IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 January 2024, accepted 29 January 2024, date of publication 6 February 2024, date of current version 13 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3363410

RESEARCH ARTICLE

RotU-Net: An Innovative U-Net With Local Rotation for Medical Image Segmentation

FUXIANG ZHANG, FENGCHAO WANG, WENFENG ZHANG, QUANZHEN WANG, YAJUN LIU, AND ZHIMING JIANG, (Member, IEEE)
Department of Critical Care Medicine, The First Affiliated Hospital of Shandong First Medical University, Jinan, Shandong 250014, China

Corresponding author: Zhiming Jiang (jiang7708@sina.com)

This work was supported in part by the Clinical Research Fund of Shandong Medical Association under Grant YXH2022ZX02090, and in part by the National Key Research and Development Program of China under Grant 2018YFC2002000.

ABSTRACT In recent years, both convolutional neural networks (CNN) and transformers have demonstrated impressive feature extraction capabilities in the field of medical image segmentation. A common approach is to utilize a combination of CNN and transformer encoders to efficiently learn both local and global features, making them widely adopted techniques in semantic segmentation of medical images. However, challenges remain due to the limited sample size of medical image datasets and the intricate foreground edge information in these images. These challenges make it difficult for models to capture key structures and information related to foreground edge details, especially when trained on smaller datasets. To address these issues, we propose a U-Net-based model called ''Rotate U-Net'' (RotU-Net). Our model design is inspired by the successful U-Net architecture, which is characterized by direct connections between encoders and decoders, and skipping connections at multiple resolutions. Meanwhile, we propose weight rotator as a feature extraction module, which enhances network to discriminate edge information in the foreground region by computing partial element correlations to improve the network to focus on the foreground region while reducing redundant information in the features. Finally, we have validated RotU-Net on the Synapse Multi-Organ Segmentation Dataset (Synapse) and the Segmentation of Multiple Myeloma Plasma Cells in Microscopic Images (SegPC). The experimental results show that RotU-Net with a very small number of parameters achieves impressive performance, which demonstrates the effectiveness and efficiency of RotU-Net.

INDEX TERMS Medical image segmentation, U-Net, weight rotation.

I. INTRODUCTION

With the development of computer vision field, advanced techniques of artificial intelligence and deep learning have been widely used in the medical field [1]. Medical image segmentation is pivotal in the field of medical image analysis and is a fundamental component of structural examination. Accurate and reliable medical image segmentation algorithms can help doctors to diagnose disease types, monitor disease progression and predict disease severity and consequences, and also accelerate the automation and intelligence of medical image processing [2], [3], [4].

Since the speedy development of deep learning, CNN have been favored by researchers in the field of medical image segmentation because of their robust feature extraction, and as a result, a multitude of CNN-based networks have been widely used in the field [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. The networks are not only able to adapt to arbitrary size inputs, but also ensure the output of accurate results by increasing the upsampling of data size. Notably, U-Net is a widely used CNN-based network [11], which is usually used in the task of segmenting organs, lesions and structures in medical images such as CT scans, MRIs, X-rays, and it achieved excellent results [12], [13], [14], [15]. Its network structure mainly has encoder and decoder, where the encoder learns the global contextual representation by downsampling the extracted features layer by layer, and the

The associate editor coordinating the review of this manuscript and approving it for publication was Francesco Mercaldo.

decoder reduces the extracted representation to the resolution used for pixel semantic segmentation by upsampling layer by layer. In addition, in order to recover spatial information that may have been lost during the downsampling process and to enhance the ability of the network to capture foreground detail, U-Net employs skip connections to link the outputs of encoders and decoders of different resolutions. Although such U-shaped convolutional networks have a powerful ability to learn representations and have shown excellent performance, they learn remote dependencies in a way that only passes through the local receptive field, which leads to a lack of remote modeling ability, for example, unsatisfactory segmentation of structures of different shapes and scales. Some researchers have tried to solve the above problems by using dilated convolution or increasing the convolution kernel [16], [17], which increases the receptive field to a certain extent but also loses part of the detail information. Therefore, there is still potential for improving the resolution of intrinsic constraints associated with convolutional kernels.

In the past few years, transformer has achieved state-of-the-art (SOTA) results in the field of natural language processing (NLP) on a variety of tasks [18]. The self-attention mechanism in transformer achieves it by modeling the importance of each element with respect to the other elements, allowing the model to dynamically assign corresponding weights to each element and to focus on important elements in the feature map, thus achieving the ability to capture the global contextual information. Thanks to the development of transformer, more and more academics have applied transformer in computer vision [19], [20], [21], [22], [23]. Transformer can help the network to perceive the relationship between the elements in the image, and improve the model performance when dealing with images with a large range of organ structures. Among them, Cao et al. proposed SwinUNet to solve the problem of poor local modeling ability of transformer, which is difficult to focus on the local information of foreground region effectively [24]. SwinUNet is a U-shaped network composed entirely of transformers, and local correlation computation is achieved by designing sliding windows, which solves the problem of poor extraction of local features and reduces the computational complexity, but it cannot change the weak local modeling of transformers. Chen et al. proposed transUNet combining U-Net and transformer [26], which has the advantages of both and is a powerful alternative for medical image segmentation. With the combination of U-Net, transformers can be used as a powerful encoder for segmentation tasks by recovering the local spatial information.The proposal of transUNet lays the foundation for subsequent research combining CNN and transformer. At the same time, the weights of the transformer need to be pre-trained using a large amount of dataset, which may lead to unsatisfactory results on small sample datasets. Although all of the above proposed networks can effectively achieve global and local elemental modeling, they produce redundant information in calculating elemental correlations, which makes the networks pay little attention to the foreground region. Meanwhile, the edge structure and information of the foreground region is beneficial for segmentation, which has often been neglected in past studies. Considering the shortcomings in past studies, we propose RotU-Net. Overall, we make the following contributions:

- We designed the weight rotator, a module that breaks the fixed spatial structure of the feature map through local image rotation, allows features at different locations to interact with each other remotely and computes some of their correlations, thus enhancing the ability of network to utilize foreground information as well as discriminate the edge structure and information in the foreground region.
- We propose a new U-Net-based network: RotU-Net, which retains the design of U-shaped structure, extracts features layer by layer by CNN and calculates the correlation of some important elements by weight rotator, which enhances our model to focus on foreground regions and better localizes and segments foreground regions.
- Our model is validated on Synapse and SegPC datasets. Results of the experiment validate and demonstrate the effectiveness of RotU-Net.

II. RELATED WORK

A. CNN-BASED NETWORK ARCHITECTURE

Before deep learning was widely used, medical image segmentation was mainly performed by manually segmenting the foreground region or by traditional machine learning methods. Since the introduction of CNN in medical image tasks, the number of CNN-based segmentation networks has exploded and achieved SOTA on various medical image segmentation datasets. e.g., FCNs [9], V-Net [10], U-Net [11], R50 U-Net [26], DARR [27] and U-Net++ [28]. Long et al. proposed fully convolutional network (FCNs) to solve the segmentation problem [9], which utilizes softmax to obtain the classification information for each pixel point and achieve pixel-level prediction. Milletari et al. proposed a 3D image segmentation method based on a fully convolutional neural network: the V-Net [10], which utilizes a V-shaped CNN to learn the complex structures in medical images, and can achieve high-precision segmentation, while the V-shaped structure provides a new method for later segmentation networks. The working principle of U-Net is mainly to learn the global contextual representation through layer-by-layer downsampling in the encoding stage, and to reduce the extracted representation through layer-by-layer upsampling in the decoding stage, while using skip connections to link the encoder and decoder with different resolution. However, with the application of U-Net, more researchers have found some problems with the U-Net structure and proposed improvement methods [26], [27], [28], [29], [30]. Although the skip connections of U-Net improves the problem of losing too much spatial information in the downsampling process, some researchers believe that such a connection brings the problem of semantic divide. So as to improve the

TABLE 1. Part of the related work overview.

Architecture	Methods	Model Characteristics
Based on CNN	FCNs [9]	In the field of medical image segmentation, the concept of full convolutional network is proposed for the first time and pixel level prediction is carried out.
	V-Net [10]	A 3D image segmentation method based on volume and full convolutional neural network.
	U-Net [11]	The U-shaped skip connection structure can concatenate low-level features and high-level semantic features.
	U-Net++ [28]	Redesigning skip connections to aggregate semantically scaled different features on decoder subnetworks, resulting in highly flexible feature fusion methods.
	U-Net3+ [29]	Full-scale skip connections combine high-level semantics from feature maps at different scales directly with low-level semantics.
	R50 Att-UNet [26]	Resnet and attention ideas, the network does a combination of both.
Based on transformer	SwinUNet [24]	A purely transformer-based U-shaped architecture with skip connections.
	DAE-Former [31]	A dual-attention mechanism based transformer architecture to capture spatial and channel relationships across feature dimensions.
	Segtran [32]	Global context and local features can be acquired simultaneously with the improved Squeese-and-Expansion transformer layer.
Based on CNN and transformer	TransUNet [25]	The transformer encodes the CNN feature map as a sequence of contexts, and the decoder upsamples the encoded features, which are then combined with the high-resolution feature map for accurate segmentation.
	TransClaw [33]	Combining CNN with transformer in the encoding stage. Global context information is efficiently captured with the help of multi-head self-attention, which compensates for the limitations of convolution.
	TDs-TransUNet-transUNet [34]	A swin transformer-based dual-scale encoder subnetwork is used to extract coarse-grained and fine-grained feature representations at different semantic scales.
	HiFormer [35]	Extracting two multi-scale feature representations with the groundbreaking swin transformer module and CNN-based encoder.

semantic divide issues of U-Net, researchers have proposed the UNet++ and UNet3+ [28], [29], and the formers solve the semantic divide problem through a complex nested connection structure, and the latter combines different levels of semantic information through a full-scale connection framework. However, the improved skip connections increase the computational complexity of the model and are not conducive to model generalization. In terms of improving U-Net to extract foreground information in feature maps. Xiao et al. proposed to use skip connections operation in the encoder of R50 U-Net to deepen the depth of downsampling and expand the sensory field [26], which is useful for segmenting foreground regions with large targets. However, the network structure with too much depth causes too much detail information to be lost during downsampling, leading to poor performance in small target segmentation. Meantime, scholars are working on solving small target segmentation. Xiao et al. added attention module to the R50 U-Net [26] to make the model focus on the foreground region, and the introduction of the attention module at the local level also accelerated the development of small target segmentation methods for medical images. Overall, a very large number of researchers have improved the segmentation results of medical images by improving the U-shaped network, and their methods have gained a certain degree of success. However, CNN-based network architectures have inherent structural drawbacks: the convolutional kernel can only capture small-sized local information aspects because of its fixed

receptive field, which leads to an inability to learn global contextual information, and it performs poorly in establishing long-range spatial dependencies, which seriously affects the segmentation results. Therefore, the U-shaped network based on CNN still has room for improvement.

B. TRANSFORMER-BASED NETWORK ARCHITECTURE

Because of transformer powerful context learning and global modeling capabilities, and it was also quickly applied in various domains. Dosovitskiy et.al first proposed vit-transformer [19], which explored the potential of transformer in computer vision. Vit-transformer has achieved SOTA performance on large datasets in various computer vision tasks, but the results achieved on small datasets are unsatisfactory. This is due to the fact that the transformer has too many parameters, which makes it difficult to converge when trained on small datasets. Therefore, researchers have made many efforts to speed up transformer training and and improve model performance in computer vision tasks. Cao et. al proposed SwinUNet [24] in order to speed up transformer training while ensuring model performance. SwinUnet is a purely transformer-based u-type codec structure with skip connections. It reduces the overall computational complexity of the model through the W-MSA module and computes the feature maps through the attention mechanism of the transformer, while the SW-MSA module is proposed to realize the information interaction of local features. Azad et al. proposed DAE-Former [31], a novel transformer architecture guided by a dual attention mechanism, whose main contribution is to maintain computational efficiency by restricting the attention mechanism to a localized region while capturing spatial and channel relationships across the feature dimension. In addition, skip connection paths are redesigned by means of a cross-attention module to ensure feature reusability and to enhance the segmentation capability of the model.Li et al. proposed Segtran [32], which centers on the possibility of simultaneously capturing global context and detail features with the help of an improved Squeese-and-Expansion Transformer layer, while using a new feature encoding approach that can apply continuous inductive bias to the image. The transformer is widely used for its excellent global modeling capabilities and superior performance on large datasets. However, the transformer tends to ignore the local information in the feature map when computing the global information through the attention mechanism. For medical image segmentation, the sample size of the dataset is limited, and at the same time, the foreground regions in medical images often require the model to extract their local information. In summary, for medical image segmentation models, using only the transformer structure will affect the final segmentation results.

C. NETWORK ARCHITECTURE WITH COMBINATION OF CNN AND TRANSFORMER

In order to improve the insufficiency of CNN global feature extraction and transformer local feature extraction, Chen et al. made the attempt to combine U-Net and transformer and proposed TransUNet [25]. It extended the U-shape network structure and skip connections of U-Net and added the last layer of downsampling to the transformer. In subsequent studies, researchers have improved the network based on TransUNet to enhance the segmentation performance. For example, Yao et al. proposed TransClaw U-Net [33], which combines a convolutional network with a transformer in the encoding stage, where the convolutional part is responsible for extracting shallow spatial features to help recover the resolution of the image after up-sampling. the transformer part is responsible for encoding the patches. The decoding phase preserves the structure of the upsampling to achieve better detail segmentation performance. However, such a network structure also brings great computational complexity and greatly occupies computational resources. Chen et al. proposed TDs-TransUNet based on swin transformer [34], of which dual-scale encoder to extract different coarse-grained and fine-grained feature representations, while the Transformer Interactive Fusion (TIF) in the network effectively establishes the dependencies between multiscale features. Heidari et al. extracted two multi-scale feature representations in HiFormer with the pioneering swin transformer module and CNN-based encoder [35], and proposed to cross-fertilize the two representations with global and local features using the DLF fusion module. In addition, several excellent CNN and transformer combination methods have emerged in the field of medical image segmentation [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], and these hybrid structured networks aim at balancing the weights of the proposed obtained local and global information, but at the expense of an excessive number of model parameters and a complex model structure.

In order to show more clearly the related work in recent years, we show some of the related work overview in Table 1.

III. METHODS

In this section, we introduce RotU-Net, and model overview is shown in Figure 1. RotU-Net mainly consists of encoding phase and decoding phase. The model components included in the encoding phase are the CNN layer, the feature layer, and the weight rotator. The decoding phase mainly recovers the feature mapping by inverse convolution. We use skip connections between the encoder and decoder to minimize the loss of spatial and semantic information during the downsampling process. In the following we explain each part of RotU-Net in-depth.

A. ENCODER OF ROTU-NET

Our encoder references the network architecture of ResNet50 in the CNN layer. ResNet50 is more likely to learn identity mapping at certain layers, whereas residual networks allow information to flow between layers, including providing feature reuse during forward propagation and mitigating gradient signal vanishing during backpropagation. Figure 2

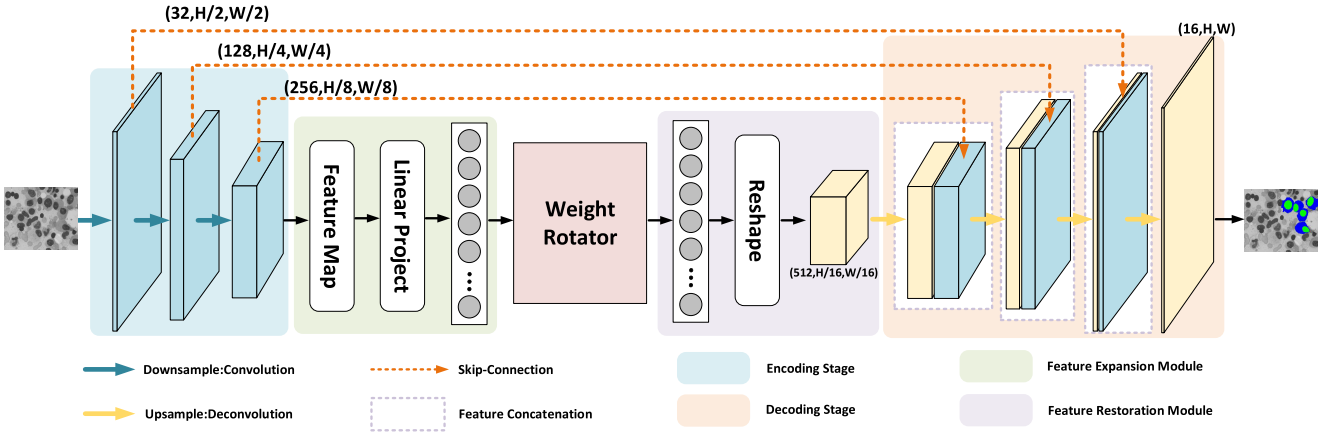


FIGURE 1. Overview of the RotU-Net.

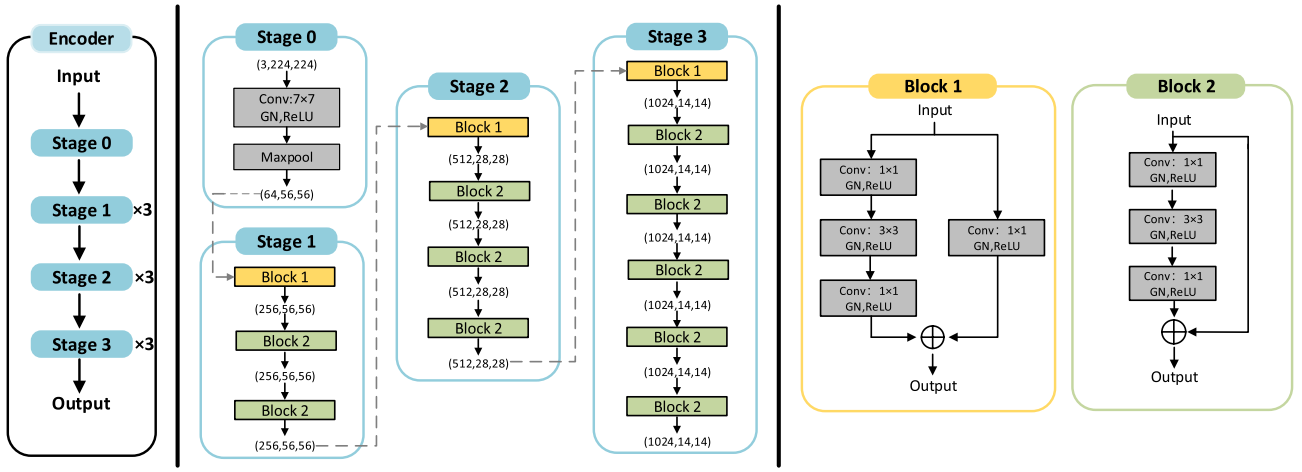


FIGURE 2. Encoder stage and convolutional blocks and of RotU-Net.

shows the encoding stages of RotU-Net, where stage 1, stage 2 and stage 3 are executed three times respectively. We use two kinds of convolutional blocks in the encoding stage, and the most used one is block 2. Block 2 adopts a two-branch structure, and the size of the convolutional kernels of the two branches are 3 and 1 respectively. We believe that different convolutional kernel sizes affect the granularity of local feature extraction, and in order to extract the foreground information more adequately, our model adopts this convolutional block structure.RotU-Net has multiple encoding stages. The size of the input encoder feature maps is $3 \times 224 \times 224$ and the size of the output feature maps is $1024 \times 14 \times 14$. The computation of the feature maps in the encoding stage is demonstrated in Figure 2. It is worth mentioning that the CNN blocks used in our network use group normalization in the normalization part compared to the convolutional blocks of ResNet. This is because the batch normalization of the ResNet can lead to failure of the experiment in case of limited computational resources. Whereas, the group normalization used in our model can be a good solution to the situation where the experimental results are poor because the input batch is too small. The number of groups chosen for group normalization is 32.

B. WEIGHT ROTATOR

As shown in Figure 3, we demonstrate the operation of weight rotator. To explain how it works, we assume that the input feature map of the module is represented as $F \in R^{1\times h\times w}$, where 1 represents the channel of the feature map, h is the length of the feature map and w is the width of the feature map. For capturing and learning the relationships of elements at different positions in the feature map, the weight rotation module takes the 1/4 feature map as the basic unit. The feature map F is divided into four feature blocks:

$$F = \begin{pmatrix} J_1 & J_2 \\ J_3 & J_4 \end{pmatrix} \tag{1}$$

The four feature blocks are represented as $J_i \in R^{1\times \frac{h}{4}\times \frac{w}{4}} (i = 1, 2, 3, 4)$. Specifically, $J_1 = \begin{pmatrix} I_{(0,0)} & I_{(0,1)} \\ I_{(1,0)} & I_{(1,1)} \end{pmatrix}$,

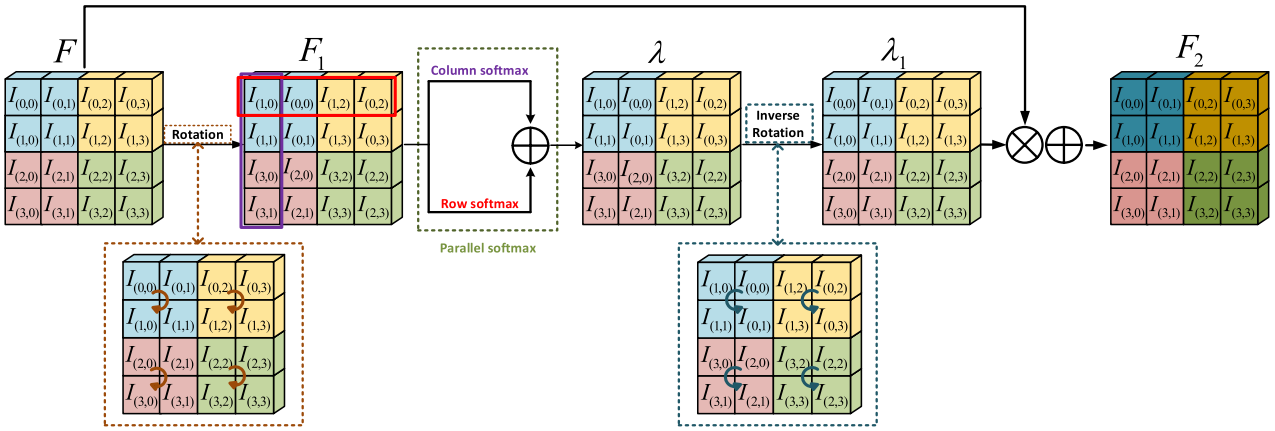


FIGURE 3. Weight Rotator: In order to describe the activation ranges of row softmax and column softmax, we take the first row and column of the feature map as an example and label them with red and purple boxes, and use the softmax for the other rows and columns.

$J_2 = \begin{pmatrix} I_{(0,2)} & I_{(0,3)} \\ I_{(1,2)} & I_{(1,3)} \end{pmatrix}$, $J_3 = \begin{pmatrix} I_{(2,0)} & I_{(2,1)} \\ I_{(3,0)} & I_{(3,1)} \end{pmatrix}$, and $J_4 = \begin{pmatrix} I_{(2,2)} & I_{(2,3)} \\ I_{(3,2)} & I_{(3,3)} \end{pmatrix}$.

Then performe a 90 degree clockwise rotation for each feature block to obtain the feature map F_1:

$$F_1 = Rotation(F) \tag{2}$$

At this point, the four feature blocks of the feature map F_1 are $J_1 = \begin{pmatrix} I_{(1,0)} & I_{(0,0)} \\ I_{(1,1)} & I_{(0,1)} \end{pmatrix}$, $J_2 = \begin{pmatrix} I_{(1,2)} & I_{(0,2)} \\ I_{(1,3)} & I_{(0,3)} \end{pmatrix}$, $J_3 = \begin{pmatrix} I_{(3,0)} & I_{(2,0)} \\ I_{(3,1)} & I_{(2,1)} \end{pmatrix}$, $J_4 = \begin{pmatrix} I_{(3,2)} & I_{(2,2)} \\ I_{(3,3)} & I_{(2,3)} \end{pmatrix}$. In order to enhance the importance of foreground information in the feature map and to strengthen the edge information in the foreground, we activate the rotated feature map F_1 using row softmax and column softmax, respectively, and subsequently add the parallel feature map to obtain the weighted feature map λ, which we define as parallel softmax:

$$\begin{aligned} \lambda &= Parallel\ softmax(F_1) \\ &= Row\ softmax(F_1) + Column\ softmax(F_1) \end{aligned} \tag{3}$$

Afterwards, the weighted feature map λ is rotated 90 degrees counterclockwise back to its original position to obtain the weighted map λ_1:

$$\lambda_1 = Inverse\ Rotation(\lambda) \tag{4}$$

Finally, so as to make model learn the elemental relationship between the weighted feature map and the feature map, the weighted feature map λ_1 is multiplied with the original feature map F and then summed to obtain the final weighted feature map F_2:

$$F_2 = F + F \times \lambda_1 \tag{5}$$

The above describes how weight rotator works. For feature map F, weight rotator breaks its intrinsic spatial structure, allows elements at different positions in the feature map to interact to capture more inter-element semantic information, and enhances the utilization of foreground information by calculating the correlation between some of the elements to enhance the attention of network to foreground regions. At the same time, the model also learns the feature representations under different transformations, which improves network to discriminate foreground edge information.

C. DECODER OF ROTU-NET

The decoding stage in this paper is the same as the TransUNet decoding stage and uses the same number of skip connection. Figure 4 shows the complete computation of the feature map in the decoding stage. The decoding stage uses an extended convolutional kernel to recover the size of the feature map where the size of the convolutional kernel is 3×3. The method used for upsampling in the decoder is the bilinear interpolation algorithm. Meanwhile, in order to minimize the missing important information caused by downsampling, the skip connection in the model concatenate the feature maps from the decoder with the feature maps from the previous layer in the channel. The U-shaped structure of the skip connection combine low-level features and high-level semantic features.

D. LOSS FUNCTION

Due to the limited size of the medical image dataset and the fact that the region to be segmented in the image is only a small part of the entire image. These issues can cause the model to overfit during training. We propose to use Dice loss function and cross-entropy loss function as loss functions to solve these problems. The main role of the Dice loss function is to solve the negative problem caused by the imbalance between foreground and background information in medical images, And it emphasizes foreground information more during the training process. The Dice loss is related to the Dice coefficient, which is used to evaluate the similarity between the label and the predicted value, the higher the Dice coefficient, the higher the similarity is proved. The Dice coefficient is shown as follows:

$$Dice = \frac{2 \times |M \cap N|}{|M| + |N|} \tag{6}$$

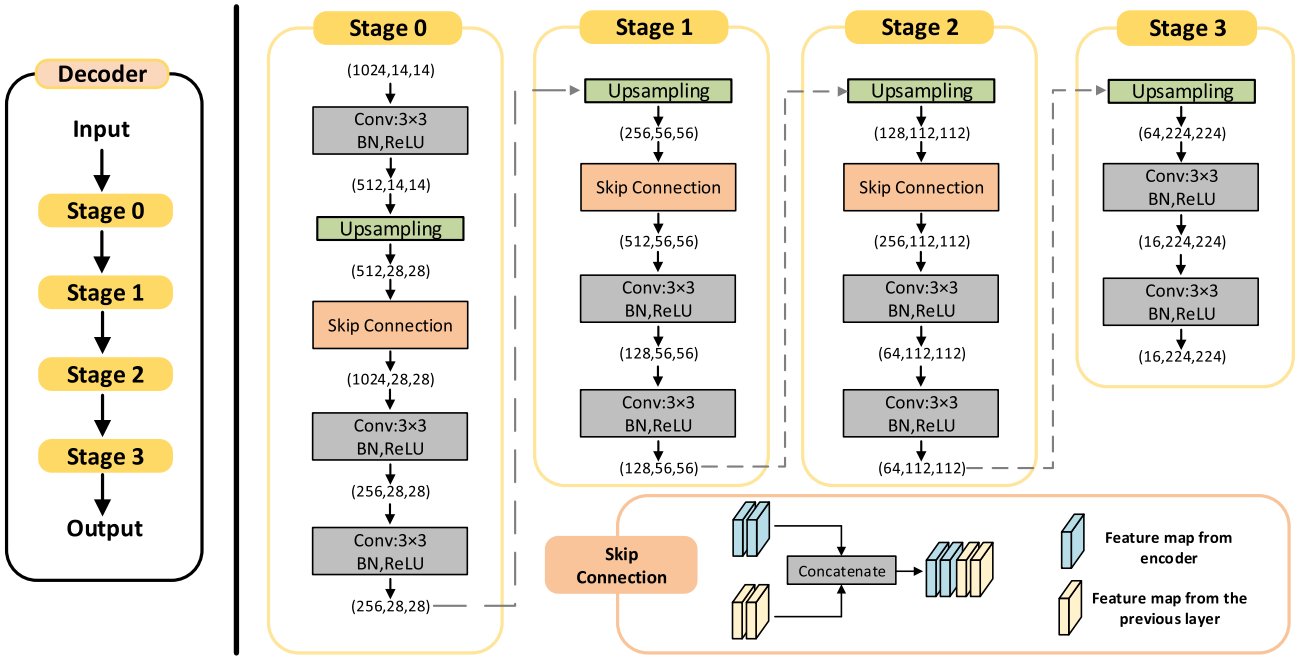


FIGURE 4. Decoder of RotU-Net.

M and N represent two sets, respectively, where $|M \cap N|$ denotes the number of elements intersecting M and N, and $|M|$ and $|N|$ denote the number of elements in the sets M and N, respectively. The Dice loss is calculated as follows:

$$L_{Dice} = 1 - Dice = 1 - \frac{2 \times |M \cap N|}{|M| + |N|} \tag{7}$$

where M and N represent ground truth and predicted value, respectively. Meanwhile, using only the Dice loss function creates the problem of loss saturation. In contrast, cross entropy loss as multiclassification loss function, treats each pixel point equally when calculating pixel loss.

$$L_{CrossEntropy} = -\sum_{x} (\mathrm{p}(x) \log q(x)) \tag{8}$$

where $p(x)$ and $q(x)$ represent ground truth and predicted values respectively. The ground truths are the labels in the datasets and the predicted values are the outputs predicted by the model. A small cross-entropy value indicates that the labeled and predicted values are similar and the model predicts better. To summarize, our loss function combines Dice loss and cross-entropy loss in the training process. In order to optimize both loss functions equally, these two loss functions are each given weight coefficients μ_1 and μ_2, both coefficients are fixed values of 0.5:

$$L_{loss} = \mu_1 \times L_{Dice} + \mu_2 \times L_{CrossEntropy} \tag{9}$$

IV. EXPERIMENTS

A. DATASETS

1) SYNAPSE MULTI-ORGAN SEGMENTATION DATASET (SYNAPSE)

This dataset includes 3779 2D axial abdominal clinical CT images extracted from 30 3D samples [25]. These images contain a varying number of organs, with some containing 8 foreground organs (aorta, gallbladder, left kidney (Kidney(L)), right kidney (Kidney(R)), liver, pancreas, spleen, stomach), some containing 4 foreground organs (aorta, left kidney (Kidney(L)), right kidney (Kidney(R)), liver), and some containing 3 foreground organs (aorta, kidney(L), kidney(R)). foreground organs (aorta, liver, stomach). Overall, the images in this dataset consisted of different categories of organs, respectively. Since the distance between foreground organs in these images varies, some organs are very close to each other and some are far away from each other. Therefore, to accurately segment the various categories of organs in this dataset, it is more important for the network to have both local and global modeling capabilities. Similar to TransUNet, this paper uses 2211 images as the training set and 1568 images as the test set.

2) SEGMENTATION OF MULTIPLE MYELOMA PLASMA CELLS IN MICROSCOPIC IMAGES (SegPC)

This dataset was taken from bone marrow aspirate sections of patients with multiple myeloma, a form of leukemia [46]. The dataset consists of 491 images, each containing several myeloma plasma cells, with segmentation targeting the cytoplasm and nucleus of the cells. We will use 398 images as a training set and the remaining 93 images as a test set. The dataset is available for download and we will provide the download address at the end of the paper.

B. EXPERIMENTAL SETUP

In our experiments, python version is 3.6 and pytorch version is 1.6. the models are trained and tested on two NVIDIA Tesla V100 GPUs. We have no data augmentation of both datasets, and set the image size of the input network to

224 × 224. During training, we use the SGD optimizer, in which momentum is set to 0.9 and the learning rate is set to 0.01. Meanwhile, we set the batchsize to 24 and the epoch to 600.

C. EVALUATION METRICS

Because medical images containing a lot of noise, the background information area tends to be larger than the foreground information area. Even if all foreground information is misclassified as background, the accuracy may still be high. So we use the Dice score (*DSC*) and Hausdorff distance (*HD*) metrics as benchmark metrics to evaluate the network performance. The similarity coefficient is a measure used to evaluate the similarity between sets and is usually quantified by the following formula:

$$DSC = \frac{2 \times |X \cap Y|}{|X| + |Y|} \times 100\% \tag{10}$$

For our experiments, X and Y represent the ground truth and predicted value, respectively.

Hausdorff distance is an evaluation metric for measuring the similarity of two sets in space, and *HD* can be defined as the discrete value obtained by quantifying 95% of the maximum difference between the labeled value and the predicted value. It is calculated as follows

$$HD = \max_{k=95\%} [d(X, Y), d(Y, X)] \tag{11}$$

where X and Y represent the ground truth and predicted value, respectively.

D. RESULT ON SYNAPSE

1) COMPARATIVE EXPERIMENTS ON SYNAPSE

In Table 2, we present experimental results for RotU-Net and other SOTA methods on Synapse. In subsequent sections, we analyze these results in light of various network properties and experimental results.

The liver, as the organ with the largest area and the most edge information among the eight foreground organs, requires the network to be able to model global features correlatively for global feature extraction. Our analysis is proved in the experimental results that the networks dominated by the transformer have better global modeling capabilities, so they achieve higher *DSC* when segmenting the liver. Taking the *DSC* of the main few networks as a comparison, SwinUNet and TransUNet, the networks with transformer as the main architecture, have an average *DSC* of 94.19% in liver segmentation, and U-Net and Att-UNet, the networks with CNN as the main architecture, have an average *DSC* of 93.5% in liver segmentation. The results show that transformer-based networks are more effective in segmenting larger organs such as the liver, which can also prove that transformer has better global modeling capability.

In the segmentation of aorta, the network composed of convolutional kernel has better *DSC*, for example, U-Net has 89.07% *DSC* in the segmentation of aorta, and Att-UNet has up to 89.55% *DSC* in the segmentation of aorta. In contrast, networks with transformer as the main architecture have lower *DSC* in segmenting the aorta, for example, *DSC* of SwinUNet for aortic segmentation is 85.47%, and TransUNet, which performs better, has a *DSC* of only 87.23%.

From the results in the Table 2, it is clear that CNN-based network has better results in segmenting the aorta. When analyzed in relation to specific medical images, this is because the aorta has the smallest area among the eight foreground organs and is located far away from other organs. In order to achieve accurate segmentation of aortic organs, the network needs to extract local features from the image. Compared to transformer-based architectures, CNN-based networks have better local feature extraction capability. Therefore, CNN-based networks perform better in segmenting smaller organs such as the aorta.

The gallbladder organ requires the network to have strong local modeling and organ boundary sensing ability when segmenting, this is because the boundary of the gallbladder organ is closely connected to the boundary of the liver organ. In the comparison experiments in Table 2. The *DSC* of U-Net and Att-UNet are 86.67% and 87.3%, which are better compared to TransUNet in gallbladder organ segmentation. The network proposed in this paper has a *DSC* of 70.51% in gallbladder segmentation, and we analyze that the weight rotator in the network enhances the ability to perceive and discriminate the structure and information of the edges of the foreground region. For pancreas segmentation, its main segmentation difficulty lies in the multiple and complex boundary information caused by the non-smooth surface of the organ. This requires the network to have global modeling capability while also requiring the network to locally model the boundary parts of the organ. Networks architectures containing CNN and transformer have better segmentation results compared to networks with only CNN, the average *DSC* of pancreas for SwinUNet, TransUNet is 56.22%, while the average *DSC* of pancreas for U-Net and Att-UNet is only 56.01%. The network proposed in this paper has a *DSC* of 64.92% in pancreas segmentation, and the comparative results demonstrate that our network adds local modeling without weakening the global modeling for foreground information.

Spleen, kidney (left), kidney (right), stomach and other organs have the same characteristics in medical image segmentation, they are relatively large in area and close to other organs. The above organ distribution characteristics require the network to have the ability of global modeling and local modeling. At the same time, the network is also required to have certain boundary recognition ability. From the experimental data in Table 2, compared with the network with only CNN, the network with both CNN and transformer has better segmentation effect. For example, the average *DSC* of TransUNet for these four foreground organs is 79.9%, while U-Net and Att-UNet are only 77.16% and 78.04%, respectively. The RotU-Net significantly improves the segmentation performance of all four organs with an

TABLE 2. DSC (%) and HD (mm) obtained for each different network architecture and ecological network on the Synapse.

Methods	DSC↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
V-Net [10]	68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [27]	69.77	-	77.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50 U-Net [26]	74.68	36.87	87.74	63.66	80.60	78.19	93.97	56.90	85.87	74.16
U-Net [11]	76.85	39.70	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
R50 Att-UNet [26]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
Att-UNet [47]	77.77	36.02	**89.55**	68.88	77.98	71.11	93.57	58.04	87.3	75.75
R50 ViT [19]	71.29	32.87	73.73	55.13	75.8	72.2	91.51	45.99	81.99	73.95
TransClaw [33]	78.09	26.38	85.87	61.38	84.83	79.36	94.28	57.65	87.74	73.55
MixTrans [37]	78.59	26.59	87.92	64.99	81.47	77.29	93.06	59.46	87.75	76.81
TransUNet [25]	77.48	31.69	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62
UCTransNet [38]	79.11	25.08	88.58	64.34	82.93	75.93	95.42	56.77	88.20	80.67
HiFormer [35]	80.39	**14.70**	86.21	65.69	**85.23**	79.77	94.61	59.52	90.99	81.08
DSGA-Net [39]	81.24	20.91	88.21	**70.87**	82.67	**82.31**	**95.76**	58.49	90.87	80.74
TransUNet++ [41]	80.87	24.79	87.03	66.78	83.38	81.49	94.71	62.49	90.69	80.73
SwinUNet [24]	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
TransDeepLab [40]	80.16	21.25	86.04	69.16	84.08	79.88	93.53	61.19	89.00	78.40
RotU-Net(ours)	**82.15**	26.95	89.03	70.51	82.74	81.79	95.29	**64.92**	**91.09**	**81.85**

average *DSC* of 84.37%, which is an improvement of 4.47% compared to TransUNet. This also proves that our model has strong local modeling and boundary structure identification capabilities while maintaining global modeling.

Figure 5 shows the predicted images of state-of-the-art models in the field of medical image segmentation in recent years. In the visualized image, we can see that the network with CNN has better performance for segmentation of small organs. The network with transformer structure can segment large organs such as liver better because of its global modeling capability, but this also weakens the local modeling and boundary perception capability of model. For example, in the second row of the prediction images, we can see that the network with transformer structure has the problem of misjudging the foreground information and blurring the boundary segmentation. The proposed weight rotator enhances the recognition of the boundary structure because it fully extracts the edge information, thus improving the overall performance of the network.

2) ABLATION EXPERIMENTS ON SYNAPSE

Table 3 shows the *DSC*(%) and *HD*(mm) of the networks row softmax, column softmax and parallel softmax with different dimensional softmax after rotation. We can see that the average score of weight rotator with column softmax is slightly higher than that of weight rotator with row softmax, which is due to the fact that after local rotation, there are more organ edges in the vertical texture of the image, which allows the network to learn more edge information and improve the *DSC*, whereas the weight rotator with parallel softmax can take care of both horizontal and vertical image textures, which allows the network to fully utilize the organ edge information and achieve the highest *DSC* of 82.15%.

Figure 6 shows the predicted images of the ablation experiment on the Synapse dataset. The weight rotator with column softmax is more sensitive to the edges of foreground segmentation than weight rotator with row softmax. For example, in the first and third row prediction images, column softmax delineates more edge structures than row softmax. While parallel softmax, which parallelizes row softmax and column softmax, almost completely distinguishes between foreground and background information, captures and utilizes more local information in the first and second rows of prediction images, reducing the area of false positive segmentation. Meanwhile, in the third and fourth rows, parallel softmax segmented the foreground structures that almost overlapped with the labels, which is due to the fact that the weight rorator of the parallel softmax network performs a two-dimensional softmax, which learns more foreground edge information in between the rows and columns. The visualization of the predicted images proves the effectiveness of weight rotator.

E. RESULT ON SegPC

1) COMPARATIVE EXPERIMENTS ON SegPC

As shown in Table 4, we show the experimental results for different network architectures on the SegPC dataset. We can see in the table that the network segmentation performance of CNN-based architecture is better than that of transformer-based structure. For example, the *DSC* of U-Net and Att-UNet are 79.62% and 78.59%, respectively, while the scores of networks with transformer architectures

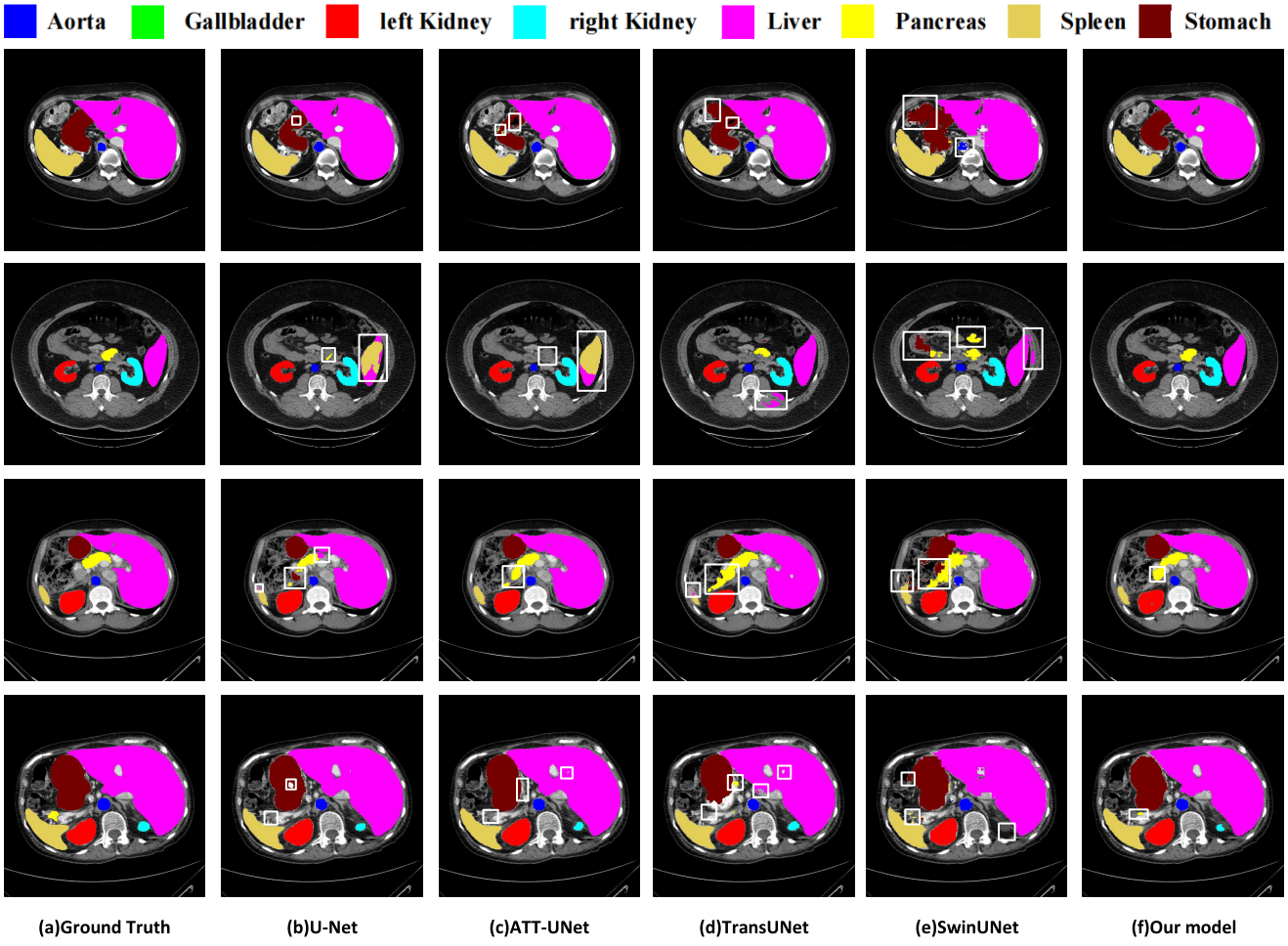


FIGURE 5. Predicted images of different networks on the Synapse dataset. The white rectangular boxes in the figure represent elements that were misclassified as elements in other categories.

TABLE 3. Results of ablation experiments of the weight rotator on Synapse.

Model	DSC↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
Row softmax	81.63	**22.80**	88.28	65.78	**83.97**	**82.29**	**95.60**	**65.00**	90.74	81.38
Column softmax	81.80	24.22	89.01	67.54	82.00	81.69	95.50	64.68	**91.42**	**82.61**
Parallel softmax	**82.15**	26.95	**89.03**	**70.51**	82.74	81.79	95.29	64.92	91.09	81.85

such as TransClaw and TransUNet are only 78.10% and 78.95%. The image of the SegPC dataset consists of two kinds of foregrounds, cytoplasmic and nucleus, which not only have a small in area, but also small in neighboring distance. By analyzing the SegPC dataset in combination with its characteristics, the model is required to provide better local modeling and boundary sensing capabilities if better segmentation is to be achieved. Therefore, the CNN-based network represented by U-Net performs better than the transformer-based network on this dataset. It is worth mentioning that our network achieves optimal segmentation results on the SegPC dataset. We analyze that the weight rorator can better model the foreground information, and at the same time fully utilize the foreground edge information so that the model as a whole can better recover the contour of the foreground.

Figure 7 shows the prediction image of the different networks we obtained on the SegPC. From the prediction images in the second and fourth rows, we can see that the CNN-based network is excellent in local modeling but poor in discriminating the connected foregrounds, which is because the connected foregrounds require the network to extract the foreground edge information in order to sense and discriminate the foreground structure. The weight

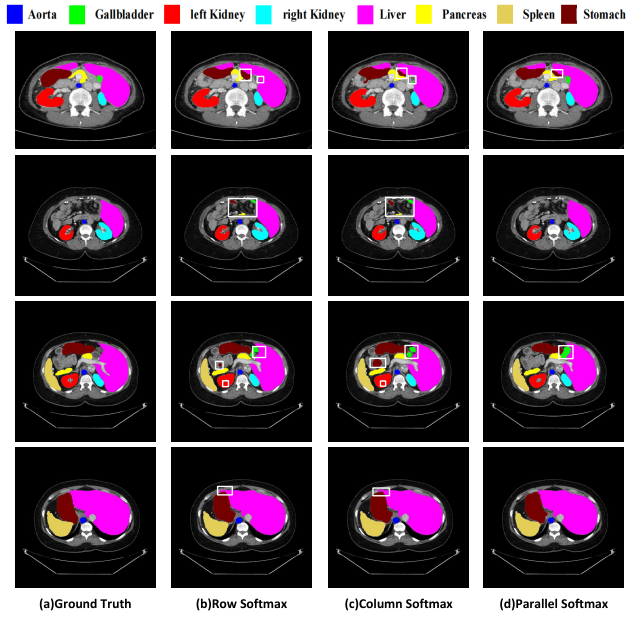


FIGURE 6. Predicted images of ablation experiments on the Synapse dataset. White rectangular boxes represent region elements that were misclassified as elements of other categories.

TABLE 4. DSC(%) and HD (mm) obtained for each different network architecture and ecological network on the SegPC.

Methods	DSC↑	HD↓	Cytoplasm	Nucleus
R50 U-Net [26]	80.55	**32.00**	80.31	80.80
U-Net [11]	79.62	33.93	79.26	79.99
R50 Att-UNet [26]	80.73	33.67	79.48	81.99
Att-UNet [47]	78.59	36.57	78.65	78.52
TransClaw [33]	78.10	35.75	77.12	79.08
MixTrans [37]	79.89	33.02	78.99	80.80
TransUNet [25]	78.95	36.29	78.12	79.78
DAEformer [35]	77.13	32.19	76.65	78.61
TransDeepLab [40]	79.03	35.22	78.37	79.70
RotU-Net(ours)	**82.01**	32.84	**81.46**	**82.57**

rorator proposed in this paper enhances the perception of foreground edges and structures by locally rotating the image without degrading the local modeling capability. So RotU-Net has better performance in segmenting tightly connected foregrounds compared to other network structures. As we can see from the predicted images in the first row, U-Net and R2-UNet do not dominate to the foregrounds at longer distances when segmenting the images, whereas Att-UNet improves the lack of global modeling ability to some extent because of the addition of the attention mechanism. Meanwhile, the models of other transformer-based structures except TransUNet segmented the foregrounds that are far away from each other through global modeling, we analyze that the reason why TransUNet failed to segment the feature points

TABLE 5. Results of ablation experiments of the proposed module on the SegPC.

Model	DSC↑	HD↓	Cytoplasm	Nucleus
Row softmax	81.85	33.11	81.19	82.51
Column softmax	81.69	34.41	80.93	82.45
Parallel softmax	**82.01**	**32.84**	**81.46**	**82.57**

at a long distance is that the convolutional blocks in this network play a greater role in extracting the image features. From the predicted images of different models, our model can better distinguish between the two types of foregrounds, which proves the effectiveness of RotU-Net.

2) ABLATION EXPERIMENTS ON SegPC

As can be seen from Table 5, weight rotator with parallel softmax outperforms weight rotator with column softmax, which in turn outperforms weight rotator with row softmax, while they both improve their performance compared to the existing SOTA models. This proves that weight rotator can indeed improve the original model underutilization of foreground structure and edge information. In addition, compared with the softmax for rows and columns alone, the parallel column softmax and row softmax can further enhance the perception and discrimination of foreground edge information in the net.

As shown in Figure 8, we demonstrate the predicted images of the ablation experiments on the SegPC dataset. Although all three have the error of mistaking background information for foreground information, parallel softmax has the lowest overall error rate. Meanwhile, we can see from the prediction images that parallel softmax has clearer boundaries in segmenting the foreground region compared to row softmax and column softmax, while having smaller misclassification regions. This is due to the local modeling capability of our proposed network as well as the boundary awareness capability.

F. COMPARISON OF NETWORK PARAMETERS

As shown in Table 6, we compare the model parameters and Floating Point Operations (FLOPs) of the proposed RotU-Net with previous SOTA methods. The results show that the RotU-Net has the lowest parameters and the FLOPs reach sub-optimal only after SwinUNet. This demonstrates the excellent performance of the model proposed in this paper despite the minimal number of parameters and operations. It also highlights the effectiveness and innovation of RotU-Net.

G. DISCUSSION

In the task of medical image segmentation, the intricate nature of diverse foreground structural information within the image, coupled with a non-uniform distribution of foreground features. These are the challenges to be faced by segmentation

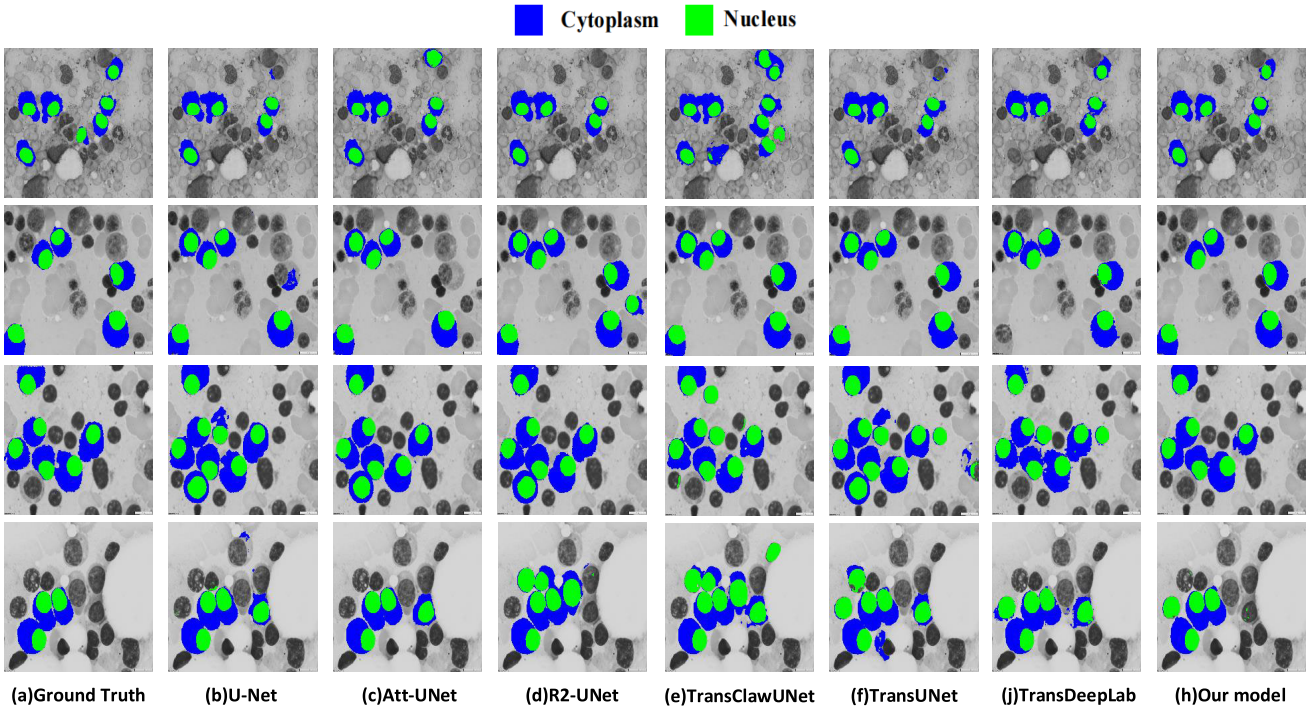


FIGURE 7. Predicted images of different networks on the SegPC dataset. The dataset includes two types of foreground information and one type of background information.

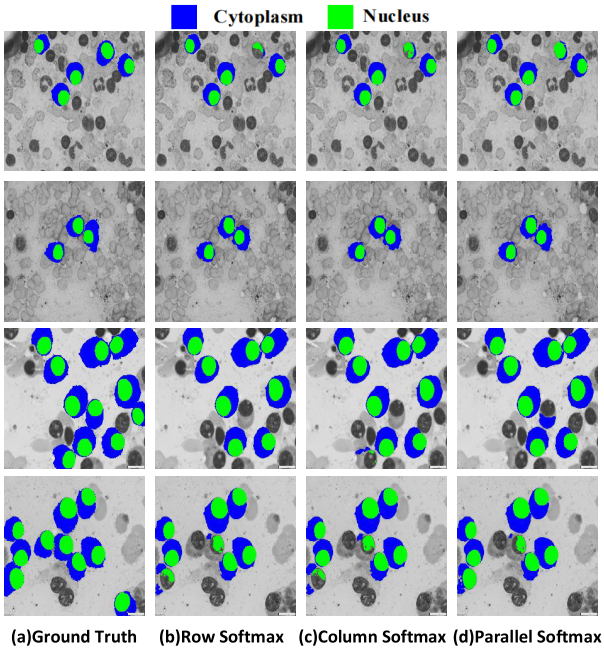


FIGURE 8. Predicted images of ablation experiments on the SegPC.

networks at this stage. Over the recent years, the widespread adoption of CNN-based and transformer-based models in the medical image domain has yielded commendable results, prompting an increasing number of researchers to improve the existing models based on CNN and transformer.

TABLE 6. Parameter (M) and FLOPs (G) between different network architectures.

Methods	Parameter↓	FLOPs↓
U-Net [11]	34.54	50.18
Att-UNet [47]	34.88	51.03
R50 U-Net [26]	85.75	37.03
R50 Att-UNet [26]	87.14	37.71
TransUNet [25]	105.32	24.63
SwinUNet [24]	27.17	**5.92**
TransDeepLab [40]	28.61	17.08
TransClaw [33]	113.02	38.08
MixTrans [37]	79.07	44.75
DAEformer [35]	48.07	26.07
RotU-Net(ours)	**25.54**	9.05

CNN-based networks mainly include U-Net and Att-UNet. U-Net structures concatenate low-level features and high-level features by skip connection, while the lightweight structure has much room for improvement, but the down-sampling process inevitably loses the edge information. Xiao et al. proposed to add an attention module to R50 U-Net, which makes the model more sensitive to foreground regions and more accurate for small target segmentation. Despite this, the fixed-size convolutional kernel imposes restrictions on the

receptive field, hindering optimal performance in multi-organ segmentation tasks. Transformer-based models such as TransUNet and SwinUNet, mainly solve the problem of global modeling. SwinUNet consists entirely of a U-shaped network of transformers, and uses the swin transformer as the basic unit for feature representation and learning remote semantic information, with excellent performance and generalization capabilities. In TransUNet architecture, the transformer encodes the CNN feature map as a contextual sequence, and the decoder upsamples the encoded features, which are then combined with the high-resolution feature map to achieve accurate localization. SwinUNet is a pure transformer network that establishes correlations between the overall elements of the feature map, but at the cost of losing local information in the feature map. This makes SwinUNet difficult for segmentation of small lesions.TransUNet uses a shallow convolutional kernel to extract local information from the feature map, and then uses transformers at deeper layers of the network to extract global information between elements in the feature map. While this allows the network to have both local and global modeling capabilities, the granularity of the local modeling still makes it difficult to discriminate the classes to which the edge information of a lesion area belongs. Therefore, how to utilize the foreground information more effectively and better identify and segment the foreground edges on the basis of local modeling and global modeling has become a problem that needs to be solved. In this paper, we introduce a novel approach by incorporating a weight rotator into the U-shaped network. This innovation leverages the local rotation of the image to disrupt the intrinsic order of the image, enabling features at different locations to interact with each other across distances. By calculating correlations, the model learns additional foreground edge information without compromising the benefits of local and global modeling. Consequently, this enhancement improves the utilization of foreground information and refines the structure and information extraction from the foreground region, enhancing the capacity of model to recognize edge structures and information within the foreground region. Moreover, it significantly bolsters the model robustness. Table 3 and Table 5 in this section demonstrate that RotU-Net has better performance compared to full convolution or transformer combined with convolution. Figure 5 and Figure 7 visualize the prediction images of the state-of-the-art model. As can be seen in the visualization, RotU-Net is more sensitive to the foreground edge information in the image. Also, we have conducted a number of ablation experiments to demonstrate the effectiveness of the proposed module. In summary, the proposed model improves the existing medical image segmentation network to some extent and improves the performance of the current network. However, RotU-Net learns by local rotation of the image, and this local rotation is chosen by human beings, how to make the model automatically learn the correlation relation aspect of different regions is the next direction of our research.

V. CONCLUSION

In this paper, we proposed a new U-shaped network: RotU-Net. The proposed network not only extracts the local information of the features by CNN, but also helps the model to better recognize and capture the boundaries of the objects by weight rotator, and computes the spatial correlation of some important features to improve the perception of the important features. We conducted experiments on Synapse and SegPC datasets and verified the effectiveness of the proposed model. Overall, RotU-Net shows the ability to effectively perceive the edge structure of the foreground region and extract edge information, and implements the computation of important feature correlations. The correlation computation method of weight rotator proposed in this paper may ignore some background information, and our future work will be devoted to achieve better feature correlation computation.

ACKNOWLEDGMENT

(Fuxiang Zhang and Fengchao Wang are co-first author.)

REFERENCES

[1] U. Sehar and M. L. Naseem, "How deep learning is empowering semantic segmentation," *Multimedia Tools Appl.*, vol. 81, no. 21, pp. 30519–30544, Sep. 2022, doi: 10.1007/s11042-022-12821-3.

[2] X. Liu, K. Gao, B. Liu, C. Pan, K. Liang, L. Yan, J. Ma, F. He, S. Zhang, S. Pan, and Y. Yu, "Advances in deep learning-based medical image analysis," *Health Data Sci.*, vol. 2021, Jun. 2021, Art. no. 8786793.

[3] T. Zhou, S. Ruan, and S. Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," 2020, *arXiv:2004.10664*.

[4] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "On the compactness, efficiency, and representation of 3D convolutional networks: Brain parcellation as a pretext task," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, vol. 10265, Boone, NC, USA, Jun. 2017, pp. 348–360.

[5] M. Xian, Y. Zhang, and H. D. Cheng, "Fully automatic segmentation of breast ultrasound images based on breast characteristics in space and frequency domains," *Pattern Recognit.*, vol. 48, no. 2, pp. 485–497, Feb. 2015.

[6] R. Huang, M. Lin, H. Dou, Z. Lin, Q. Ying, X. Jia, W. Xu, Z. Mei, X. Yang, Y. Dong, J. Zhou, and D. Ni, "Boundary-rendering network for breast lesion segmentation in ultrasound images," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102478.

[7] X. Zhang, Z. Xiao, H. Fu, Y. Hu, J. Yuan, Y. Xu, R. Higashita, and J. Liu, "Attention to region: Region-based integration-and-recalibration networks for nuclear cataract classification using AS-OCT images," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102499.

[8] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, Aug. 2019.

[9] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[10] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Stanford, CA, USA, Oct. 2016, pp. 565–571.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, Munich, Germany, Oct. 2015, pp. 234–241.

[12] L. Yu, J. Z. Cheng, Q. Dou, X. Yang, H. Chen, J. Qin, and P. A. Heng, "Automatic 3D cardiovascular MR segmentation with densely-connected volumetric convnets," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 10434, Quebec City, QC, Canada, 2017, pp. 287–295.

[13] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.

[14] Q. Yu, L. Xie, Y. Wang, Y. Zhou, E. K. Fishman, and A. L. Yuille, "Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 8280–8289.

[15] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal CT scans," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 10433, Quebec City, QC, Canada, Sep. 2017, pp. 693–701.

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

[17] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "CE-Net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[20] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2020, *arXiv:2012.12877*.

[21] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3146–3154.

[22] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7794–7803.

[23] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA, 2019, pp. 7354–7363.

[24] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Comput. Vis. ECCV Workshops*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 205–218.

[25] 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," 2021, *arXiv:2102.04306*.

[26] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-UNet for high-quality retina vessel segmentation," in *Proc. 9th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Hangzhou, China, Oct. 2018, pp. 327–331.

[27] S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, and A. Yuille, "Domain adaptive relational reasoning for 3D multi-organ segmentation," in *Proc. 23rd Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, vol. 12261, Lima, Peru, 2020, pp. 656–666.

[28] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Int. Workshop 8th Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, vol. 11045, Granada, Spain, Sep. 2018, pp. 3–11.

[29] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. W. Chen, and J. Wu, "UNet3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 1055–1059.

[30] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation," 2018, *arXiv:1802.06955*.

[31] R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, and D. Merhof, "DAE-Former: Dual attention-guided efficient transformer for medical image segmentation," in *Predictive Intelligence in Medicine* (Lecture Notes in Computer Science), vol. 14277, I. Rekik, E. Adeli, S. H. Park, C. Cintas, and G. Zamzmi, Eds. Cham, Switzerland: Springer, 2023.

[32] S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, and R. Goh, "Medical image segmentation using squeeze-and-expansion transformers," 2021, *arXiv:2105.09511*.

[33] C. Yao, M. Hu, Q. Li, G. Zhai, and X.-P. Zhang, "Transclaw U-Net: Claw U-Net with transformers for medical image segmentation," in *Proc. 5th Int. Conf. Inf. Commun. Signal Process. (ICICSP)*, Shenzhen, China, Nov. 2022, pp. 280–284.

[34] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual Swin Transformer U-Net for medical image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[35] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "HiFormer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2023, pp. 6191–6201.

[36] B. Chen, Y. Liu, Z. Zhang, G. Lu, and A. W. K. Kong, "TransAttUnet: Multi-level attention-guided U-Net with transformer for medical image segmentation," 2021, *arXiv:2107.05274*.

[37] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong, "Mixed transformer U-Net for medical image segmentation," 2021, *arXiv:2111.04734*.

[38] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 2441–2449.

[39] J. Sun, J. Zhao, X. Wu, C. Tang, S. Wang, and Y. Zhang, "DSGA-Net: Deeply separable gated transformer and attention strategy for medical image segmentation network," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 5, May 2023, Art. no. 101553.

[40] R. Azad, M. Heidari, M. Shariatnia, E. K. Aghdam, S. Karimijafarbigloo, E. Adeli, and D. Merhof, "TransDeepLab: Convolution-free transformer-based DeepLab V^{3+} for medical image segmentation," in *Proc. 5th Int. Workshop Predictive Intell. MEdicine*, Singapore, Berlin, Germany: Springer-Verlag, 2022, pp. 91–102.

[41] L. Xu, L. Wang, Y. Li, and A. Du, "Big model and small model : Remote modeling and local information extraction module for medical image segmentation," *Appl. Soft Comput.*, vol. 136, Mar. 2023, Art. no. 110128.

[42] R. Azad, Y. Jia, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Enhancing medical image segmentation with TransCeption: A multi-scale feature fusion approach," 2023, *arXiv:2301.10847*.

[43] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," 2021, *arXiv:2103.03024*.

[44] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," 2021, *arXiv:2102.08005*.

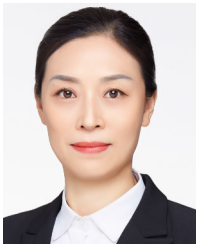
[45] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, "NnFormer: Interleaved transformer for volumetric segmentation," 2021, *arXiv:2109.03201*.

[46] A. Gupta, P. Mallick, O. Sharma, R. Gupta, and R. Duggal, "PCSeg: Color model driven probabilistic multiphase level set based tool for plasma cell segmentation in multiple myeloma," *PLoS One*, vol. 13, no. 12, Dec. 2018, Art. no. e0207908, doi: 10.1371/journal.pone.0207908.

[47] S. Katakis, N. Barotsis, A. Kakotaritis, G. Economou, E. Panagiotopoulos, and G. Panayiotakis, "Automatic extraction of muscle parameters with attention UNet in ultrasonography," *Sensors*, vol. 22, no. 14, p. 5230, Jul. 2022.

FUXIANG ZHANG received the master's degree in internal medicine from Zunyi Medical University, China, in 2013. He is currently with the Intensive Care Medicine Department, The First Affiliated Hospital of Shandong First Medical University, where he is an attending Physician. His research interests include health information and big data analysis, the mechanism of sepsis, and transplantation immunity.

FENGCHAO WANG received the master's degree in microbiology from Liaoning University, in 2022. He is currently a Research Assistant with the Department of Critical Care Medicine, The First Affiliated Hospital of Shandong First Medical University, Jinan, Shandong, China. His research interests include molecular biology, analytical chemistry, and functional nucleic acids.



WENFENG ZHANG received the master's degree in obstetrics and gynecology from Shandong University, in 2007. She is currently the Deputy Chief Physician of the Department of Obstetrics and Gynecology, The First Affiliated Hospital of Shandong First Medical University. Her research interests include gestational diabetes, gestational obesity weight management, and cesarean section beauty suture.

QUANZHEN WANG received the M.D. degree in geriatric medicine from Shandong University, in 2016. She is currently with the Intensive Care Medicine Department, The First Affiliated Hospital of Shandong First Medical University, where she is an attending Physician. Her research interests include the mechanisms of organ damage in sepsis and medical big data analysis.

YAJUN LIU received the with a bachelor's degree in clinical medicine from Binzhou Medical University, in 2022. She is currently pursuing the master's degree with the Department of Clinical Medicine, Shandong First Medical University. Her research interests include the mechanism and treatment of acute respiratory distress syndrome, the occurrence and treatment of sepsis, and medical big data analysis.

ZHIMING JIANG (Member, IEEE) received the M.D. degree in internal medicine from Shandong University, in 2021. He is currently the Director of the Intensive Care Medicine Department, The First Affiliated Hospital of Shandong First Medical University, where he is also a Master Supervisor and a Chief Physician. His research interests include the mechanisms of acute respiratory distress syndrome/acute lung injury and medical big data analysis.