiv:2105.05537 (2021).

[8] Zhou, Zongwei, et al. "Unet++: A nested u-net architecture for medical image segmentation." 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 International Publishing, 2018.

[9] Huang, Huimin, et al. "Unet 3+: A full-scale connected unet for medical image segmentation." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[10] Oktay, Ozan, et al. "Attention u-net: Learning where to look for the pancreas." arXiv preprint arXiv:1804.03999 (2018).

[11] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[12] Liu, Ze, et al. "Swin transformer v2: Scaling up capacity and resolution." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[13] Wang, Wenhai, et al. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[14] Wang, Wenhai, et al. "Pvt v2: Improved baselines with pyramid vision transformer." Computational Visual Media 8.3 (2022): 415-424.
[15] Yuan, Li, et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
[16] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).
[17] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[18] He, Kaiming, et al. "Identity mappings in deep residual networks." Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016.
[19] Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
[20] Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
[21] LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017