

# CiT-Net: Convolutional Neural Networks Hand in Hand with Vision Transformers for Medical Image Segmentation

Tao Lei<sup>1,2</sup>, Rui Sun<sup>1</sup>, Xuan Wang<sup>3</sup>, Yingbo Wang<sup>1</sup>, Xi He<sup>1</sup>, Asoke Nandi<sup>4</sup>

<sup>1</sup>Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology

<sup>2</sup>Department of Geriatric Surgery, First Affiliated Hospital, Xi'an Jiaotong University

<sup>3</sup>Unmanned System Research Institute, Northwestern Polytechnical University

<sup>4</sup>Department of Electronic and Electrical Engineering, Brunel University London

leitao@sust.edu.cn, siri0920@163.com, wangxuan@nwpu.edu.cn, {wangyingbo, xihe}@sust.edu.cn,  
Asoke.Nandi@brunel.ac.uk

## Abstract

The hybrid architecture of convolutional neural networks (CNNs) and Transformer are very popular for medical image segmentation. However, it suffers from two challenges. First, although a CNNs branch can capture the local image features using vanilla convolution, it cannot achieve adaptive feature learning. Second, although a Transformer branch can capture the global features, it ignores the channel and cross-dimensional self-attention, resulting in a low segmentation accuracy on complex-content images. To address these challenges, we propose a novel hybrid architecture of convolutional neural networks hand in hand with vision Transformers (CiT-Net) for medical image segmentation. Our network has two advantages. First, we design a dynamic deformable convolution and apply it to the CNNs branch, which overcomes the weak feature extraction ability due to fixed-size convolution kernels and the stiff design of sharing kernel parameters among different inputs. Second, we design a shifted-window adaptive complementary attention module and a compact convolutional projection. We apply them to the Transformer branch to learn the cross-dimensional long-term dependency for medical images. Experimental results show that our CiT-Net provides better medical image segmentation results than popular SOTA methods. Besides, our CiT-Net requires lower parameters and less computational costs and does not rely on pre-training. The code is publicly available at <https://github.com/SR0920/CiT-Net>.

## 1 Introduction

Medical image segmentation refers to dividing a medical image into several specific regions with unique properties. Medical image segmentation results can not only achieve abnormal detection of human body regions but also be used to guide clinicians. Therefore, accurate medical image segmentation has become a key component of computer-aided diagnosis and treatment, patient condition analysis, image-guided

surgery, tissue and organ reconstruction, and treatment planning. Compared with common RGB images, medical images usually suffer from the problems such as high density noise, low contrast, and blurred edges. So how to quickly and accurately segment specific human organs and lesions from medical images has always been a huge challenge in the field of smart medicine.

In recent years, with the rapid development of computer hardware resources, researchers have continuously developed many new automatic medical image segmentation algorithms based on a large number of experiments. The existing medical image segmentation algorithms can be divided into two categories: based on convolutional neural networks (CNNs) and based on the Transformer networks.

The early traditional medical image segmentation algorithms are based on manual features designed by medical experts using professional knowledge [Suetens, 2017]. These methods have a strong mathematical basis and theoretical support, but these algorithms have poor generalization for different organs or lesions of the human body. Later, inspired by the full convolutional networks (FCN) [Long *et al.*, 2015] and the encoder-decoder, Ronneberger *et al.* designed the U-Net [Ronneberger *et al.*, 2015] network that was first applied to medical image segmentation. After the network was proposed, its symmetric U-shaped encoder and decoder structure received widespread attention. At the same time, due to the small number of parameters and the good segmentation effect of the U-Net network, deep learning has made a breakthrough in medical image segmentation. Then a series of improved medical image segmentation networks are inspired based on the U-Net network, such as 2D U-Net++ [Zhou *et al.*, 2018], ResDO-UNet [Liu *et al.*, 2023], SGU-Net [Lei *et al.*, March 2023], 2.5D RIU-Net [Lv *et al.*, 2022], 3D Unet [Çiçek *et al.*, 2016], V-Net [Milletari *et al.*, 2016], etc. Among them, Alom *et al.* designed R2U-Net [Alom *et al.*, 2018] by combining U-Net, ResNet [Song *et al.*, 2020], and recurrent neural network (RCNN) [Girshick *et al.*, 2014]. Then Gu *et al.* introduced dynamic convolution [Chen *et al.*, 2020] into U-Net proposed CA-Net [Gu *et al.*, 2020]. Based on U-Net, Yang *et al.* proposed DCU-Net [Yang *et al.*, 2022] by referring to the idea of residual connection and deformable convolution [Dai *et al.*, 2017]. Lei *et al.* [Lei *et al.*, 2022] proposed a network ASE-Net based on adversarial consistency learning and dynamic

convolution.

The rapid development of CNNs in the field of medical image segmentation is largely due to the scale invariance and inductive bias of convolution operation. Although this fixed receptive field improves the computational efficiency of CNNs, it limits its ability to capture the relationship between distant pixels in medical images and lacks the ability to model medical images in a long range.

Aiming at the shortcomings of CNNs in obtaining global features of medical images, scholars have proposed a Transformer architecture. In 2017, Vaswani et al. [Vaswani *et al.*, 2017] proposed the first Transformer network. Because of its unique structure, Transformer obtains the ability to process indefinite-length input, establish long-range dependency modeling, and capture global information. With the excellent performance of Transformer in NLP fields, ViT [Dosovitskiy *et al.*, 2020] applied Transformer to the field of image processing for the first time. Then Chen et al. put forward TransUNet [Chen *et al.*, 2021], which initiates a new period of Transformer in the field of medical image segmentation. Valanarasu et al. proposed MedT [Valanarasu *et al.*, 2021] in combination with the gating mechanism. Cao et al. proposed a pure Transformer network Swin-Unet [Cao *et al.*, 2021] for medical image segmentation, in combination with the shifted-window multi-head self-attention (SW-MSA) in Swin Transformer [Liu *et al.*, 2021b]. Subsequently, Wang et al. designed the BAT [Wang *et al.*, 2021a] network for dermoscopic images segmentation by combining the edge detection idea [Sun *et al.*, 2022]. Hatamizadeh et al. proposed Swin UNETR [Tang *et al.*, 2022] network for 3D brain tumor segmentation. Wang et al. proposed the UCTransNet [Wang *et al.*, 2022] network that combines the channel attention with Transformer.

These methods can be roughly divided into based on the pure Transformer architecture and based on the hybrid architecture of CNNs and Transformer. The pure Transformer network realizes the long-range dependency modeling based on self-attention. However, due to the lack of inductive bias of the Transformer itself, Transformer cannot be widely used in small-scale datasets like medical images [Shamshad *et al.*, 2022]. At the same time, Transformer architecture is prone to ignore detailed local features, which reduces the separability between the background and the foreground of small lesions or objects with large-scale changes in the medical image.

The hybrid architecture of CNNs and Transformer realizes the local and global information modeling of medical images by taking advantage of the complementary advantages of CNNs and Transformer, thus achieving a better medical image segmentation effect [Azad *et al.*, 2022]. However, this hybrid architecture still suffers from the following two problems. First, it ignores the problems of organ deformation and lesion irregularities when modeling local features, resulting in weak local feature expression. Second, it ignores the correlation between the feature map space and the channels when modeling the global feature, resulting in inadequate expression of self-attention. To address the above problems, our main contributions are as follows:

- A novel dynamic deformable convolution (DDConv) is

proposed. Through task adaptive learning, DDConv can flexibly change the weight coefficient and deformation offset of convolution itself. DDConv can overcome the problems of fixation of receptive fields and sharing of convolution kernel parameters, which are common problems of vanilla convolution and its variant convolutions, such as Atrous convolution and Involution, etc. Improves the ability to perceive tiny lesions and targets with large-scale changes in medical images.

- A new (shifted)-window adaptive complementary attention module ((S)W-ACAM) is proposed. (S)W-ACAM realizes the cross-dimensional global modeling of medical images through four parallel branches of weight coefficient adaptive learning. Compared with the current popular attention mechanisms, such as CBAM and Non-Local, (S)W-ACAM fully makes up for the deficiency of the conventional attention mechanism in modeling the cross-dimensional relationship between spatial and channels. It can capture the cross-dimensional long-distance correlation features in medical images, and enhance the separability between the segmented object and the background in medical images.
- A new parallel network structure based on dynamically adaptive CNNs and cross-dimensional feature fusion Transformer is proposed for medical image segmentation, called CiT-Net. Compared with the current popular hybrid architecture of CNNs and Transformer, CiT-Net can maximize the retention of local and global features in medical images. It is worth noting that CiT-Net not only abandons pre-training but also has fewer parameters and less computational costs, which are 11.58 M and 4.53 GFLOPs respectively.

Compared with the previous vanilla convolution [Ronneberger *et al.*, 2015], dynamic convolution [Chen *et al.*, 2020] [Li *et al.*, 2021], and deformable convolution [Dai *et al.*, 2017], our DDConv can not only adaptively change the weight coefficient and deformation offset of the convolution according to the medical image task, but also better adapt to the shape of organs and small lesions with large-scale changes in the medical image, and additionally, it can improve the local feature expression ability of the segmentation network. Compared with the self-attention mechanism in the existing Transformer architectures [Cao *et al.*, 2021] [Wang *et al.*, 2021a], our (S)W-ACAM requires fewer parameters and less computational costs while it's capable of capturing the global cross-dimensional long-range dependency in the medical image, and improving the global feature expression ability of the segmentation network. Our CiT-Net does not require a large number of labeled data for pre-training, but it can maximize the retention of local details and global semantic information in medical images. It has achieved the best segmentation performance on both dermoscopic images and liver datasets.

## 2 Method

### 2.1 Overall Architecture

The fusion of local and global features are clearly helpful for improving medical image segmentation. CNNs capture lo-

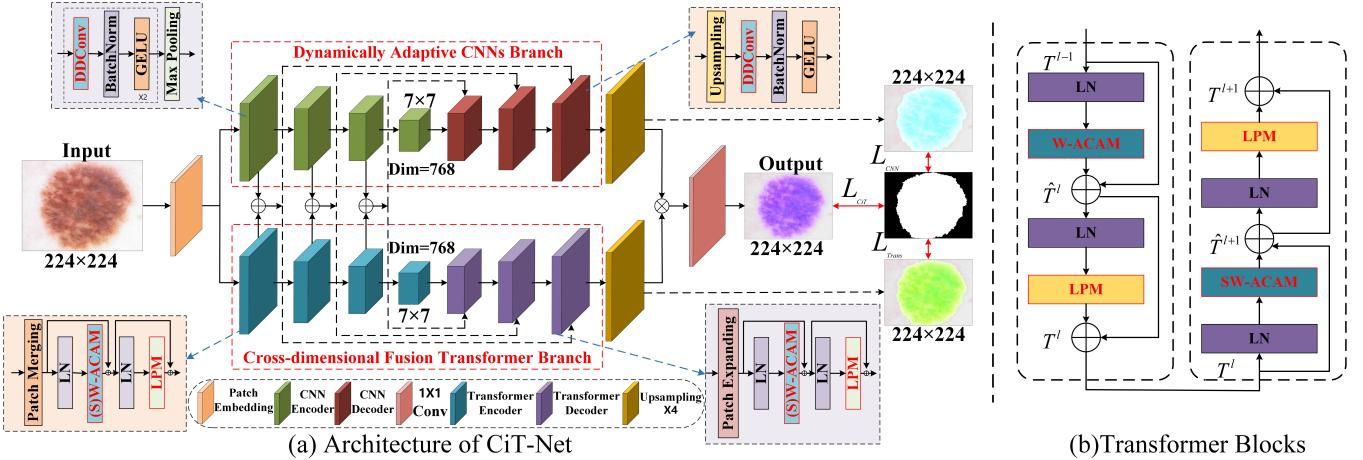


Figure 1: (a) The architecture of CiT-Net. CiT-Net consists of a dual-branch interaction between dynamically adaptive CNNs and cross-dimensional feature fusion Transformer. The DDConv in the CNNs branch can adaptively change the weight coefficient and deformation offset of the convolution itself, which improves the segmentation accuracy of irregular objects in medical images. The (S)W-ACAM in the Transformer branch can capture the cross-dimensional long-range dependency in medical images, improving the separability of segmented objects and backgrounds in medical images. The lightweight perceptron module (LPM) greatly reduces the parameters and calculations of the original Transformer network by using the Ghost strategy. (b) Two successive Transformer blocks. W-ACAM and SW-ACAM are cross-dimensional self-attention modules with shifted windows and compact convolutional projection configurations.

cal features in medical images through convolution operation and hierarchical feature representation. In contrast, the Transformer network realizes the extraction of global features in medical images through the cascaded self-attention mechanism and the matrix operation with context interaction. In order to make full use of local details and global semantic features in medical images, we design a parallel interactive network architecture CiT-Net. The overall architecture of the network is shown in Figure 1 (a). CiT-Net fully considers the complementary properties of CNNs and Transformer. During the forward propagation process, CiT-Net continuously feeds the local details extracted by the CNNs to the decoder of the Transformer branch. Similarly, CiT-Net also feeds the global long-range relationship captured by the Transformer branch to the decoder of the CNNs branch. Obviously, the proposed CiT-Net provides better local and global feature representation than pure CNNs or Transformer networks, and it shows great potential in the field of medical image segmentation.

Specifically, CiT-Net consists of a patch embedding model, dynamically adaptive CNNs branch, cross-dimensional fusion Transformer branch, and feature fusion module. Among them, the dynamically adaptive CNNs branch and the cross-dimensional fusion Transformer branch follow the design of U-Net and Swin-Unet, respectively. The dynamically adaptive CNNs branch consists of seven main stages. By using the weight coefficient and deformation offset adaptive DDConv in each stage, the segmentation network can better understand the local semantic features of medical images, better perceive the subtle changes of human organs or lesions, and improve the ability of extracting multi-scale change targets in medical images. Similarly, the cross-dimensional fusion Transformer branch also consists of seven main stages. By using (S)W-ACAM attention in each stage, as shown in Figure 1 (b), the segmentation network can better understand the

global dependency of medical images to capture the position information between different organs, and improve the separability of the segmented object and the background in the medical images.

Although our CiT-Net can effectively improve the feature representation of medical images, it requires a large number of training data and network parameters due to the dual-branch structure. As the conventional Transformer network contains a lot of MLP layers, which not only aggravates the training burden of the network but also makes the number of model parameters rise sharply, resulting in the slow training of the model. Inspired by the idea of the Ghost network [Han *et al.*, 2020], we redesign the MLP layer in the original Transformer and proposed a lightweight perceptron module (LPM). The LPM can help our CiT-Net not only achieve better medical image segmentation results than MLP but also greatly reduced the parameters and computational complexity of the original Transformer block, even the Transformer can achieve good results without a lot of labeled data training. It is worth mentioning that the dual-branch structure involves mutually symmetric encoders and decoders so that the parallel interaction network structure can maximize the preservation of local features and global features in medical images.

## 2.2 Dynamic Deformable Convolution

Vanilla convolution has spatial invariance and channel specificity, so it has a limited ability to change different visual modalities when dealing with different spatial locations. At the same time, due to the limitations of the receptive field, it is difficult for vanilla convolution to extract features of small targets or targets with blurred edges. Therefore, vanilla convolution inevitably has poor adaptability and weak generalization ability for complex medical images. Although the ex-

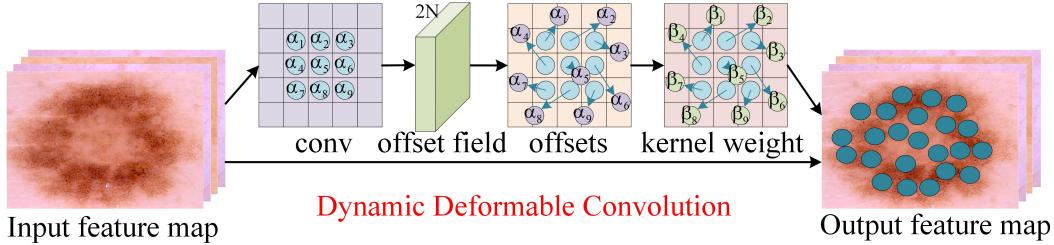


Figure 2: The module of the proposed DDConv. Compared with the current popular convolution strategy, DDConv can dynamically adjust the weight coefficient and deformation offset of the convolution itself during the training process, which is conducive to the feature capture and extraction of irregular targets in medical images.  $\alpha$  and  $\beta$  represent the different weight values of DDConv in different states.

isting deformable convolution [Dai *et al.*, 2017] and dynamic convolution [Chen *et al.*, 2020] [Li *et al.*, 2021] outperforms vanilla convolution to a certain extent, they still have the unsatisfied ability to balance the performance and size of networks when dealing with medical image segmentation.

In order to solve the shortcomings of current convolution operations, this paper proposes a new convolution strategy, DDConv, as shown in Figure 2. It can be seen that DDConv can adaptively learn the kernel deformation offset and weight coefficients according to the specific task and data distribution, so as to realize the change of both the shapes and the values of convolution kernels. It can effectively deal with the problems of large data distribution differences and large target deformation in medical image segmentation. Also, DDConv is plug-and-play and can be embedded in any network structure.

The shape change of the convolutional kernel in DDConv is based on the network learning of the deformation offsets. The segmentation network first samples the input feature map  $X$  using a square convolutional kernel  $S$ , and then performs a weighted sum with a weight matrix  $M$ . The square convolution kernel  $S$  determines the range of the receptive field, e.g., a  $3 \times 3$  convolution kernel can be expressed as:

$$S = \{(0,0), (0,1), (0,2), \dots, (2,1), (2,2)\}, \quad (1)$$

then the output feature map  $Y$  at the coordinate  $\varphi_n$  can be expressed as:

$$Y(\varphi_n) = \sum_{\varphi_m \in S} S(\varphi_m) \cdot X(\varphi_n + \varphi_m), \quad (2)$$

when the deformation offset  $\Delta\varphi_m = \{m = 1, 2, 3, \dots, N\}$  is introduced in the weight matrix  $M$ ,  $N$  is the total length of  $S$ . Thus the Equation (2) can be expressed as:

$$Y(\varphi_n) = \sum_{\varphi_m \in S} S(\varphi_m) \cdot X(\varphi_n + \varphi_m + \Delta\varphi_m). \quad (3)$$

Through network learning, an offset matrix with the same size as the input feature map can be finally obtained, and the matrix dimension is twice that of the input feature map.

To show the convolution kernel of DDConv is dynamic, we first present the output feature map of vanilla convolution:

$$y = \sigma(W \cdot x), \quad (4)$$

where  $\sigma$  is the activation function,  $W$  is the convolutional kernel weight matrix and  $y$  is the output feature map. In contrast, the output of the feature map of DDConv is:

$$\hat{y} = \sigma((\alpha_1 \cdot W_1 + \dots + \alpha_n \cdot W_n) \cdot x), \quad (5)$$

where  $n$  is the number of weight coefficients,  $\alpha_n$  is the weight coefficients with learnable parameters and  $\hat{y}$  is the output feature map generated by DDConv. DDConv achieves dynamic adjustment of the convolution kernel weights by linearly combining different weight matrices according to the corresponding weight coefficients before performing the convolution operation.

According to the above analysis, we can see that DDConv realizes the dynamic adjustment of the shape and weights of the convolution kernel by combining the convolution kernel deformation offset and the convolution kernel weight coefficient with a minimal number of calculation. Compared with directly increasing the number and size of convolution kernels, the DDConv is simpler and more efficient. The proposed DDConv not only solves the problem of poor adaptive feature extraction ability of fixed-size convolution kernels but also overcomes the defect that different inputs share the same convolution kernel parameters. Consequently, our DDConv can be used to improve the segmentation accuracy of small targets and large targets with blurred edges in medical images.

### 2.3 Shifted Window Adaptive Complementary Attention Module

The self-attention mechanism is the core computing unit in Transformer networks, which realizes the capture of long-range dependency of feature maps by utilizing matrix operations. However, the self-attention mechanism only considers the dependency in the spatial dimension but not the cross-dimensional dependency between spatial and channels [Hong *et al.*, 2021]. Therefore, when dealing with medical image segmentation with low contrast and high density noise, the self-attention mechanism is easy to confuse the segmentation targets with the background, resulting in poor segmentation results.

To solve the problems mentioned above, we propose a new cross-dimensional self-attention module called (S)W-ACAM. As shown in Figure 3, (S)W-ACAM has four parallel branches, the top two branches are the conventional dual attention module [Liu *et al.*, 2021a] and the bottom

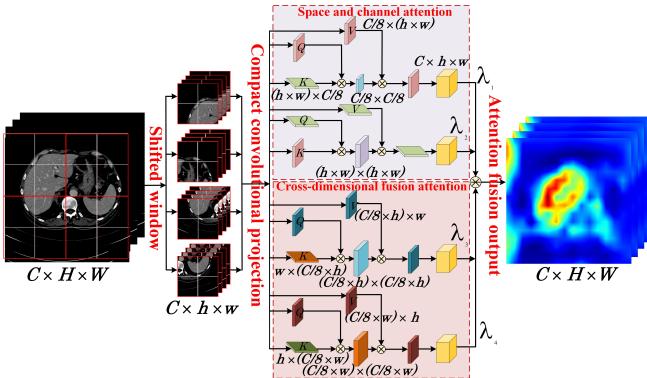


Figure 3: The module of the proposed (S)W-ACAM. Unlike conventional self-attention, (S)W-ACAM has the advantages of spatial and channel attention, and can also capture long-distance correlation features between spatial and channels. Through the shifted window operation, the spatial resolution of the image is significantly reduced, and through the compact convolutional projection operation, the channel dimension of the image is also significantly reduced. Thus, the overall computational costs and complexity of the network are reduced.  $\lambda_1, \lambda_2, \lambda_3$  and  $\lambda_4$  are learnable weight parameters.

two branches are cross-dimensional attention modules. Compared to popular self-attention modules such as spatial self-attention, channel self-attention, and dual self-attention, our proposed (S)W-ACAM can not only fully extract the long-range dependency of spatial and channels, but also capture the cross-dimensional long-range dependency between spatial and channels. These four branches complement each other, provide richer long-range dependency relationships, enhance the separability between the foreground and background, and thus improve the segmentation results for medical images.

The standard Transformer architecture [Dosovitskiy *et al.*, 2020] uses the global self-attention method to calculate the relationship between one token and all other tokens. This calculation method is complex, especially in the face of high-resolution and intensive prediction tasks like medical images where the computational costs will increase exponentially. In order to improve the calculation efficiency, we use the shifted window calculation method similar to that in Swin Transformer [Liu *et al.*, 2021b], which only calculates the self-attention in the local window. However, in the face of our (S)W-ACAM four branches module, using the shifted window method to calculate self-attention does not reduce the overall computational complexity of the module. Therefore, we also designed the compact convolutional projection. First, we reduce the local size of the medical image through the shifted window operation, then we compress the channel dimension of feature maps through the compact convolutional projection, and finally calculate the self-attention. It is worth mentioning that this method can not only better capture the global high-dimensional information of medical images but also significantly reduce the computational costs of the module. Suppose an image contains  $h \times w$  windows, each window size is  $M \times M$ , then the complexity of the (S)W-ACAM, the global MSA in the original Transformer, and the (S)W-MSA

in the Swin Transformer are compared as follows:

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C, \quad (6)$$

$$\Omega((S)W-MSA) = 4hwC^2 + 2M^2hwC, \quad (7)$$

$$\Omega((S)W-ACAM) = \frac{hwC^2}{4} + M^2hwC. \quad (8)$$

if the former term of each formula is a quadratic function of the number of patches  $hw$ , the latter term is linear when  $M$  is fixed (the default is 7). Then the computational costs of (S)W-ACAM are smaller compared with MSA and (S)W-MSA.

Among the four parallel branches of (S)W-ACAM, two branches are used to capture channel correlation and spatial correlation, respectively, and the remaining two branches are used to capture the correlation between channel dimension  $C$  and space dimension  $H$  and vice versa (between channel dimension  $C$  and space dimension  $W$ ). After adopting the shifted window partitioning method, as shown in Figure 2 (b), the calculation process of continuous Transformer blocks is as follows:

$$T^l = W\text{-ACAM}(LN(T^{l-1})) + T^{l-1}, \quad (9)$$

$$T^l = LPM(LN(\hat{T}^l)) + \hat{T}^l, \quad (10)$$

$$\hat{T}^{l+1} = SW\text{-ACAM}(LN(T^l)) + T^l, \quad (11)$$

$$T^{l+1} = LPM(LN(\hat{T}^{l+1})) + \hat{T}^{l+1}. \quad (12)$$

where  $\hat{T}^l$  and  $T^l$  represent the output features of (S)W-ACAM and LPM, respectively. W-ACAM represents window adaptive complementary attention, SW-ACAM represents shifted window adaptive complementary attention, and LPM represents lightweight perceptron module. For the specific attention calculation process of each branch, we follow the same principle in Swin Transformer as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{C/8}} + B\right)V, \quad (13)$$

where relative position bias  $B \in \mathbb{R}^{M^2 \times M^2}$ ,  $Q, K, V \in \mathbb{R}^{M^2 \times \frac{C}{8}}$  are query, key, and value matrices respectively.  $\frac{C}{8}$  represents the dimension of query/key, and  $M^2$  represents the number of patches.

After four parallel attention branches  $Out_1, Out_2, Out_3$  and  $Out_4$  are calculated, the final feature fusion output is:

$$Out = \lambda_1 \cdot Out_1 + \lambda_2 \cdot Out_2 + \lambda_3 \cdot Out_3 + \lambda_4 \cdot Out_4, \quad (14)$$

where  $\lambda_1, \lambda_2, \lambda_3$  and  $\lambda_4$  are learnable parameters that enable adaptive control of the importance of each attention branch for spatial and channel information in a particular segmentation task through the back-propagation process of the segmentation network.

Different from other self-attention mechanisms, the (S)W-ACAM in this paper can fully capture the correlation between spatial and channels, and reasonably use the context information of medical images to achieve long-range dependence

	<b>Method</b>	<b>DI</b> ↑	<b>JA</b> ↑	<b>SE</b> ↑	<b>AC</b> ↑	<b>SP</b> ↑	<b>Para.</b> (M) ↓	<b>GFLOPs</b>
<b>CNNs</b>	U-Net [Ronneberger <i>et al.</i> , 2015]	86.54	79.31	88.56	93.16	96.44	34.52	65.39
	R2UNet [Alom <i>et al.</i> , 2018]	87.92	80.28	90.92	93.38	96.33	39.09	152.82
	Attention Unet [Oktay <i>et al.</i> , 2018]	87.16	79.55	88.52	93.17	95.62	34.88	66.57
	CENet [Gu <i>et al.</i> , 2019]	87.61	81.18	90.71	94.03	96.35	29.02	11.79
	CPFNet † [Feng <i>et al.</i> , 2020]	90.18	82.92	91.66	94.68	96.63	30.65	<b>9.15</b>
<b>Transformer</b>	Swin-Unet † [Cao <i>et al.</i> , 2021]	89.26	80.47	90.36	94.45	96.51	41.40	11.63
	TransUNet † [Chen <i>et al.</i> , 2021]	89.39	82.10	91.43	93.67	96.54	105.30	15.21
	BAT † [Wang <i>et al.</i> , 2021a]	90.21	83.49	91.59	94.85	96.57	45.56	13.38
	CvT † [Wu <i>et al.</i> , 2021]	88.23	80.21	87.60	93.68	96.28	21.51	20.53
	PVT [Wang <i>et al.</i> , 2021b]	87.31	79.99	87.74	93.10	96.21	28.86	14.92
	CrossForm [Wang <i>et al.</i> , 2021c]	87.44	80.06	88.25	93.39	96.40	38.66	13.57
<b>CiT-Net-T (our)</b>		<b>90.72</b>	<b>84.59</b>	<b>92.54</b>	<b>95.21</b>	<b>96.83</b>	<b>11.58</b>	<b>4.53</b>
<b>CiT-Net-B (our)</b>		<b>91.23</b>	<b>84.76</b>	<b>92.68</b>	<b>95.56</b>	<b>98.21</b>	<b>21.24</b>	13.29

Table 1: Performance comparison of the proposed method against the SOTA approaches on the ISIC2018 benchmarks. **Red** indicates the best result, and **blue** displays the second-best.

† indicates the model is initialized with pre-trained weights on the ImageNet21K. “Para.” refers to the number of parameters. “GFLOPs” is calculated under the input scale of  $224 \times 224$ . Since the dermoscopic images are 2D medical images, the comparison methods are all 2D networks.

modeling. Since our (S)W-ACAM effectively overcomes better feature representation of the defect that the conventional self-attention only focuses on the spatial self-attention of images and ignores the channel and cross-dimensional self-attention, it achieves the best image suffers from large noise, low contrast, and complex background.

## 2.4 Architecture Variants

We have built a CiT-Net-T as a base network with a model size of 11.58 M and a computing capacity of 4.53 GFLOPs. In addition, we built the CiT-Net-B network to make a fair comparison with the latest networks such as CvT [Wu *et al.*, 2021] and PVT [Wang *et al.*, 2021b]. The window size is set to 7, and the input image size is  $224 \times 224$ . Other network parameters are set as follows:

- CiT-Net-T: *layer number* = {2, 2, 6, 2, 6, 2, 2},  $H = \{3, 6, 12, 24, 12, 6, 3\}$ ,  $D = 96$
- CiT-Net-B: *layer number* = {2, 2, 18, 2, 18, 2, 2},  $H = \{4, 8, 16, 32, 16, 8, 4\}$ ,  $D = 96$ ,

$D$  represents the number of image channels when entering the first layer of the dynamically adaptive CNNs branch and the cross-dimensional fusion Transformer branch, *layer number* represents the number of Transformer blocks used in each stage, and  $H$  represents the number of multiple heads in self-attention.

## 3 Experiment and Results

### 3.1 Datasets

We conducted experiments on the skin lesion segmentation dataset ISIC2018 from the International Symposium on Biomedical Imaging (ISBI) and the Liver Tumor Segmentation Challenge dataset (LiTS) from the Medical Image Computing and Computer Assisted Intervention Society (MICCAI). The ISIC2018 contains 2,594 dermoscopic images for training, but the ground truth images of the testing set

have not been released, thus we performed a five-fold cross-validation on the training set for a fair comparison. The LiTS contains 131 3D CT liver scans, where 100 scans of which are used for training, and the remaining 31 scans are used for testing. In addition, all images are empirically resized to  $224 \times 224$  for efficiency.

### 3.2 Implementation Details

All the networks are implemented on NVIDIA GeForce RTX 3090 24GB and PyTorch 1.7. We utilize Adam with an initial learning rate of 0.001 to optimize the networks. The learning rate decreases in half when the loss on the validation set has not dropped by 10 epochs. We used mean squared error loss (MSE) and Dice loss as loss functions in our experiment.

### 3.3 Evaluation and Results

In this paper, we selected the mainstream medical image segmentation networks U-Net [Ronneberger *et al.*, 2015], Attention Unet [Oktay *et al.*, 2018], Swin-Unet [Cao *et al.*, 2021], PVT [Wang *et al.*, 2021b], CrossForm [Wang *et al.*, 2021c] and the proposed CiT-Net to conduct a comprehensive comparison of the two different modalities datasets, ISIC2018 and the LiTS.

In the experiment of the ISIC2018 dataset, we made an overall evaluation of the mainstream medical image segmentation network by using five indicators: Dice (DI), Jaccard (JA), Sensitivity (SE), Accuracy (AC), and Specificity (SP). Table 1 shows the quantitative analysis of the results of the proposed CiT-Net and the current mainstream CNNs and Transformer networks in the ISIC2018 dataset. From the experimental results, we can conclude that our CiT-Net has the minimum number of parameters and the lowest computational costs, and can obtain the best segmentation effect on the dermoscopic images without adding pre-training. Moreover, our CiT-Net-T network has only 11.58 M of parameters and 4.53 GFLOPs of computational costs, but still achieves the second-best segmentation effect. Our CiT-Net-B network,

<b>Method</b>	<b>DI</b> $\uparrow$	<b>VOE</b> $\downarrow$	<b>RVD</b> $\downarrow$	<b>ASD</b> $\downarrow$	<b>RMSD</b> $\downarrow$	<b>Para. (M)</b> $\downarrow$	<b>GFLOPs</b>	
CNNs	U-Net [Ronneberger <i>et al.</i> , 2015]	93.99 $\pm$ 1.23	11.13 $\pm$ 2.47	3.22 $\pm$ 0.20	5.79 $\pm$ 0.53	123.57 $\pm$ 6.28	34.52	65.39
	R2UNet [Alom <i>et al.</i> , 2018]	94.01 $\pm$ 1.18	11.12 $\pm$ 2.37	2.36 $\pm$ 0.15	5.23 $\pm$ 0.45	120.36 $\pm$ 5.03	39.09	152.82
	Attention Unet [Oktay <i>et al.</i> , 2018]	94.08 $\pm$ 1.21	10.95 $\pm$ 2.36	3.02 $\pm$ 0.18	4.95 $\pm$ 0.48	118.67 $\pm$ 5.31	34.88	66.57
	CENet [Gu <i>et al.</i> , 2019]	94.04 $\pm$ 1.15	11.03 $\pm$ 2.31	6.19 $\pm$ 0.16	4.11 $\pm$ 0.51	115.40 $\pm$ 5.82	29.02	<b>11.79</b>
	3D Unet [Çiçek <i>et al.</i> , 2016]	94.10 $\pm$ 1.06	11.13 $\pm$ 2.23	<b>1.42<math>\pm</math>0.13</b>	2.61 $\pm$ 0.45	36.43 $\pm$ 5.38	40.32	66.45
	V-Net [Milletari <i>et al.</i> , 2016]	94.25 $\pm$ 1.03	10.65 $\pm$ 2.17	1.92 $\pm$ 0.11	2.48 $\pm$ 0.38	38.28 $\pm$ 5.05	65.17	55.35
Transformer	Swin-Unet † [Cao <i>et al.</i> , 2021]	95.62 $\pm$ 1.32	9.73 $\pm$ 2.16	2.78 $\pm$ 0.21	2.35 $\pm$ 0.35	38.85 $\pm$ 5.42	41.40	11.63
	TransUNet † [Chen <i>et al.</i> , 2021]	95.79 $\pm$ 1.09	9.82 $\pm$ 2.10	1.98 $\pm$ 0.15	2.33 $\pm$ 0.41	37.22 $\pm$ 5.23	105.30	15.21
	CvT † [Wu <i>et al.</i> , 2021]	95.81 $\pm$ 1.25	9.66 $\pm$ 2.31	1.77 $\pm$ 0.16	2.34 $\pm$ 0.29	36.71 $\pm$ 5.09	21.51	20.53
	PVT [Wang <i>et al.</i> , 2021b]	94.56 $\pm$ 1.15	9.75 $\pm$ 2.19	1.69 $\pm$ 0.12	2.42 $\pm$ 0.34	37.35 $\pm$ 5.16	28.86	14.92
	CrossForm [Wang <i>et al.</i> , 2021c]	94.63 $\pm$ 1.24	9.72 $\pm$ 2.24	1.65 $\pm$ 0.15	2.39 $\pm$ 0.31	37.21 $\pm$ 5.32	38.66	13.57
	<b>CiT-Net-T (our)</b>	<b>96.48<math>\pm</math>1.05</b>	<b>9.53<math>\pm</math>2.11</b>	1.45 $\pm$ 0.12	<b>2.29<math>\pm</math>0.33</b>	<b>36.21<math>\pm</math>4.97</b>	<b>11.58</b>	<b>4.53</b>
	<b>CiT-Net-B (our)</b>	<b>96.82<math>\pm</math>1.22</b>	<b>9.46<math>\pm</math>2.33</b>	<b>1.38<math>\pm</math>0.13</b>	<b>2.21<math>\pm</math>0.35</b>	<b>36.08<math>\pm</math>4.88</b>	<b>21.24</b>	13.29

Table 2: Performance comparison of the proposed method against the SOTA approaches on the LiTS-Liver benchmarks. **Red** indicates the best result, and **blue** displays the second-best.

† indicates the model initialized with pre-trained weights on ImageNet21K. “Para.” refers to the number of parameters. “GFLOPs” is calculated under the input scale of  $224 \times 224$ . Compared with the comparison experiment oin the ISIC2018 dataset, 3D Unet and V-Net are introduced into the comparison experiment oin the LiTS-Liver dataset.

BAT, CvT, and CrossForm have similar parameters or computational costs, but in the ISIC2018 dataset, the division Dice value of our CiT-Net-B is 1.02%, 3.00%, and 3.79% higher than that of the BAT, CvT, and CrossForm network respectively. In terms of other evaluation indicators, our CiT-Net-B is also significantly better than other comparison methods.

In the experiment of the LiTS-Liver dataset, we conducted an overall evaluation of the mainstream medical image segmentation network by using five indicators: DI, VOE, RVD, ASD and RMSD. Table 2 shows the quantitative analysis of the results of the proposed CiT-Net and the current mainstream networks in the LiTS-Liver dataset. It can be seen from the experimental results that our CiT-Net has great advantages in medical image segmentation, which further verifies the integrity of CiT-Net in preserving local and global features in medical images. It is worth noting that the CiT-Net-B and CiT-Net-T networks have achieved good results in medical image segmentation in the first and second place, with the least number of model parameters and computational costs. The division Dice value of our CiT-Net-B network without pre-training is 1.20%, 1.03%, and 1.01% higher than that of the Swin-Unet, TransUNet, and CvT network with pre-training. In terms of other evaluation indicators, our CiT-Net-B is also significantly better than other comparison methods.

<b>Backbone</b>	<b>DDConv</b>	<b>(S)W-ACAM</b>	<b>LPM</b>	<b>Para. (M)</b>	<b>DI (%)</b> $\uparrow$
U-Net+Swin-Unet				46.92	87.45
U-Net+Swin-Unet	✓			48.25	89.15
U-Net+Swin-Unet		✓		30.26	89.62
U-Net+Swin-Unet			✓	15.45	88.43
U-Net+Swin-Unet	✓	✓		32.16	90.88
U-Net+Swin-Unet	✓		✓	16.93	89.12
U-Net+Swin-Unet		✓	✓	9.67	89.46
CiT-Net-T (our)	✓	✓	✓	11.58	90.72

Table 3: Ablation experiments of DDConv, (S)W-ACAM and LPM in CiT-Net in the ISIC2018 dataset.

### 3.4 Ablation Study

In order to fully prove the effectiveness of different modules in our CiT-Net, we conducted a series of ablation experiments on the ISIC2018 dataset. As shown in Table 3, we can see that the Dynamic Deformable Convolution (DDConv) and (Shifted) Window Adaptive Complementary Attention Module ((S)W-ACAM) proposed in this paper show good performance, and the combination of these two modules, CiT-Net shows the best medical image segmentation effect. At the same time, the Lightweight Perceptron Module (LPM) can significantly reduce the overall parameters of the CiT-Net.

## 4 Conclusion

In this study, we have proposed a new architecture CiT-Net that combines dynamically adaptive CNNs and cross-dimensional fusion Transformer in parallel for medical image segmentation. The proposed CiT-Net integrates the advantages of both CNNs and Transformer, and retains the local details and global semantic features of medical images to the maximum extent through local relationship modeling and long-range dependency modeling. The proposed DDConv overcomes the problems of fixed receptive field and parameter sharing in vanilla convolution, enhances the ability to express local features, and realizes adaptive extraction of spatial features. The proposed (S)W-ACAM self-attention mechanism can fully capture the cross-dimensional correlation between feature spatial and channels, and adaptively learn the important information between spatial and channels through network training. In addition, by using the LPM to replace the MLP in the traditional Transformer, our CiT-Net significantly reduces the number of parameters, gets rid of the dependence of the network on pre-training, avoids the challenge of the lack of labeled medical image data and easy over-fitting of the network. Compared with popular CNNs and Transformer medical image segmentation networks, our CiT-Net shows significant advantages in terms of operational efficiency and segmentation effect.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62271296, 62201334 and 62201452, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2021JC-47, and in part by the Key Research and Development Program of Shaanxi under Grants 2022GY-436 and 2021ZDLGY08-07.

## References

- [Alom *et al.*, 2018] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018.
- [Azad *et al.*, 2022] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *arXiv preprint arXiv:2211.14830*, 2022.
- [Cao *et al.*, 2021] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.
- [Chen *et al.*, 2020] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020.
- [Chen *et al.*, 2021] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- [Çiçek *et al.*, 2016] Özgün Çiçek, Ahmed Abdulkadir, Søren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.
- [Dai *et al.*, 2017] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.
- [Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [Feng *et al.*, 2020] Shuanglang Feng, Heming Zhao, Fei Shi, Xuena Cheng, Meng Wang, Yuhui Ma, Dehui Xiang, Weifang Zhu, and Xinjian Chen. Cpfnet: Context pyramid fusion network for medical image segmentation. *IEEE transactions on medical imaging*, 39(10):3008–3018, 2020.
- [Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [Gu *et al.*, 2019] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019.
- [Gu *et al.*, 2020] Ran Gu, Guotai Wang, Tao Song, Rui Huang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE transactions on medical imaging*, 40(2):699–711, 2020.
- [Han *et al.*, 2020] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020.
- [Hong *et al.*, 2021] Luminzi Hong, Risheng Wang, Tao Lei, Xiaogang Du, and Yong Wan. Qau-net: Quartet attention u-net for liver and liver-tumor segmentation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.
- [Lei *et al.*, 2022] Tao Lei, Dong Zhang, Xiaogang Du, Xuan Wang, Yong Wan, and Asoke K Nandi. Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. *IEEE Transactions on Medical Imaging*, 2022.
- [Lei *et al.*, March 2023] Tao Lei, Rui Sun, Xiaogang Du, Huazhu Fu, Changqing Zhang, and Asoke K Nandi. Sgu-net: Shape-guided ultralight network for abdominal image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(3):1431–1442, March, 2023.
- [Li *et al.*, 2021] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12321–12330, 2021.
- [Liu *et al.*, 2021a] Xin Liu, Guobao Xiao, Luanyuan Dai, Kun Zeng, Changcai Yang, and Riqing Chen. Scsa-net: Presentation of two-view reliable correspondence learning via spatial-channel self-attention. *Neurocomputing*, 431:137–147, 2021.
- [Liu *et al.*, 2021b] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using

- shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.
- [Liu *et al.*, 2023] Yanhong Liu, Ji Shen, Lei Yang, Guibin Bian, and Hongnian Yu. Resdo-unet: A deep residual network for accurate retinal vessel segmentation from fundus images. *Biomedical Signal Processing and Control*, 79:104087, 2023.
- [Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [Lv *et al.*, 2022] Peiqing Lv, Jinke Wang, and Haiying Wang. 2.5 d lightweight riu-net for automatic liver and tumor segmentation from ct. *Biomedical Signal Processing and Control*, 75:103567, 2022.
- [Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.
- [Oktay *et al.*, 2018] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Matthias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.
- [Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [Shamshad *et al.*, 2022] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*, 2022.
- [Song *et al.*, 2020] Jiarui Song, Beibei Li, Yuhao Wu, Yixin Shi, and Aohan Li. Real: A new resnet-alstm based intrusion detection system for the internet of energy. In *2020 IEEE 45th Conference on Local Computer Networks (LCN)*, pages 491–496. IEEE, 2020.
- [Suetens, 2017] Paul Suetens. *Fundamentals of medical imaging*. Cambridge university press, 2017.
- [Sun *et al.*, 2022] Rui Sun, Tao Lei, Qi Chen, Zexuan Wang, Xiaogang Du, Weiqiang Zhao, and A Nandi. Survey of image edge detection. *Frontiers in Signal Processing*, 2(1):1–13, 2022.
- [Tang *et al.*, 2022] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.
- [Valanarasu *et al.*, 2021] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2021.
- [Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [Wang *et al.*, 2021a] Jiacheng Wang, Lan Wei, Liansheng Wang, Qichao Zhou, Lei Zhu, and Jing Qin. Boundary-aware transformers for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 206–216. Springer, 2021.
- [Wang *et al.*, 2021b] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.
- [Wang *et al.*, 2021c] Wenxiao Wang, Lu Yao, Long Chen, Binbin Lin, Deng Cai, Xiaofei He, and Wei Liu. Crossformer: A versatile vision transformer hinging on cross-scale attention. *arXiv preprint arXiv:2108.00154*, 2021.
- [Wang *et al.*, 2022] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2441–2449, 2022.
- [Wu *et al.*, 2021] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.
- [Yang *et al.*, 2022] Xin Yang, Zhiqiang Li, Yingqing Guo, and Dake Zhou. Dcu-net: a deformable convolutional neural network based on cascade u-net for retinal vessel segmentation. *Multimedia Tools and Applications*, 81(11):15593–15607, 2022.
- [Zhou *et al.*, 2018] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming 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*, pages 3–11. Springer, 2018.