IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 January 2023, accepted 6 February 2023, date of publication 13 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244197

RESEARCH ARTICLE

Medical Image Segmentation Based on Transformer and HarDNet Structures

TONGPING SHEN[1,2], (Member, IEEE), AND HUANQING XU[1,3], (Member, IEEE)

[1]School of Information Engineering, Anhui University of Chinese Medicine, Hefei 230012, China
[2]Graduate School, Angeles University Foundation, Angeles 2009, Philippines
[3]School of Electrical and Information Engineering, Tianjin University, Tianjin 300000, China

Corresponding author: Tongping Shen (shentp2010@ahtcm.edu.cn)

This work was supported in part by the Excellent Young Talents in Anhui Universities Project gxyq2022026, in part by the Anhui Province Quality Engineering Project 2021jyxm0801, and in part by the Natural Science Foundation of Anhui University of Chinese Medicine under Grant 2020zrzd18.

ABSTRACT Medical image segmentation is a crucial way to assist doctors in the accurate diagnosis of diseases. However, the accuracy of medical image segmentation needs further improvement due to the problems of many noisy medical images and the high similarity between background and target regions. The current mainstream image segmentation networks, such as TransUnet, have achieved accurate image segmentation. Still, the encoders of such segmentation networks do not consider the local connection between adjacent chunks and lack the interaction of inter-channel information during the upsampling of the decoder. To address the above problems, this paper proposed a dual-encoder image segmentation network, including HarDNet68 and Transformer branch, which can extract the local features and global feature information of the input image, allowing the segmentation network to learn more image information, thus improving the effectiveness and accuracy of medical segmentation. In this paper, to realize the fusion of image feature information of different dimensions in two stages of encoding and decoding, we propose a feature adaptation fusion module to fuse the channel information of multi-level features and realize the information interaction between channels, and then improve the segmentation network accuracy. The experimental results on CVC-ClinicDB, ETIS-Larib, and COVID-19 CT datasets show that the proposed model performs better in four evaluation metrics, Dice, Iou, Prec, and Sens, and achieves better segmentation results in both internal filling and edge prediction of medical images. Accurate medical image segmentation can assist doctors in making a critical diagnosis of cancerous regions in advance, ensure cancer patients receive timely targeted treatment, and improve their survival quality.

INDEX TERMS Deep learning, medical image segmentation, transformer, HarDNet, feature fusion.

I. INTRODUCTION

Medical images have become an essential source of information for physicians in medical activities such as disease diagnosis, surgery planning, and post-operative evaluation. Medical image segmentation can extract critical information from specific tissue images. The segmented images are provided to physicians for quantitative and qualitative analysis, helping them understand diseased tissues' location and structural characteristics more intuitively and comprehensively to develop better treatment plans and improve treatment results. Clinicians often segment and annotate images manually or semi-manually, which is not only expensive and tedious but also adds a significant burden to the work of clinicians. Moreover, medical images are prone to problems such as unclear edges and inconspicuous contrast, resulting in inaccurate image segmentation results. Image segmentation of lesion regions by deep learning techniques has been the research focus for many years. It has been applied to many organ tissues, such as retinal vessels [1], lung nodules [2], [3], liver tumors [4], brain tumors [5], etc.

The earliest application of deep learning in the field of medical images is the fully convolutional networks (FCN) [6]. FCN leads to serious information loss due to the

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

inconsistent structure of upper and lower sampling layers. With the advent of FCN, Ronneberger proposed the U-Net network's classical image segmentation algorithm [7].

Zhou et al. [8] designed an encoder-decoder network structure UNet ++, which employs an intensely supervised training approach that allows supervised learning of the model's multi-branch output. The model applies more jump connections between high-dimensional and low-dimensional information and reduces the feature error between semantic information. Li et al. [9] proposed a segmentation network using an attention mechanism to obtain multiple dimensions of medical image feature dimensional information. Tang et al. and Manal et al. [10], [11] proposed segmentation networks with dense blocks replacing the convolutional layer in U-Net, improving segmentation accuracy and parameter efficiency. Luo et al. [12] proposed an attention dense Unet (ADUnet), which incorporates dense connections on top of convolutional layers for fine vessel image segmentation.

He et al. [13] replaced the convolutional layer of U-Net with a residual neural network (ResNet). The residual network replaced the jump connection using residual connection paths and codecs. Lu et al. [14] added a circular residual module to the U-Net network to enhance the network's ability to extract feature information for solving pancreatic image segmentation. Gu et al. [15] used a ResNet-34 residual block to replace the original U-Net encoder block as a fixed feature extractor. Liu et al. [16] further optimized the Res-Unet structure by improving the residual modules in encoding and decoding and increasing the number of layers in the segmentation network. Wang et al. [17] and others added residual connections to each layer of convolution in the segmentation network to deepen the network depth and added attention models to improve the segmentation accuracy.

Aamer et al. [18] proposed Attention Residual Unet, an architecture that integrates the residual attention module and the convolutional block of the Unet network. Yu et al. [19] used the ResNet-34 structure for the encoding part of the U-Net structure to deepen the model depth and training speed to improve the segmentation network performance. Hari et al. [20] replaced the encoding part of the U-Net structure with the ResNet-50 structure for automatic detection and segmentation operations of brain tumor images. Cui et al. [21] used a multiscale input structure in the coding layer and a multiscale attention module at the jump connections to effectively improve the accuracy of heart segmentation. Li et al. [22] used a multiscale input module to obtain global feature information of the image.

Almasni et al. [23] and Wang et al. [24] used pyramid pooling modules to fuse global contextual feature information at multiple spatial scales to enhance the detailed representation of the network encoder. Hu et al. [25] used inflated convolution to replace the original convolution and expand the range of convolution layers to improve the model segmentation accuracy and segmentation effect. Ge et al. [26] combined maximum pooling with inflated convolution in the UNet bottleneck layer, intermittently using inflated convolution to obtain image feature information to improve network segmentation. Lan et al. [27] proposed a new image segmentation network that uses lightweight hybrid attention blocks in the encoder to effectively enhance image features and suppress scattered noise in the encoding stage.

Li et al. [9] introduced attention gates between upsampling and downsampling to improve the model segmentation performance by the attention module to merge the image feature information of different hierarchical dimensions. Gu et al. [28] proposed a segmentation network based on an attention mechanism module that focuses on different regions of image feature information. Wang et al. [29] placed the channel attention model in the jump path of the U-Net network to ensure that the segmentation network obtains channel feature information and improves the segmentation effect. Wang et al. [17] added a channel attention mechanism in the jump connection for accurate and efficient medical image segmentation.

Although the improved U-Net network structure can effectively capture the local and global feature information of medical images, the continuous downsampling operation causes the loss of image location information and global dependencies between pixels, which affects the medical image segmentation accuracy [30].

Therefore, medical image segmentation should focus on the extraction of global feature information of images and the fusion of different levels of image feature information in the encoding and decoding stages.

Transformer structure can construct global contextual information, widely used in natural language processing [31]. Dosovitskiy et al. [32] first proposed the Vision Transformer (ViT) structure and applied the Transformer module to the image processing field with good results. Chen et al. [33] combined Transformer and U-Net structures and used them for medical image segmentation aspects. Li et al. [34] proposed the X-Net structure to use the Transformer as an encoder for the backbone segmentation network. Zhang et al. [35] proposed the TransFuse network with a parallel CNN and Transformer feature extraction module designed. Tang et al. [36] proposed a self-supervised pre-trained segmentation model based on the Swin Transformer and proposed the Swin UNETR. Zhou et al. [37] proposed the nnFormer network, where the Transformer and CNN are used alternately in the network. The feature information at each scale is extracted for multi-scale supervised learning to ensure that the multi-scale feature representation is as accurate as possible. Wu et al. [38] proposed the D-Former network, which consists of a Local Scope Module and a Global Scope Module for the cavity transformer.

Luo et al. [39] and Lin et al. [40] used the U-Net network and the Swin Transformer together as the coding part of the backbone network. Hatamizadeh et al. [41] proposed Unetr, which consists of a pure Transformer to form an encoder that

extracts sequences from different layers of the encoder and captures global multiscale features. Li et al. [42] proposed UConvTrans, a two-branch U-shaped network, which splices the output of CNN branches with Transformer branches to achieve an interactive fusion of global and local features. Zhang et al. [43], by fusing two parallel CNN branches and Transformer branches, the images' global dependencies and local detail features can be obtained by a shallower number of layers.

In the above Transformer-based encoding process, there needs to be more information interaction of images within local regions, and geometric features such as lines, edges, and shapes of images are ignored in the process of image slice recombination. In image decoding, the encoded features are upsampled and stitched with high-resolution features without considering the correlation between the two channels and positions.

In the coding stage, the HarDNet68 module is used in this paper. The HarDNet68 module is a harmonic dense connectivity network proposed by Chao et al. in 2019 [44]. Compared with structures such as ResNet and DenseNet [45], HarDNet reduces the number of connections between layers and increases the channel width to improve the model training speed and accuracy, considering that the amount of computation and memory access can affect the model performance.

Considering the problems of Transformer structure, we propose to integrate two modules, HarDNet68 and Transformer, for extracting local and global features of medical images, respectively, and then fuse the two feature information before using them for medical image segmentation. Also, to facilitate the fusion of image features with different dimensions in both the encoding and decoding stages, we propose the feature adaptation fusion module.

We perform model validation on three public datasets, CVC-ClinicDB [46], ETIS-Larib [47], and COVID-19 CT [48] datasets, and analyze them in comparison with other classical segmentation networks.

The main innovations of this paper are as follows.

(1) We adopt the image segmentation network with dual encoders of HarDNet68 and Transformer module, which can extract not only local feature information of low- dimension medical images but also global feature information of high-dimension incremental medical images to improve the model segmentation effect.

(2) We adopt the HarDNet68 module instead of the traditional CNN module in the coding stage. HarDNet68 network can learn more feature information of medical images and reduce the computation, thus improving the operation speed, the segmentation effect, and the accuracy of medical images.

(3) We add a feature information fusion module to the jump connection path of the image segmentation network to fuse the image information of different feature dimensions in the encoding and decoding stages to improve the model segmentation effect.

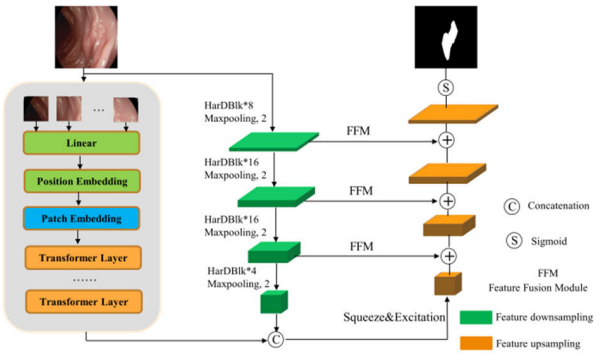


FIGURE 1. Transformer and HardNet network model structure.

FIGURE 2. Vision transformer block.

II. THE PROPOSED ARCHITECTURE

The medical image segmentation network proposed in this paper mainly consists of three parts, HarDNet68, Transformer, and feature information fusion module, and the network structure is shown in Fig. 1. HarDNet68 module is responsible for acquiring local feature information of the input image after multi-layer convolution operation. Transformer module chunks the input image and then acquires global feature information of the medical image. The input image's global and local feature information is directly fused by the summation operation function, allowing the segmentation network to learn the feature information of the input image in multiple dimensions. The fused image feature information is transmitted to the decoder through a simple squeeze and excitation module [49] to activate the effective channels and suppress the useless ones. In the jump path of the image segmentation network, we propose the feature information fusion module to fuse multi-level image feature information of different dimensions to enhance the model expression capability, compensate for the information interaction between channels that the model lacks, enhance the sensitivity of the model to the key information between channels, and thus improve the segmentation network accuracy.

A. TRANSFORMER

Vaswani et al. [31] proposed the Transformer architecture to solve the problem of limited parallel computation during natural language processing. Dosovitskiy et al. [32] proposed the ViT model across domains for computer vision image classification tasks. The Transformer structure in the ViT network is shown in Figure 2.

The self-attention mechanism is an optimization and improvement on the attention mechanism, which mainly relies on the internal correlation of image feature information and assigns different weights to different pixel information.

The self-attention mechanism transforms the input image information into three different matrices of information Q, K, and V. The query matrix Q is multiplied with the transpose of the key value matrix K to derive the similarity matrix QK^T between them, and if the larger value indicates more correlation.

The weight matrix is obtained by 5-normalizing the similarity matrix with the softmax function. Finally the attention of the input matrix is obtained by multiplying the weight matrix with the value matrix, and the final result is shown in equation (1).

$$Attention(Q, K, V) = SoftMax(\frac{QK^T}{\sqrt{d_K}})V \tag{1}$$

where Q, K, and V denote the three matrices obtained after the same input is calculated with different parameters. The moderating smoothing factor $\sqrt{d_K}$ for the k dimensions, the dimension of the key, prevents the multiplication result from being too large. The generated vector is passed through a softmax activation function to normalize the scores and ensure that all the values are positive.

The multi-headed self-attentive mechanism is the core component of Transformer, which is a combination of n self-attentive modules, where W_i^Q, W_i^K, W_i^V denotes the ith self-attentive linear transformation matrix, respectively, which are multiplied with the input vectors X_i to obtain the projections on different spaces to enhance the representational power of the model and obtain the corresponding Q, K, V. Then all the output matrices are stitched together and multiplied with the linear transformation matrix W^o to obtain the final self-attentive output matrix.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

$$Multi(Q, K, V) = Concat(head_i, \cdots, head_h)W^o \tag{3}$$

Multi-layer perceptron: The multi-layer perceptron mainly consists of a linear combination of two fully connected layers and a linear activation layer ReLU, W_1, b_1 and W_2, b_2 denote the weights and biases of the two fully connected layers, respectively, and is calculated as shown in equation (4).

$$MLP(X) = \max(0, XW_1 + b_1)W_2 + b_2 \tag{4}$$

Position embedding: Unlike natural language processing, an image is only a single individual, and in obtaining its self-attentiveness, it is necessary to first segment the whole image into small fixed-sized pieces by convolutional neural network CNN and then splice a learnable position embedding matrix to learn the position embedding information of the image, and then after layer normalization (LN) and multilayer perceptron (MLP), finally, each layer Transformer can be expressed as shown in equation (5) and (6).

$$O'_l = Multi\,(LN\,(O_{l-1})) + O_l \tag{5}$$

$$O_l = MLP\left(LN\left(O'_l\right)\right) + O'_l \tag{6}$$

where O_{l-1} represents the output of the previous Transformer layer. O_l as the input of the next Transformer layer, and so on.

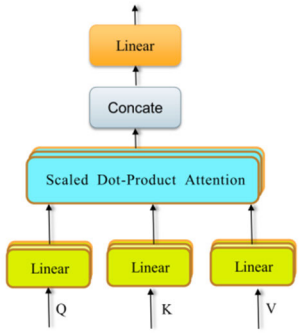


FIGURE 3. Multi-head attention.

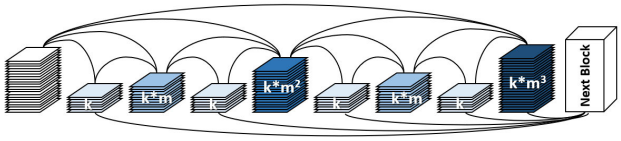


FIGURE 4. HarDNet network structure [44].

Before the medical image is input to the Transform layer, the image needs to be segmented for the operation. We take an input image $X \in \mathbb{R}^{3\times H\times W}$ with H × W resolution and segment it into $N = \frac{H}{P} \times \frac{W}{P}$ blocks for labeling, P is the size of each image segmentation block, and N is the total number of each image segmentation block. Then, the projection function converts the two-dimensional image into a one-dimensional embedded image sequence $X_e \in \mathbb{R}^{N\times C_e}$ with C_e dimension. Finally, the resulting one-dimensional image sequence is transferred to the Transform layer.

B. HardNet

The HarDNet model is an improvement on the DenseNet structure, which can reduce the total number of model parameters and increase the speed of network operation, as shown in Figure 4.

HarDNet uses a sparse connection, assuming that layer k is connected to layer $k-2^n$, 2^n, which can divide k integer, where $n \geq 0, k-2^n \geq 0$. With this connection, if 2^n is processed, layer 1 to layer 2^n-1 can be cleared from memory, reducing the amount of model parameter computation.

We input n image data $x \in \mathbb{R}^{H\times W\times C\times n}$ into a CNN feature extraction network composed of HarDNet, where n denotes the number of images and C denotes the number of channels of images. To realize the transformation of image data $x \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C\times n}$ from high-resolution images to low-resolution images and complete the coarse extraction of image features. At the same time, the segmentation network will pass the feature information of low dimensionality in the image to the upsampling process. Reduce the image features in the encoding and stage process, resulting in information loss.

C. FEATURE FUSION MODULE

The fusion module fuses the outputs of the two branches in the decoder to complete the decoding process from high-level features to segmentation masks. By upsampling, the

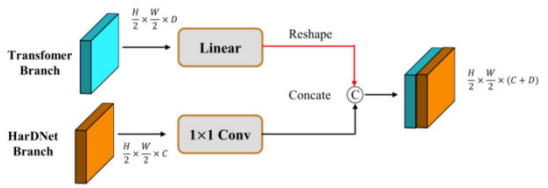


FIGURE 5. Feature fusion structure.

high-level features are spliced with the middle and low-level features of the same scale stored in the hybrid encoder module to prevent partial image detail information during the image upsampling process and ensure the accuracy of the restored image. The output of the Transformer branch first passes through the linear layer, then performs the Reshape operation to save the image feature information. After the output of the HarDNet68 branch undergoes 1 × 1 convolution, it is spliced with the features after the reshape of the Transformer branch in the channel dimension. After a simple extrusion and excitation (SE) module, the effective channel is activated, and the useless channel is uppressed. The element summation operation is directly performed to fuse the feature map, and it is used as the input layer for sampling on the decoding side of the segmentation network. The model structure is shown in Figure 5.

D. LOSS FUNCTION

The loss function is a neural network to measure the degree of loss and error, and it is the index of a neural network to find the optimal weight parameters. Due to the relatively small size of the lesion tissues, the segmentation effect is not satisfactory by directly using the cross-entropy loss function. The loss function adopted in this paper is the combined loss function of Cross Entropy loss [7] and Dice loss [50]. This function can combine the advantages of the two functions to make the network better find the optimal parameters for optimization learning. The cross-entropy loss function evaluates the loss incurred when classifying pixel points in the image segmentation process, and the smaller the value, the better the segmentation model.

$$\mathcal{L}_{\text{Celoss}} = \frac{1}{N}\sum_{i=1} - \sum_{c=1}^{C} y_i \lg(p_i) \tag{7}$$

where C is the label and y_i refers to whether it is category i. If it is that category, $y_i = 1$, otherwise $y_i = 0$.

$\mathcal{L}_{\text{Diceloss}}$ is the loss function of Dice loss. It is mainly used to measure the degree similarity between the segmented image predicted by the model and the real segmented image, and the value range is [0,1]. The calculation formula of the function is shown in Eq. (8). $|X \cap Y|$ represents the number of intersections between a real segmented image and a model predicted image, $|X|$ and $|Y|$ represent the number of real segmented images and model predicted images respectively.

$$\mathcal{L}_{\text{Diceloss}} = 1 - \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{8}$$

The total loss function proposed in this paper is $\mathcal{L}_{\text{Total}}$, The formula is shown in Eq. (9).

$$\mathcal{L}_{\text{Total}} = \frac{\mathcal{L}_{\text{Diceloss}} + \mathcal{L}_{\text{Celoss}}}{2} \tag{9}$$

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATASETS

The CVC-ClinicDB dataset comes from the Clinical Hospital of Barcelona, and consists of 612 polyp images extracted from 31 colonoscopy videos. Each picture in the dataset has GroundTruth manually marked by experts for segmentation experiments in medical image processing. The original size of the image is 383*288.

The ETIS-Larib dataset contains 196 colonoscopic images of 36 polyp types and their corresponding ground-truth segmentation labels, all of which have a resolution of 1225 × 966. Compared with the CVC-ClinicDB dataset, ETIS-Larib has a smaller area of interest and a higher similarity between the target and background features.

The COVID-19 CT contains a series of CT images of lung image segmentation and the corresponding label data, released by Kaggle in 2019. The original image size is 512*512, and the image size is resized to 256*256 as needed.

The above three datasets are divided into the training set, validation set, and test set according to the ratio of 80%, 10%, and 10%.

B. TRAINING AND MEASUREMENT METRICS

The operating environment of this model: CPU main frequency is 3.6GHz, graphics card is RTX2080T, memory is 24G. The operating system is Win 10 Professional, and the deep learning framework is Pytorch for 1.6.

During the network training process, We set the model training count to 200 and the batch size to 6. The network optimizer to Adam's algorithm and the initial learning rate to 0.0001. A Dropout strategy with a ratio of 0.5 is used to prevent the network from overfitting during training.

In this paper, the performance indicators widely used in the medical image segmentation neighborhood are used to evaluate the segmentation results quantitatively: the Dice coefficient, Intersection over Union (Iou), Sensitivity, and Precision. Several evaluation metrics of segmentation are used in the experiments, and the specific definitions of these metrics are given below.

The Dice coefficient can measure the similarity between the predicted segmentation label map and the real segmentation label map, and the larger the value, the higher the similarity between the two.

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \tag{10}$$

The Iou can reflect the overlap rate between the predicted segmentation label map and the real segmentation label map. The larger the value, the more overlap between the two sets

TABLE 1. Comparison of segmentation effects of different methods on CVC-ClinicDB.

Data set	Methods	Dice	Iou	Prec	Sens	Time	Parameters
CVC-Clinic DB	UNet[7]	0.823	0.809	0.882	0.893	~3.1h	7M
	UNet++[8]	0.794	0.811	0.891	0.859	~4.3h	9M
	ResUNet++ [51]	0.899	0.849	0.886	0.873	~2.4h	4M
	Attention-Unet[52]	0.820	0.837	0.904	0.901	~4.9h	11M
	UACANet [53]	0.916	0.870	0.893	0.912	~7.1h	69M
	Ours	**0.931**	**0.892**	**0.916**	**0.924**	~5.3h	19M

TABLE 2. Comparison of segmentation effects of different methods on ETIS-Larib.

Data set	Methods	Dice	Iou	Prec	Sens	Time	Parameters
ETIS-Larib	UNet[7]	0.398	0.335	0.808	0.631	~1.2h	7M
	UNet++[8]	0.401	0.344	0.897	0.679	~2.5h	9M
	ResUNet++ [51]	0.628	0.567	0.865	0.706	~0.8h	4M
	Attention-Unet[52]	0.697	0.617	0.884	0.723	~2.8h	11M
	UACANet [53]	0.694	0.615	0.893	0.739	~4.9h	69M
	Ours	**0.774**	**0.691**	**0.912**	**0.754**	~3.1h	19M

and the higher the similarity.

$$\text{Iou} = \frac{TP}{TP + FP + FN} \tag{11}$$

The precision can reflect whether the pixel consistency between the predicted segmentation label map and the real segmentation label map is strict and accurate. The calculation is the proportion of correctly segmented target pixels to all predicted target class pixels.

$$\text{Prec} = \frac{TP}{TP + FP} \tag{12}$$

The sensitivity calculates the proportion of correctly segmented target pixels to the target class pixels in the real segmentation label map. The larger the value, the lower the proportion of missed detection.

$$\text{Sens} = \frac{TP}{TP + FN} \tag{13}$$

where TP is the pixels correctly segmented in the medical segmentation results, TN is the pixels incorrectly segmented in the medical segmentation results, FP is the background pixels incorrectly treated as medical pixels in the medical segmentation results, and FN is the medical pixels incorrectly treated as background pixels in the segmentation results.

C. ANALYSIS OF EXPERIMENTAL RESULTS

The segmentation results on the CVC-ClinicDB, ETIS-Larib and COVID-19 CT datasets were compared with those of Unet, PraNet, UNet++, Attention-Unet, PraNet, UACANet networks, respectively. Table 1 to Table 3 shows the final obtained methods of each of the segmentation performance metrics.

From Table 1, for the CVC-ClinicDB dataset, the Dice, Iou, Prec, and Sens of the U-Net network are 0.823, 0.809, 0.882, and 0.893, respectively. Compared with the U-Net, our model proposed in this paper increased by 10.8%, 8.3%, 3.4%, and 3.1% on Dice, Iou, Prec, and Sens, respectively.

From Table 2, for the ETIS-Larib dataset, the Dice, Iou, Prec, and Sens of the U-Net network are 0.398, 0.335, 0.808, and 0.631, respectively. Compared with the UACANet, our model proposed in this paper increased by 8.0%, 7.6%, 1.9%, and 1.5% on Dice, Iou, Prec, and Sens, respectively.

From Table 3, for the COVID-19 CT dataset, the Dice, Iou, Prec, and Sens of the U-Net network are 0.798, 0.825, 0.967, and 0.949, respectively. Compared with U-Net, our model proposed in this paper increased by 15.5%, 14.9%, 3.1%, and 4.7% on Dice, Iou, Prec, and Sens, respectively.

We also compare the proposed method in this paper with several recently proposed segmentation methods for analysis. Table 4 shows the performance of different segmentation methods on the CVC-ClinicDB, ETIS-Larib, and COVID-19 CT datasets. The overall performance of the model in this paper is comparable to the performance of the latest model. The models achieved relatively good results on all three datasets.

Our model proposed it has achieved excellent segmentation results in several evaluation indexes on the three datasets.

The segmentation experiments are carried out on CVC-ClinicDB, ETIS-Larib, and COVID-19 CT datasets and compared with the Unet, UNet++, ResUNet++, Attention-Unet, and UACANet networks. The paper runs the six comparison networks in the same experimental environment, and their visual effects are shown in Figure 6 to Figure 8.

Figure 6 shows the segmentation effect of various networks on the CVC-ClinicDB dataset. The third and eighth columns are the result diagram of six network segmentations. From

TABLE 3. Comparison of segmentation effects of different methods on COVID-19 CT.

Data set	Methods	Dice	Iou	Prec	Sens	Time	Parameters
COVID-19 CT	UNet[7]	0.798	0.825	0.967	0.949	~1.1h	7M
	UNet++[8]	0.801	0.834	0.989	0.992	~1.4h	9M
	ResUNet++ [51]	0.828	0.867	0.986	0.996	~0.6h	4M
	Attention-Unet[52]	0.897	0.911	0.990	**0.997**	~1.7h	11M
	UACANet [53]	0.894	0.935	0.995	0.993	~3.9h	69M
	Ours	**0.953**	**0.974**	**0.998**	0.996	~2h	19M

TABLE 4. Comparisons against existing approaches on CVC-ClinicDB, ETIS-Larib and COVID-19 CT.

Datasets	Methods	DICE	IOU	PRC	SENS
CVC-ClinicDB	FCBFormer[54]	0.947	0.902	**0.952**	**0.944**
	DuAT[55]	0.948	0.906	-	-
	ESFPNet[56]	**0.949**	**0.907**	-	-
	Ours	0.931	0.892	0.916	0.924
ETIS-Larib	PVT-CASCADE[57]	0.801	0.726	-	-
	DuAT[55]	0.822	0.74	-	-
	ESFPNet	**0.823**	**0.748**	-	-
	Ours	0.774	0.691	**0.912**	**0.754**
COVID-19 CT	BCDU-Net[58]	0.979	0.948	0.975	**0.998**
	R2U-Net[59]	0.943	**0.975**	0.973	0.983
	MDA-Net[60]	**0.986**	0.954	**0.986**	-
	Ours	0.953	0.974	0.998	0.996

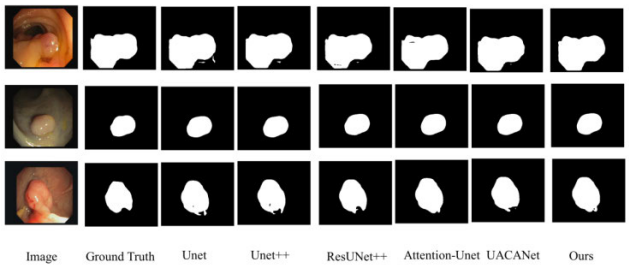


FIGURE 6. Model segmentation results in the CVC-ClinicDB dataset.

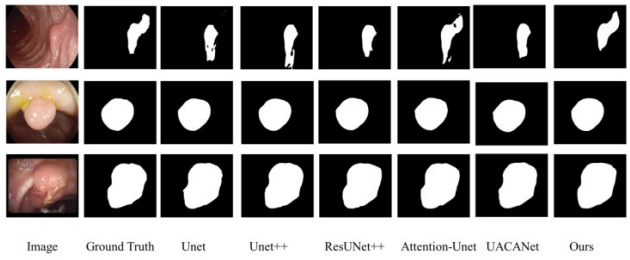


FIGURE 7. Model segmentation results in the ETIS-Larib dataset.

Figure 6, the proposed algorithm can completely distinguish lesion regions with blurred boundaries, while other algorithms have some omissions for segmentation targets with blurred edges. The visualization results can well overcome the problem of similar color polyps and backgrounds, detect polyp tissue with different shapes, sizes, and colors, and divide the region and boundary more clearly and accurately without missing phenomena.

Figure 7 shows the segmentation effect of various networks on the ETIS-Larib dataset. The third and eighth columns are the resulting diagram of six network segmentations. From Figure 7, several networks can segment the edge information of lesion regions images, but our model proposed is better than other networks in dealing with the edge part. The segmentation boundary is clearer, the structure is relatively complete, and it achieves the best segmentation performance.

Through Figure 6 and Figure 7, we can also find that all the neural network models segment well for polyps with relatively smooth edges. However, the segmentation results in the first row in Figure 5 shows that for polyps with similar backgrounds, the segmentation results of other models could be better compared with the model in this paper. There are problems such as image edge loss.

The method in this paper has almost no problem with missing polyp segmentation, less miss-segmented areas, clearer edge contours, and better internal coherence of the segmented image.

Figure 8 shows the segmentation effect of various networks on the COVID-19 CT data set. The third and eighth columns are the resulting diagram of six network segmentations. From Figure 8, the U-Net has learned too many redundant features. There are always obvious noise points; several other networks

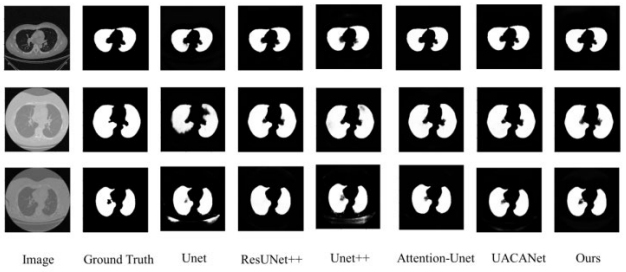


FIGURE 8. Model segmentation results in the COVID-19 CT dataset.

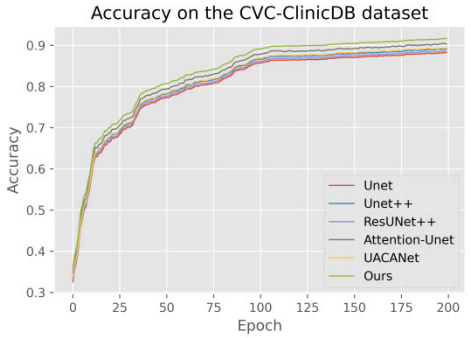


FIGURE 9. Accuracy of the CVC-ClinicDB dataset.

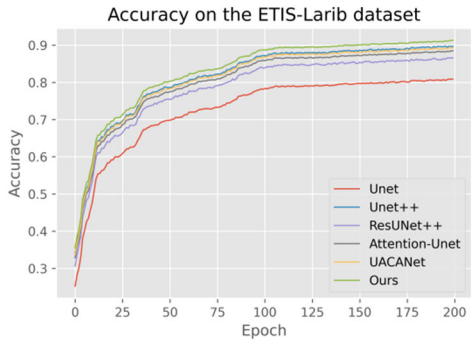


FIGURE 10. Accuracy of the ETIS-Larib dataset.

also have good segmentation performance on the segmentation boundary, but it pays too much attention to the image boundary, thus ignoring the internal features of the image. However, our model proposed in this paper retains more image details, and the segmentation results are consistent with the standard segmented images.

We also compare the accuracy and loss of the different models on the three datasets, as shown in Figure 9 to 11.

The comparative analysis of the three figures shows that the model proposed in our paper converges fast on the three datasets. Finally, the model is almost converged and achieved a high accuracy rate.

IV. DISCUSSION

In recent years, many experts have applied the Transform module to the medical image field and the mainstream

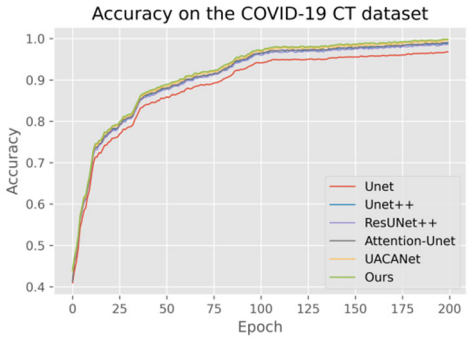


FIGURE 11. Accuracy of the COVID-19 CT dataset.

segmentation network collection to improve the effectiveness of medical image segmentation, such as the TransUnet network. However, the network model based on the Transformer structure tends to ignore geometric features such as lines, edges, and shapes of the image during the encoding and decoding stages and cannot consider the correlation information between channels and positions. Therefore, we propose a dual module of HarDNet68 and Transformer for simultaneous image feature information extraction. The HarDNet68 module improves the DenseNet network structure, which runs faster and is directly used for local feature information extraction of medical images. The Transformer module can consider global information and is used to extract global feature information of medical images. Finally, the global and local feature information of medical images are fused and then transmitted to the next layer of the medical segmentation network to improve the segmentation performance of the network.

It can be seen from the objective evaluation index, and the segmentation effect diagram that the network proposed in this paper has achieved the best performance in the evaluation indexes of Dice, Iou, Prec, and Sens. And also is better than the comparison method in the text in terms of visual effect. We propose a feature adaptation module to facilitate the fusion of image features at the encoding and decoding stages. By introducing a simple Squeeze and Excite module, activate effective channels and suppress useless channels, directly perform element-wise summation operations to fuse feature maps of corresponding layers in local and global contexts. Fuse channels of multi-level features information, enhance the expressive ability of the model, make up for the information interaction between channels that the model lacks, and enhance the model's sensitivity to the key information between channels, thereby improving the accuracy of the segmentation network.

Although we have extensively evaluated the network's performance on three different datasets, CVC-ClinicDB, ETIS-Larib, and COVID-19 CT, we considered two different types of image datasets of polyps and lungs for model validation, but our network still needs to improve. First of all, due to objective factors, we did not try to verify the influence of the connection mode of different attention modules on the network structure; secondly, the images used in the experiment were all 2D images, and we did not try to verify it on the 3D medical image segmentation dataset. We may do the above work in the future.

V. CONCLUSION

Since the Transform module in the existing medical image segmentation network does not consider the local connection between adjacent blocks, and the lack of interaction between channel information in the upsampling process, we propose a dual module of HarDNet68 and Transformer for simultaneous image feature information extraction. The HarDNet68 module is an improvement on the DenseNet network structure, which runs faster and is directly used for local feature information extraction of medical images. The Transformer module can consider global information and is used to extract global feature information of medical images. To realize the fusion of image feature information of different dimensions in two stages of encoding and decoding, a feature adaptation fusion module is proposed to fuse the channel information of multi-level features and realize the information interaction between channels and then improve the segmentation network accuracy. On CVC-ClinicDB, ETIS-Larib, and COVID-19 CT datasets, the Dice of the method in this paper reached 0.931, 0.774, and 0.953, respectively, and the IoU reached 0.892, 0.691 and 0.974, respectively. The segmentation effect was better than that of the comparison method. In future work, we will further optimize the model structure, apply it to more medical datasets, and improve the generalization performance of the network model.

REFERENCES

[1] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su, "DUNet: A deformable network for retinal vessel segmentation," *Knowl.-Based Syst.*, vol. 178, pp. 149–162, Aug. 2019.

[2] Q. Zhang, M. Chen, and Y. Qin, "Lung nodule segmentation based on 3D ResUNet network," *Chin. J. Med. Phys.*, vol. 36, no. 11, pp. 1356–1361, 2019.

[3] S. H. Zhong, M. L. Wang, X. M. Guo, Y. Zhang, and Y. N. Zheng, "Study on the improved VNet network based pulmonary nodule segmentation method," *Chin. J. Sci. Instrum.*, vol. 41, no. 9, pp. 206–215, 2020.

[4] S. Li, G. K. F. Tso, and K. He, "Bottleneck feature supervised U-Net for pixel-wise liver and tumor segmentation," *Expert Syst. Appl.*, vol. 145, May 2020, Art. no. 113131.

[5] X. Feng, N. J. Tustison, S. H. Patel, and C. H. Meyer, "Brain tumor segmentation using an ensemble of 3D U-Nets and overall survival prediction using radiomic features," *Frontiers Comput. Neurosci.*, vol. 14, p. 25, Apr. 2020.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Munich, Germany: Springer, 2015, pp. 234–241.

[8] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2019.

[9] C. Li, Y. Tan, W. Chen, X. Luo, Y. He, Y. Gao, and F. Li, "ANU-Net: Attention-based nested U-Net to exploit full resolution features for medical image segmentation," *Comput. Graph.*, vol. 90, pp. 11–20, Aug. 2020.

[10] P. Tang, C. Zu, M. Hong, R. Yan, X. Peng, J. Xiao, X. Wu, J. Zhou, L. Zhou, and Y. Wang, "DA-DSUnet: Dual attention-based dense SU-Net for automatic head-and-neck tumor segmentation in MRI images," *Neurocomputing*, vol. 435, pp. 103–113, May 2021.

[11] M. AlGhamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, "DU-Net: Convolutional network for the detection of arterial calcifications in mammograms," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3240–3249, Oct. 2020.

[12] Z. Luo, Y. Zhang, L. Zhou, B. Zhang, J. Luo, and H. Wu, "Micro-vessel image segmentation based on the AD-UNet model," *IEEE Access*, vol. 7, pp. 143402–143411, 2019.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[14] L. Lu, L. Jian, J. Luo, and B. Xiao, "Pancreatic segmentation via ringed residual U-Net," *IEEE Access*, vol. 7, pp. 172871–172878, 2019.

[15] 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.

[16] S. Liu, Y. Li, J. Zhou, J. Hu, N. Chen, Y. Shang, Z. Chen, and T. Li, "Segmenting nailfold capillaries using an improved U-Net network," *Microvascular Res.*, vol. 130, Jul. 2020, Art. no. 104011.

[17] Z. Wang, Y. Zou, and P. X. Liu, "Hybrid dilation and attention residual U-Net for medical image segmentation," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104449.

[18] A. A. Rahman, B. Biswal, P. G. Pavani, S. Hasan, and M. V. S. Sairam, "Robust segmentation of vascular network using deeply cascaded AReN-UNet," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102953.

[19] S. Yu, D. Xiao, S. Frost, and Y. Kanagasingam, "Robust optic disc and cup segmentation with deep learning for glaucoma detection," *Comput. Med. Imag. Graph.*, vol. 74, pp. 61–71, Jun. 2019.

[20] H. M. Rai, K. Chatterjee, and S. Dashkevich, "Automatic and accurate abnormality detection from brain MR images using a novel hybrid UnetResNext-50 deep CNN model," *Biomed. Signal Process. Control*, vol. 66, Apr. 2021, Art. no. 102477.

[21] H. Cui, C. Yuwen, L. Jiang, Y. Xia, and Y. Zhang, "Multiscale attention guided U-Net architecture for cardiac segmentation in short-axis MRI images," *Comput. Methods Programs Biomed.*, vol. 206, Jul. 2021, Art. no. 106142.

[22] F. Li, W. Li, S. Qin, and L. Wang, "MDFA-Net: Multiscale dual-path feature aggregation network for cardiac segmentation on multi-sequence cardiac MR," *Knowl.-Based Syst.*, vol. 215, Mar. 2021, Art. no. 106776.

[23] M. A. Al-Masni and D.-H. Kim, "CMM-Net: Contextual multi-scale multi-level network for efficient biomedical image segmentation," *Sci. Rep.*, vol. 11, no. 1, p. 10191, May 2021.

[24] D. Wang, G. Hu, and C. Lyu, "Multi-path connected network for medical image segmentation," *J. Vis. Commun. Image Represent.*, vol. 71, Aug. 2020, Art. no. 102852.

[25] X. Hu and H. Wang, "Efficient fast semantic segmentation using continuous shuffle dilated convolutions," *IEEE Access*, vol. 8, pp. 70913–70924, 2020.

[26] R. Ge, H. Cai, X. Yuan, F. Qin, Y. Huang, P. Wang, and L. Lyu, "MD-UNET: Multi-input dilated U-shape neural network for segmentation of bladder cancer," *Comput. Biol. Chem.*, vol. 93, Aug. 2021, Art. no. 107510.

[27] Y. Lan and X. Zhang, "Real-time ultrasound image despeckling using mixed-attention mechanism based residual UNet," *IEEE Access*, vol. 8, pp. 195327–195340, 2020.

[28] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Feb. 2021.

[29] Y. L. Wang, Z. J. Zhao, S. Y. Hu, and F. L. Chang, "CLCU-Net: Cross-level connected U-shaped network with selective feature aggregation attention module for brain tumor segmentation," *Comput. Methods Programs Biomed.*, vol. 207, Aug. 2021, Art. no. 106154.

[30] P. Qin, C. Li, J. Chen, and R. Chai, "Research on improved algorithm of object detection based on feature pyramid," *Multimedia Tools Appl.*, vol. 78, no. 1, pp. 913–927, Jan. 2019.

[31] 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.*, 2017, p. 30.

[32] 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," in *Proc. Int. Conf. Learn. Represent.*, Vienna, Austria, 2020, pp. 1–22.

[33] 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*.

[34] Y. Li, Z. Wang, L. Yin, Z. Zhu, G. Qi, and Y. Liu, "X-Net: A dual encoding-decoding method in medical image segmentation," *Vis. Comput.*, vol. 37, pp. 1–11, Nov. 2021, doi: 10.1007/s00371-021-02328-7.

[35] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Strasbourg, France, 2021, pp. 14–24.

[36] Y. Tang, D. Yang, W. Li, H. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of swin transformers for 3D medical image analysis," 2021, *arXiv:2111.14791*.

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

[38] Y. Wu, K. Liao, J. Chen, J. Wang, D. Z. Chen, H. Gao, and J. Wu, "D-former: A U-shaped dilated transformer for 3D medical image segmentation," 2022, *arXiv:2201.00462*.

[39] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between CNN and transformer," 2021, *arXiv:2112.04894*.

[40] A. Lin, B. Chen, J. Xu, Z. Zhang, and G. Lu, "DS-TransUNet: Dual swin transformer U-Net for medical image segmentation," 2021, *arXiv:2106.06716*.

[41] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2022, pp. 574–584.

[42] Q. Li, Y. B. Huangpu, J. Y. Jiang, Z. F. Yang, P. Chen, and Z. H. Wang, "UConvTrans: A dual-flow cardiac image segmentation network by global and local information integration," *J. Shanghai Jiaotong Univ.*, pp. 1–12, Sep. 2022.

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

[44] P. Chao, C.-Y. Kao, Y. Ruan, C.-H. Huang, and Y.-L. Lin, "HarDNet: A low memory traffic network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3552–3561.

[45] H. Gao, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 4700–4708.

[46] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, Jul. 2015.

[47] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, 2014.

[48] J. Ma, Y. Wang, X. An, C. Ge, Z. Yu, J. Chen, Q. Zhu, G. Dong, J. He, Z. He, T. Cao, Y. Zhu, Z. Nie, and X. Yang, "Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation," *Med. Phys.*, vol. 48, no. 3, pp. 1197–1210, Mar. 2021.

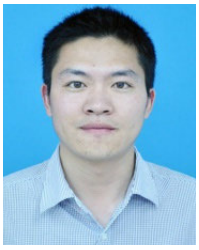
[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[50] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, Apr. 2022.

[51] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. D. Lange, P. Halvorsen, and H. D. Johansen, "ResUNet++: An advanced architecture for medical image segmentation," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, San Diego, CA, USA, Dec. 2019, pp. 225–255.

[52] Y. Sun, F. Bi, Y. Gao, L. Chen, and S. Feng, "A multi-attention UNet for semantic segmentation in remote sensing images," *Symmetry*, vol. 14, no. 5, p. 906, Apr. 2022.

[53] T. Kim, H. Lee, and D. Kim, "UACANet: Uncertainty augmented context attention for polyp segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2021, pp. 2167–2175.
[54] E. Sanderson and J. M. Bogdan, "FCN-transformer feature fusion for polyp segmentation," in *Proc. Annu. Conf. Med. Image Understand. Anal.*, Cambridge, U.K., 2022, pp. 892–907.
[55] F. Tang, Q. Huang, J. Wang, X. Hou, J. Su, and J. Liu, "DuAT: Dual-aggregation transformer network for medical image segmentation," 2022, *arXiv:2212.11677*.
[56] Q. Chang, D. Ahmad, J. Toth, R. Bascom, and W. E. Higgins, "ESFPNet: Efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video," 2022, *arXiv:2207.07759*.
[57] M. M. Rahman and R. Marculescu, "Medical image segmentation via cascaded attention decoding," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2023, pp. 6222–6231.
[58] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM U-Net with Densley connected convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 1–10.
[59] M. Z. Alom, C. Yakopcic, T. M. Taha, and V. K. Asari, "Nuclei segmentation with recurrent residual convolutional neural networks based U-Net (R2U-Net)," in *Proc. IEEE Nat. Aerosp. Electron. Conf.*, New York, NY, USA, Jul. 2018, pp. 228–233.
[60] X. G. Peng and D. L. Peng, "MDA-Net: A medical image segmentation network combining dual-path attention mechanism," *Small Microcomput. Syst.*, pp. 1–9, Jul. 2022.



TONGPING SHEN (Member, IEEE) received the B.S. degree in information management and information systems from the Anhui University of Chinese Medicine, Hefei, China, in 2007, and the M.S. degree in intelligence from Anhui University, Hefei, in 2010. He is currently pursuing the Ph.D. degree in information technology with Angeles University Foundation, Angeles, Philippines.

He is currently an Associate Professor with the School of Pharmaceutical Information Engineering, Anhui University of Traditional Chinese Medicine. He mainly researches on Chinese medicine informatization, publishes more than 30 papers, authorizes four invention patents, and presides over several scientific research projects.

HUANQING XU (Member, IEEE) received the Ph.D. degree from Tianjin University. Currently, he is a Teacher in medical information engineering with the Anhui University of Traditional Chinese Medicine. His research interests include medical image segmentation, machine learning, and deep learning.