ACMS-TransNet: Polyp Segmentation Network Based on Adaptive Convolution and Multi-Scale Global Context

Ping Sun, Jiansheng Wu, Zixue Zhao and Han Gao

***Abstract*—Accurate segmentation of colorectal polyps is critical for the early diagnosis and treatment of colorectal cancer. With the advancement of computer vision technologies, the U-Net framework, characterized by its encoder-decoder architecture, has been widely applied to medical image segmentation tasks. However, it still has several limitations. Traditional convolution operations exhibit constraints in modeling spatial and channel features, making it challenging to detect small polyp targets effectively. Additionally, the use of simple skip connections between the encoder and decoder lacks effective modeling of global multi-scale contextual information, resulting in difficulties in fusing multi-scale feature information efficiently. In order to solve these problems, this study proposes a novel model, ACMS-TransNet. The model introduces an Adaptive Convolution Block at critical levels of the encoder and decoder, dynamically adjusting the receptive fields of convolution kernels to accommodate features at different scales, thereby enhancing small object detection capabilities. Additionally, it incorporates a skip connection design combining the MHCIA module and the SEMS-FFN module. By exploring multi-scale global contextual information, this design establishes stronger associations between the encoder and decoder. The MHCIA module facilitates inter-channel information interaction through a multi-head attention mechanism, improving the network's ability to capture global contextual information. Meanwhile, the SEMS-FFN module integrates four parallel deep convolutions at different scales with saliency feature extraction techniques, effectively capturing multi-scale information during feature fusion and enhancing the network's capability to extract features across various scales. Experimental results demonstrate that the ACMS-TransNet model achieved a Dice coefficient of 90.46% and an IoU of 83.69% on the Kvasir-SEG dataset, and a Dice coefficient of 94.45% and an IoU of 89.68% on the CVC-ClinicDB dataset. These findings validate the efficiency and accuracy of the proposed model in colorectal polyp segmentation tasks, providing robust technical support for the early detection and treatment of colorectal cancer.**

***Index Terms*—Deep Learning, Polyp Segmentation, Attention Mechanism, Adaptive Convolution, Multi-Scale Features.**

I. Introduction

COLON polyps are important precursors to colon cancer, and early detection and removal of these polyps are key measures in preventing colon cancer [1]. Colonoscopy is currently the most commonly used diagnostic method; however, it relies heavily on the experience of the doctor, which can lead to missed diagnoses and misdiagnoses. As a result, automated colon polyp segmentation technology has become a research focus. The goal is to use computer-aided diagnosis systems to accurately segment colon polyps, which can significantly improve diagnostic efficiency and accuracy, reduce the workload of doctors, and enhance treatment outcomes and survival rates for patients. Polyp segmentation is mainly to accurately identify and isolate polyp areas from colonoscopy images. However, polyp segmentation is challenging due to the variety of shapes, sizes, colors and textures. In recent years, with the rapid development of deep learning technology [2], U-Net [3] based medical image segmentation methods have achieved significant results in the task of colon and rectal polyp segmentation. However, existing methods still have limitations in handling small target lesions, complex boundaries, and insufficient feature expression. On the one hand, traditional convolutional operations have limited modeling capabilities for spatial and channel features, making it difficult to effectively handle complex medical image features, and small polyp targets are easily overlooked by the network. On the other hand, single-scale features are unable to fully express global context information, affecting the robustness of segmentation. In addition, the simple skip connection impairs the shallow semantic information in the encoder as well as the fine-grained details of the segmentation target when transferring the encoder information. The simple skip connection has limitations in capturing global information and long-distance dependencies, and it is difficult to make full use of global context information to effectively fuse multi-scale feature information. Therefore, it is easy to lose semantic information in the process of information transmission, resulting in a decrease in segmentation accuracy.

In order to solve the above problems, this chapter proposes a colon polyp segmentation network based on adaptive convolution and multi-scale global context (ACMS-TransNet). An innovative network structure integrating ACB module, MHCIA module and SEMS-FFN module. By organically combining the advantages of different modules, the network can perform efficient feature extraction and information fusion at different scales and feature levels, effectively improving the accuracy and efficiency of colon polyp segmentation. Specifically, the contributions of this paper are as follows:

1) Adaptive convolution operations are introduced at the key levels of the encoder and decoder to dynamically adjust the receptive field of the convolution kernel to adapt to

Manuscript received October 8, 2024; revised December 29, 2024. This work was supported by the Special Fund for Scientific Research Construction of the University of Science and Technology Liaoning.

Ping Sun is a postgraduate student of the University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: sunping97@163.com)

Jiansheng Wu is a professor of the University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author, e-mail: ssewu@163.com).

Zixue Zhao is a postgraduate student of the University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: mszzx1998@163.com).

Han Gao is a postgraduate student of the University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 1789502643@qq.com).

features of different scales and improve the ability to detect small targets. Multiple convolution filter branches are used to perform feature changes in spaces of different scales, effectively collecting contextual information at each spatial position, improving the accuracy of feature representation, and generating richer and more diverse output features.

2) In the skip link part, the MSCF-SETrans network is designed, which combines the multi-head channel interaction attention module (MHCIA) and the multi-scale feedforward neural network module (SEMS-FFN). Among them, the MHCIA module improves Q and K in the multi-head attention mechanism, allowing Q to perform attention mechanism calculation on K in the global scope. Through cross-channel feature aggregation, the information interaction between different channels is enhanced and the dependency between different channels is captured. The SEMS-FFN module combines four parallel deep convolutions of different scales with the channel attention mechanism (SE-Block), captures multi-scale global information in the feature fusion process, and improves the accuracy of segmentation details.

II. RELATED WORK

Medical image segmentation is a crucial research area in computer vision. It is to automatically identify and segment regions of interest from complex medical images, such as organs and lesions. Early methods in medical image segmentation primarily relied on traditional image processing techniques, such as thresholding, region growing, and edge detection [4]. While these methods perform well in some simple scenarios, they struggle with complex structures and multi-scale information. They are often affected by noise and image variability, making it challenging to achieve accurate segmentation results. Because of its strong ability of representation learning, CNN [5]–[7] has been introduced into the field of medical image segmentation. In 2015, Jonathan Long et al. proposed Fully Convolutional Networks [8]. FCN replaced fully connected layers with convolutional layers, allowing it to handle input images of any size and perform pixel-level classification, addressing the semantic-level image segmentation problem. In the same year, Ronneberger et al. built on FCNs to propose U-Net, an end-to-end encoder-decoder architecture designed for medical image segmentation. The U-Net architecture extracted hierarchical feature representations of images through the encoder and then reconstructed these features into segmentation predictions of the input image using the decoder. It introduced the concept of skip connections between the encoder and decoder to address the shortcomings of FCNs in preserving pixel spatial location and contextual information. This approach resolved the issues of local and global feature loss and achieved significant results in medical image segmentation tasks. The introduction of U-Net brought a revolutionary change to medical image segmentation. In 2018, Zhou et al. proposed a new model, U-Net++ [9], which introduced additional skip connections to enhance feature transfer. Zhang et al. proposed ResUNet [10], which combined residual connections with the U-Net structure, effectively addressing gradient vanishing and model degradation issues in deep neural networks. That same year, Ozan Oktay et al. proposed Attention U-Net [11], a hybrid structure that incorporated attention gates into the skip paths. These attention gates selectively passed significant features to the decoder while suppressing redundant information, allowing for precise reconstruction of segmentation maps. In 2019, Nabil Ibtehaz and M. Sohel Rahman et al. proposed the MultiResUNet [12], which introduced multi-scale residual modules. Despite the significant achievements of Convolutional Neural Networks in medical image segmentation, they still face limitations in handling complex global contextual information and multi-scale features. The Transformer [13] model, due to its outstanding performance in natural language processing tasks, was gradually introduced into the field of computer vision. The Transformer captured long-range dependencies between elements in a sequence through self-attention mechanisms, addressing the shortcomings of Convolutional Neural Networks in global information modeling. It enhanced the model's ability to capture global contextual information, achieving groundbreaking results in medical image segmentation tasks and significantly improving accuracy [14]. In 2021, Chen et al. proposed TransUNet [15]. This model introduced the Transformer into the encoder part of U-Net, enhancing the capture of global contextual information through self-attention mechanisms, and significantly improved performance in medical image segmentation. In 2022, Wang et al. proposed UCTransNet [16], which considered attention mechanisms from a channel perspective. The channel-level fusion Transformer replaced the traditional skip connections in U-Net, achieving better integration of semantic information between the encoder and decoder. These methods collectively advanced the field of medical image segmentation, not only enhancing segmentation accuracy but also providing diverse and innovative solutions to tackle various challenging scenarios.

III. RESEARCH METHOD

A. Overall Architecture

The network structure of the ACMS-TransNet is shown in Fig. 1. It consists of an encoder, a decoder, and skip connections. Specifically, the ACB module is introduced in both the encoder and decoder of the network. Utilizing multiple convolution filter branches that operate at different spatial scales, effectively enhances the contextual information surrounding each spatial location, enabling dynamic calibration of the input features. This approach expands the receptive field of the convolutional layers, improving the accuracy of feature representation and capturing more diverse and rich feature information. To further enhance the model's ability to effectively extract multi-scale global contextual information, better fuse the semantic information between the encoder and decoder, and improve the fine-grained details of the segmentation targets, the (MSCF-SETrans) network is designed in the skip connection section. This network includes the MHCIA module and SEMS-FFN module, which effectively integrates features from different scales. Specifically, the input feature map $Img \in R^{C\times H\times W}$ first passes through the ACB module for adaptive adjustment, expanding the receptive field and strengthening important information. Then, it undergoes a series of convolutions and max pooling operations to extract the encoding layer information$E_i \in R^{C_i\times \frac{H}{2^{i-1}}\times \frac{W}{2^{i-1}}} (i = 1, 2, 3, 4, 5)$. Before the final downsampling, the ACB module prevents key information from being lost during the downsampling process.

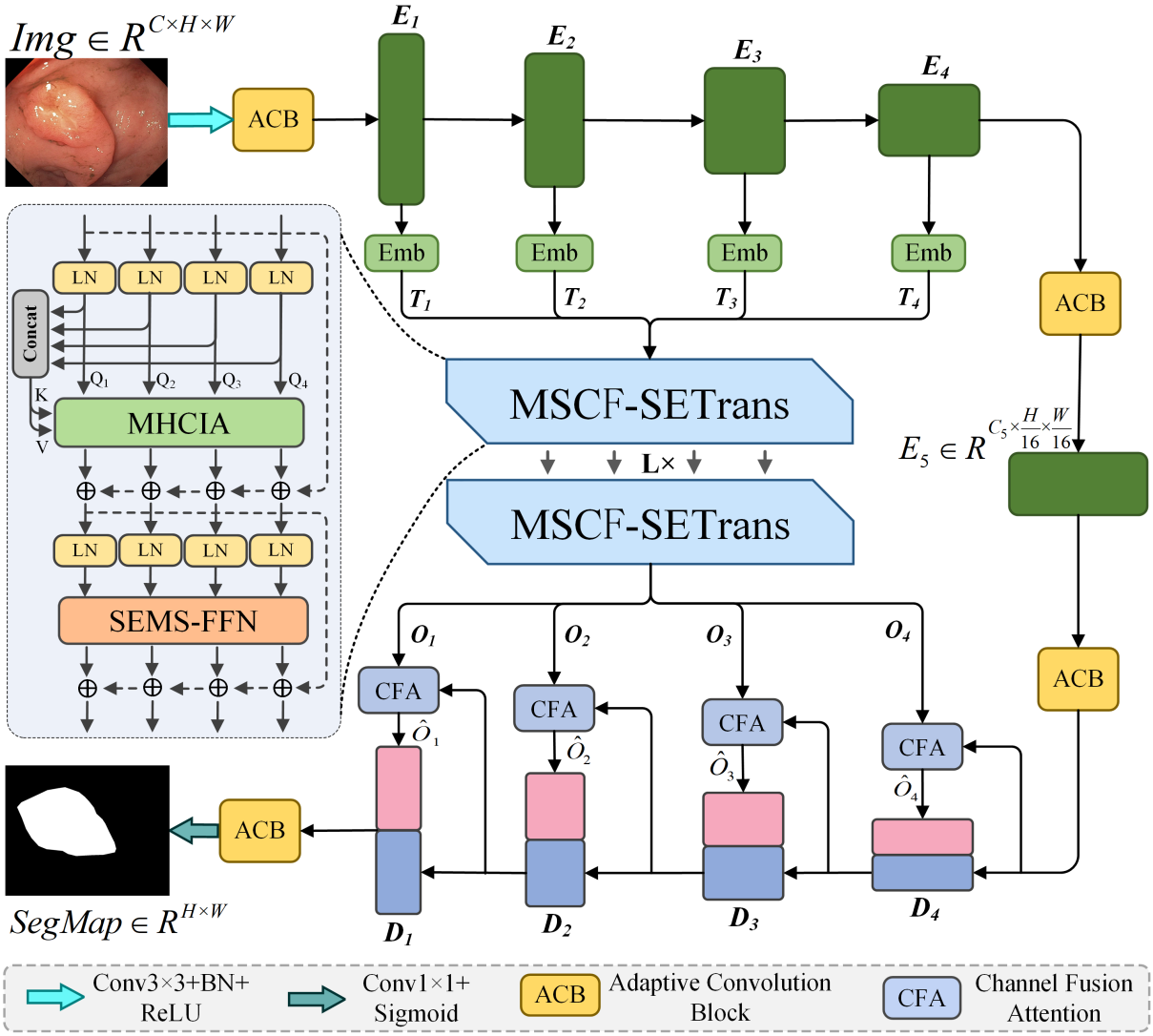

Fig. 1. ACMS-TransNet Network Architecture.

Similarly, before the first upsampling, the ACB module enhances the processing capability of high-level semantic information, effectively recovering details and reducing information loss. Meanwhile, the outputs of the first four encoding layers $E_i(i = 1, 2, 3, 4)$ are processed through patch-embedding and position-embedding to generate 2D patch sequences $T_i \in R^{d \times C_i}(i = 1, 2, 3, 4)$, which are then mapped to the multi-head attention mechanism as $Q_{\mathrm{i}} \in R^{C_i \times d}$. The sequences $T_i \in R^{d \times C_i}(i = 1, 2, 3, 4)$ are concatenated to form $T_{\sum_{i=1}^{4}} = (T_1, T_2, T_3, T_4)$, which is uesd as the keys (K) and values (V) for the attention mechanism. After passing through L layers of MSCF-SETrans, the feature tensors $O_i \in R^{C \times H \times W}(i = 1, 2, 3, 4)$ are obtained. To effectively connect to the decoder and eliminate feature ambiguity, the feature tensor O_i of the i-th layer and the feature map $D_i \in R^{C \times H \times W}(i = 1, 2, 3, 4)$ of the i-th decoder layer are reconstructed using the Channel Fusion Attention (CFA) module to obtain $\hat{O}_i$. These reconstructed features are concatenated with the corresponding upsampled features of the decoder layers and passed through convolution to output the decoded layer information. Finally, at the end of the decoder, the ACB module significantly enhances the detail recovery features, and the final segmentation result $SegMap \in R^{H \times W}$ is obtained through a 1×1 convolution followed by a Sigmoid function.

B. Adaptive Convolution Block

In order to effectively enhance the contextual awareness at each spatial location, expand the receptive field, and reduce information loss, the ACB [17] module is introduced in both the encoder and decoder. The network structure of this module is shown in Fig.2. It uses three convolution filter branches $k_i(i = 1, 2, 3)$ to realize feature transformation in two different scale Spaces of input features, in which each convolution filter has different effects.

Firstly, the input feature $X \in R^{C \times H \times W}$ is subsampled by k_1 branch averaging pooling, which shrinks it to smaller scale space $S_1 \in R^{C \times \frac{H}{r} \times \frac{W}{r}}$, enlarges the receptive field and generates low-resolution embedded features.The calculation of S_1 is as follows:

$$S_1 = AvgPool_r(X) \tag{1}$$

These embedded features undergo feature transformation through convolution operations to generate reference signals,

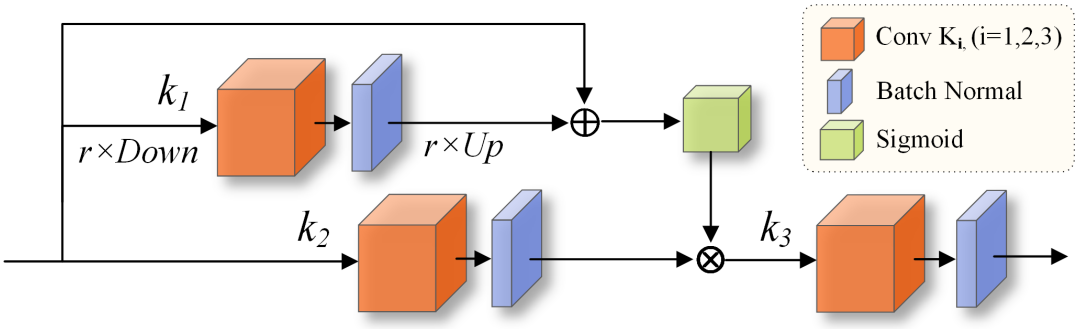

Fig. 2. ACB network structure diagram.

which are then normalized and upsampled back to the original scale space $S_2 \in R^{C\times H\times W}$ using bilinear interpolation. After residual connection with the original input image and passing through a Sigmoid function, adaptive calibration weights are generated:

$$S_2 = Up(BN(S_1 * K_1)) \tag{2}$$

In the k_2 branch, Convolution operations are performed on the original features to capture fine-grained information, which is then multiplied by the weights output from the k_1 branch to achieve dynamic calibration of the input features. Finally, the calibrated and fused feature map is passed through the k_3 branch for another convolution operation to integrate the previously extracted features, producing the final output feature map Y. The related computation formulas are as follows:

$$X_1 = \sigma(S_2 + X) \tag{3}$$

$$X_2 = BN(X * K_2) \times X_1 \tag{4}$$

$$Y = BN(X_2 * K_3) \tag{5}$$

In the formulas above: r represents the step size, Up represents the bilinear interpolation operation, $K_i(i = 1, 2, 3)$ represents the convolution kernels of each branch, "$*$" represents the convolution operation, σ indicates the Sigmoid function, and "$\times$" signifies element-wise multiplication.

C. MSCF-SETrans Network

This paper designs the MSCF-SETrans network in the skip connections of the model. This network combines the strengths of Transformer and convolutional neural networks (CNNs) by calculating the correlations between information from different encoding layers to capture target-specific information. By exploring multi-scale global contextual information, it establishes a connection between the encoder and decoder, replacing the simple skip connections. This reduces the semantic gap between the encoder and decoder, enabling more effective fusion of features from different scales. Compared to traditional Transformers, the MSCF-SETrans network retains local detail information while enhancing global contextual information extraction, resulting in richer feature representations. The structure of the MSCF-SETrans network

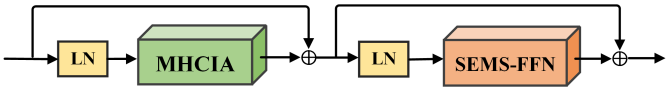

Fig. 3. MSCF-SETrans network structure diagram.

is shown in Fig.3 and primarily consists of the MHCIA module, SEMS-FFN module, and layer normalization (LN).

Input feature X passes through MHCIA and SEMS-FFN modules successively, before which LN layer normalization operation is used to maintain the consistency of feature distribution in different layers, and then the output results of these two modules are respectively residual with the feature maps before normalization operation to obtain the final output feature Y.

1) Multi-Head Channel Interaction Attention: The Multi-Head Self-Attention mechanism is one of the core components of the Transformer model. It captures the dependencies between different positions when processing sequence data [13]. Its computational complexity increases with the growth of spatial or channel dimensions. Moreover, directly applying simple dot-product operations to the flattened query vector Q and key vector K may reduce the correlation between feature channels. To solve the above problems, we propose an MHCIA module, Its structure is shown in Fig.4. The outputs

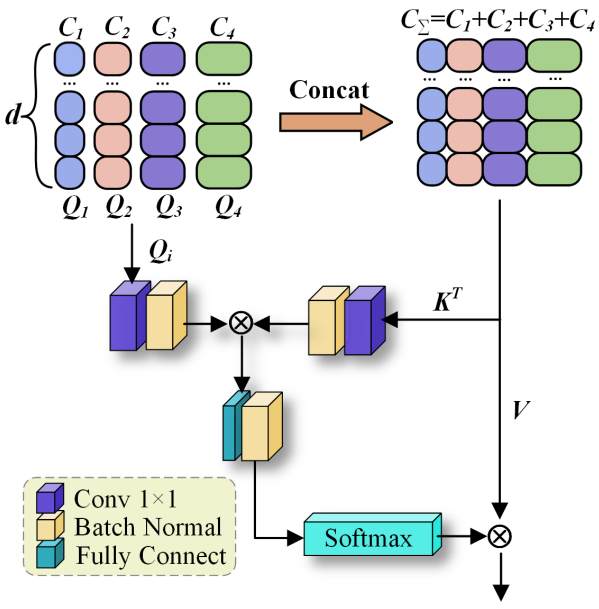

Fig. 4. MHCIA network structure diagram

of the first four encoder layers are each mapped to four distinct query matrices $Q_i \in \mathbb{R}^{C_i \times d}$ (for $i = 1, 2, 3, 4$). These matrices are then concatenated to form $K \in \mathbb{R}^{C_\Sigma \times d}$ and $V \in \mathbb{R}^{C_\Sigma \times d}$. Before multiplying Q_i and K, both are passed through 1×1 convolution and batch normalization, which effectively reduces computational complexity, enhances feature interaction between different channels, and captures local contextual information. After the dot-product operation, a fully connected layer extends the channel dimensions to better align with the feature information in the value vectors V. The input feature vectors carry out the cross-attention mechanism along the channel axis, and the similarity matrix is generated through Q_i, K, V, and the value vector V is weighted. The formula for calculating the improved attention mechanism is as follows:

$$Z_i = \frac{BN(c(Q_i) \times BN(c(K^T)))}{\sqrt{d_k}} \tag{6}$$

$$Attention(Q_i, K, V) = Softmax(BN(FC(Z_i)))V \tag{7}$$

Where d_k is the dimension of the key vector, c is the 1×1 convolution.This module aggregates multi-level features by interacting K and V with Q_i, capturing the dependencies between different channels, weighting important feature channels, and suppressing less important ones.

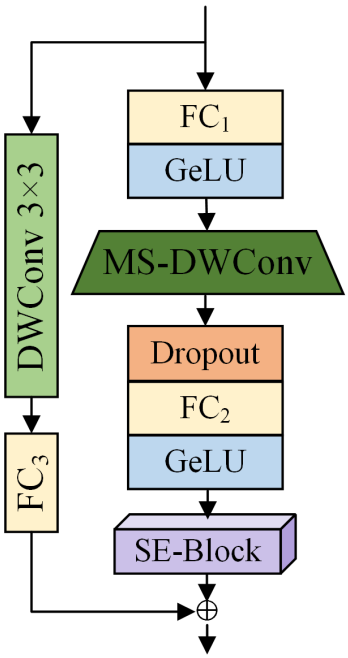

Fig. 5. SEMS-FFN network structure diagram.

2) SE-Multiscale Feedforward Neural Network: The SEMS-FFN module is another important component of the MSCF-SETrans. This module builds upon the standard feedforward neural network by incorporating a Squeeze-and-Excitation block(SE-Block) [18] and multi-scale depthwise convolution operations(MS-DWConv) [19]. Its structure shown in Fig.5, begins with the first fully connected layer FC_1, which maps the input features to a higher-dimensional space. The GeLU activation function and the multi-scale depthwise convolution module further enhance the nonlinear feature representation. The second fully connected layer FC_2 maps the features back to their original dimensions. Subsequently, the Squeeze-and-Excitation (SE) block further models the relationships between different channels and adaptively recalibrates the feature responses across channels. Finally, residual connections with depthwise separable convolutions [20] and a fully connected layer FC_3 capture local information, enabling the model to extract local features more effectively.

After the first fully connected layer, multi-scale depthwise convolutions [19] are introduced, as shown in Fig.6. Four parallel depthwise convolutions of different scales are applied, each processing one-quarter of the channels with kernel sizes of $\{1, 3, 5, 7\}$, to perform multi-scale token aggregation. This design leverages receptive fields of varying scales to capture multi-scale information from the input feature map, thereby enhancing the understanding and fusion of features.

After the second fully connected layer, SE-Block is introduced to calculate adaptive weights for each channel, automatically selecting the feature channels most relevant to the current task. This highlights useful information and reduces redundancy. The Squeeze operation generates a global feature descriptor through global average pooling, compressing the spatial dimensions of each channel and calculating the global average value for each channel to obtain the channel descriptor vector. Let i and j represent the height and width of the pixel spatial coordinates of the feature map X_c, respectively:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{c,i,j}, (c = 1, 2, \cdots C) \tag{8}$$

The Excitation operation processes the channel descriptor vector through two fully connected layers to generate channel weights, adaptively recalibrating the feature responses of each channel. Finally, the channel weights are applied back to the input feature map. Let s represent the generated channel weights, W_1 and W_2 represent the weight matrice of the fully

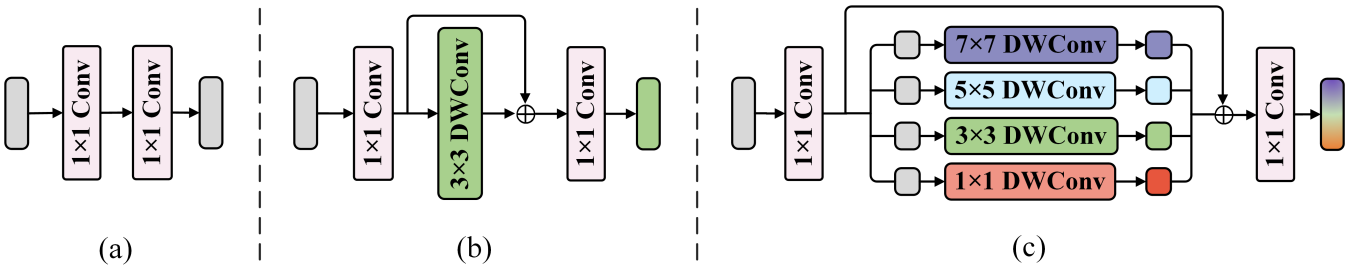

Fig. 6. (a) The FFN focuses solely on processing channel information of the feature map. (b) The FFN further aggregates token information within the region. (c) Our MS-DWConv performs multi-scale token aggregation through four parallel depthwise convolutions.

connected layers, and σ represent the Sigmoid function.

$$s_c = \sigma(W_2 \cdot (\mathrm{Re}LU(W_1 \cdot z_c + b_1) + b_2)) \tag{9}$$

$$\tilde{x} = s_c \times X_c \tag{10}$$

To sum up, the calculation formula of the EMS-FFN network is as follows:

$$X_1 = SE(FC_2(D(M(G(FC_1(X)))))) \tag{11}$$

$$X_2 = FC_3(DW(X)) \tag{12}$$

$$Y = X_1 + X_2 \tag{13}$$

Here,SE stands for the Squeeze-and-Excitation Block, FC stands for Fully Connected, D stands for Dropout, M stands for Multi-scale Depthwise Convolution, G stands for the $GeLU$ activation function, and DW stands for 3×3 Depthwise Separable Convolution.

IV. Experiments and Results

A. Dataset Description

To validate the performance of the proposed network, we used two colonoscopy polyp image datasets, namely the Kvasir-SEG dataset and the CVC-ClinicDB dataset.

1)Kvasir-SEG is a dataset for the MediaEval2020 competition [21]. The dataset contains 1000 colonoscopy images with polyps and their corresponding segmentation masks, including polyps of different sizes, shapes and positions, the image resolution of the dataset ranges from 332×487 to 1920×1072 pixels, providing a diverse resource of high-quality images.

2)CVC-ClinicDB is the official data set for the training phase of the MICCAI2015 Colonoscopy Video Automated Polyp Detection Subchallenge [22]. The dataset consisted of 612 static polyp images extracted from colonoscopy videos and corresponding hand-labeled polyp masks from 29 different sequences with an image resolution of 384×288.

B. Evaluation Metrics

To quantitatively evaluate the segmentation results, this paper adopts five evaluation metrics: Dice [23], IoU, Recall, Precision, and Accuracy. The closer these values are to 1, the better the performance. Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. The corresponding formulas are as follows:

1) Dice: The Dice coefficient is a statistical measure used to evaluate the similarity between two sample sets:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{14}$$

2) IoU: The IoU used to measure the overlap between the predicted result and the ground truth:

$$IoU = \frac{TP}{TP + FP + FN} \tag{15}$$

3) Recall: Recall measures the proportion of actual positive samples that the model correctly identifies as positive:

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

4) Precision: Precision refers to the proportion of positive data that a model can correctly predict among all predicted positive samples:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

5) Accuracy: Accuracy refers to the proportion of positive data samples predicted by the model in the total sample:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

C. Experimental Environment and Parameters

The ACMS-TransNet proposed in this study is implemented using the PyTorch1.8 framework. The operating system is Ubuntu18.04, and the GPU used is an NVIDIA GeForce RTX 3060 with 12GB of VRAM. The programming environment includes Python3.8 and CUDA 11.8. The Adam optimizer is used with an initial learning rate of 10^{-4}, the model is trained using a combination of cross-entropy loss and Dice loss functions. The input image resolution is set to 224×224, with a patch size of 16 and a batch size of 4.

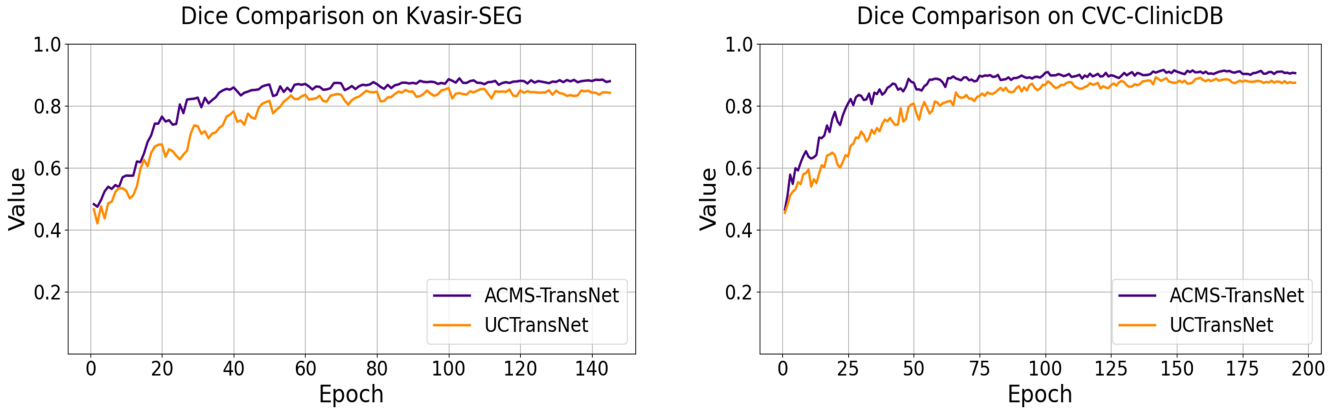

Fig. 7. Train the Dice coefficient curve

D. Experimental Results and Analysis

The Dice coefficient is a widely used metric for evaluating the performance of segmentation models. During training, early stopping is applied by monitoring the changes in the Dice coefficient on the validation set to prevent overfitting and save computational resources. If the Dice coefficient does not improve for 50 consecutive epochs, the training is terminated early. Fig.7 presents the training curves of the baseline model UCTransNet and the proposed ACMS-TransNet on the Kvasir-SEG and CVC-ClinicDB datasets. These curves provide an intuitive visualization of the models' convergence rates and trends in final segmentation performance. Notably, ACMS-TransNet demonstrates significant advantages in both convergence speed and segmentation accuracy during the training process.

To further validate the segmentation performance of ACMS-TransNet on colorectal polyp images, a series of comparative experiments were conducted with state-of-the-art medical image segmentation models. These models include U-Net, U-Net++, Attention U-Net, ResUNet++, DoubleU-Net [24], TransUNet, UCTransNet, and DA-TransUNet [25]. The comparison results for the Kvasir-SEG dataset are shown in Table I, and the comparison results for the CVC-ClinicDB dataset are shown in Table II.

From the results in the table, it is evident that the proposed model achieves significant improvements across all evaluation metrics on both datasets, the segmentation performance is obviously better than the classical method. For the Kvasir-SEG dataset, the proposed model achieves the following results for the five evaluation metrics: 90.46%, 83.69%, 92.48%, 90.35%, and 96.66%. These results exceed those of the comparison models across all metrics. Specifically, compared to the baseline model UCTransNet, the proposed model shows improvements of 2.4%, 2.64%, 1.06%, 1.71%, and 0.5%, respectively. Compared to the classic U-Net model, it demonstrates enhancements of 7.27%, 11.78%, 10.47%, 7.52%, and 1.85%.For the CVC-ClinicDB dataset, the proposed model achieves the following results for the five evaluation metrics: 94.45%, 89.68%, 95.74%, 93.43%, and 98.97%. These results show improvements over the baseline model UCTransNet by 0.93%, 1.54%, 1.1%, 0.42%, and 0.21%, respectively. Compared to the classic U-Net model, the proposed model demonstrates enhancements of 7.16%, 8.75%, 6.38%, 5.18%, and 0.48%. This verifies the excellence of the proposed ACMS-TransNet.

To more intuitively compare different segmentation methods, three groups of data samples were randomly selected from the Kvasir-SEG and CVC-ClinicDB datasets. The segmentation effect of the Kvasir-SEG dataset is shown in Fig.8, and the segmentation effect of the CVC-ClinicDB dataset is shown in Fig.9. The comparison images demonstrate that ACMS-TransNet achieves higher accuracy in medical image segmentation.

E. Ablation Experiment

To comprehensively validate the effectiveness of the proposed modules and design strategies, as well as their impact on overall segmentation performance, we perform ablation studies focusing on different kernel sizes of MS-DWConv, the number of layers in the skip connection MSCF-SETrans, and the contribution of various innovative module combinations.Using UCTransNet as the baseline model, we first conduct experiments on Kvasir-SEG with dynamic convolution kernels of different scales. The experimental results, shown in Table III, indicate that the four-channel parallel multi-scale depthwise convolutions $\{1, 3, 5, 7\}$ achieve the best performance.

TABLE III
ABLATION ON CONVOLUTION KERNEL SCALES IN MS-DWCONV

Scales	Dice(%)	IoU(%)	Params(M)	Flops(G)
$\{3\}$	88.03	80.97	65.87	32.68
$\{1, 3\}$	88.05	81.02	65.95	32.76
$\{1, 3, 5\}$	88.09	81.04	66.15	32.85
$\{1, 3, 5, 7\}$	88.14	81.07	66.24	32.98
$\{1, 3, 5, 7, 9\}$	88.11	81.05	66.92	33.24

TABLE I
COMPARISON RESULTS OF DIFFERENT METHODS ON KVASIR-SEG

Method	Dice(%)	IoU(%)	Recall(%)	Precision(%)	Accuracy(%)
U-Net	83.19	71.91	82.01	85.83	94.81
U-Net++	83.59	72.43	81.86	86.75	94.94
Attn-UNet	84.01	73.17	81.37	88.18	95.11
ResUNet++	82.26	79.30	80.55	84.64	92.97
DoubleU-Net	81.30	73.30	84.40	86.10	-
TransUNet	86.78	79.91	87.31	87.69	96.29
UCTransNet	88.06	81.05	91.42	88.64	96.16
DA-TransUNet	88.47	81.02	-	-	-
ACMS-TransNet(ours)	90.46	83.69	92.48	90.35	96.66

TABLE II
COMPARISON RESULTS OF DIFFERENT METHODS ON CVC-CLINICDB

Method	Dice(%)	IoU(%)	Recall(%)	Precision(%)	Accuracy(%)
U-Net	87.29	80.93	89.36	88.25	98.49
U-Net++	87.64	81.16	85.97	91.99	98.42
Attn-UNet	89.58	83.59	90.15	90.40	98.66
ResUNet++	85.46	78.13	85.33	87.19	98.21
DoubleU-Net	92.39	86.11	84.57	93.32	-
TransUNet	89.63	83.66	91.27	89.29	98.67
UCTransNet	93.52	88.14	94.64	93.01	98.76
DA-TransUNet	89.47	82.51	-	-	-
ACMS-TransNet(ours)	94.45	89.68	95.74	93.43	98.97

Compared to a single scale, adding 1×1 and 5×5 convolution kernels improves the handling of details and local features, while the 7×7 kernel effectively captures broader contextual information, enhancing the multi-scale representation of the feature map. However, introducing a 9×9 convolution kernel results in feature redundancy and increased computational complexity, leading to a decline in model performance.

Subsequently, experiments are conducted on the Kvasir-SEG dataset to evaluate the effect of the number of MSCF-SETrans layers in the skip connections. The number of layers L is set to 2, 4, 8, 12, as shown in Table IV. The results indicate that the model performs best when L is set to 4. Further increasing the number of layers, however, leads to performance degradation.

TABLE IV
ABLATION ON THE NUMBER OF LAYERS IN MSCF-SETRANS(UNIT: %)

Layers	Dice	IoU	Recall	Precision	Accuracy
2	78.89	69.38	87.11	77.85	92.76
4	88.25	81.10	91.49	89.65	96.33
8	88.01	80.89	90.65	88.69	96.09
12	87.97	80.08	90.23	87.85	95.91

Finally, ablation experiments are conducted on both the Kvasir-SEG and CVC-ClinicDB datasets to evaluate the impact of different combinations of the ACB module, MHCIA module, and SEMS-FFN module. These experiments are designed to verify the contribution of each module to the overall performance of the network model. The experimental results are shown in TableV and TableVI, indicate that the best performance is achieved by combining the three modules (the full version of the ACMS-TransNet model). In contrast, using any of these modules alone or in combination is less effective than the combination of all three. The individual use of ACB, MHCIA, and SEMS-FFN modules can each improve the overall performance of the model to some extent, with ACB showing the most significant improvement. Although MHCIA and SEMS-FFN can bring some improvement when used individually, the enhancement is relatively weak, especially for SEMS-FFN, where some metrics perform worse than the baseline model. However, the combination of SEMS-FFN and MHCIA shows some improvement, indicating that their combination can better leverage their strengths. It is worth noting that the combination of MHCIA or SEMS-FFN with ACB does not

TABLE V
ABLATION EXPERIMENT OF MODULES ON KVASIR-SEG

Base	ACB	MHCIA	SEMS-FFN	Dice(%)	IoU(%)	Recall(%)	Precision(%)	Accuracy(%)
✓				88.06	81.05	91.42	88.64	96.16
✓	✓			89.73	83.03	91.86	90.04	96.50
✓		✓		88.19	81.09	91.72	88.48	96.07
✓			✓	88.14	81.07	91.41	88.36	96.01
✓	✓	✓		88.58	82.15	91.25	88.43	95.95
✓	✓		✓	88.92	82.66	90.84	89.95	96.16
✓		✓	✓	88.25	81.10	90.49	89.65	96.33
✓	✓	✓	✓	90.46	83.69	92.48	90.35	96.66

TABLE VI
ABLATION EXPERIMENT OF MODULES ON CVC-CLINICDB

Base	ACB	MHCIA	SEMS-FFN	Dice(%)	IoU(%)	Recall(%)	Precision(%)	Accuracy(%)
✓				93.52	88.14	94.64	93.01	98.76
✓	✓			93.81	88.60	94.99	93.91	98.86
✓		✓		93.63	88.35	93.63	94.07	98.77
✓			✓	93.58	88.21	94.56	93.38	98.81
✓	✓	✓		93.65	88.45	95.45	91.06	98.79
✓	✓		✓	93.77	88.52	94.78	91.89	98.80
✓		✓	✓	94.20	89.31	94.73	94.22	98.91
✓	✓	✓	✓	94.45	89.68	95.79	93.41	98.97

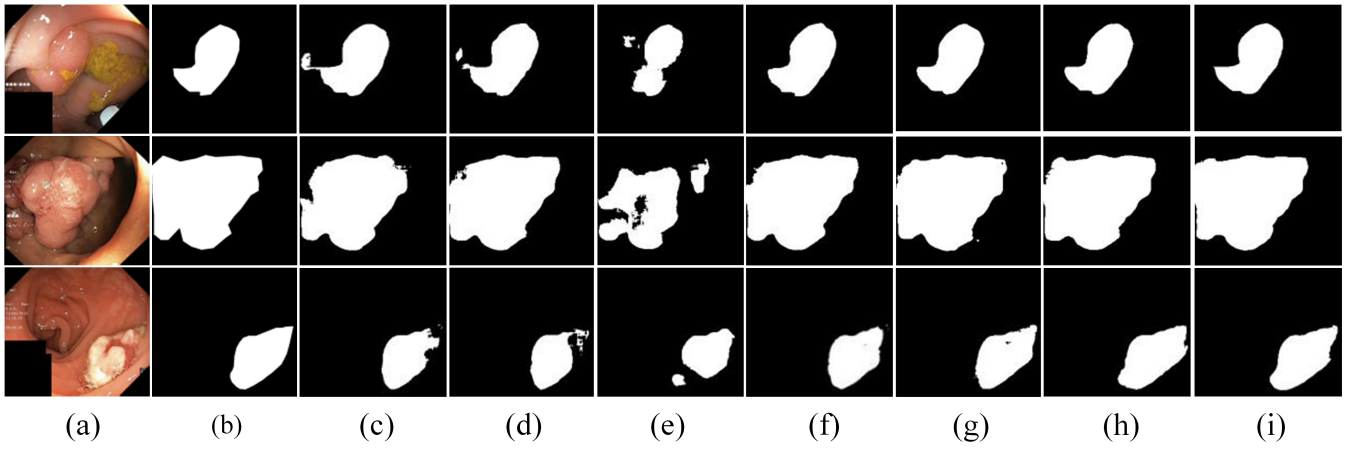

Fig. 8. Segmentation effect diagram of Kvasir-SEG. Column(a) shows the original images, column(b) shows the ground truth masks, and columns(c) to (i) show the segmentation results from U-Net, U-Net++, ResUNet++, Attention U-Net, TransUNet, UCTransNet, and the ACMS-TransNet, respectively.

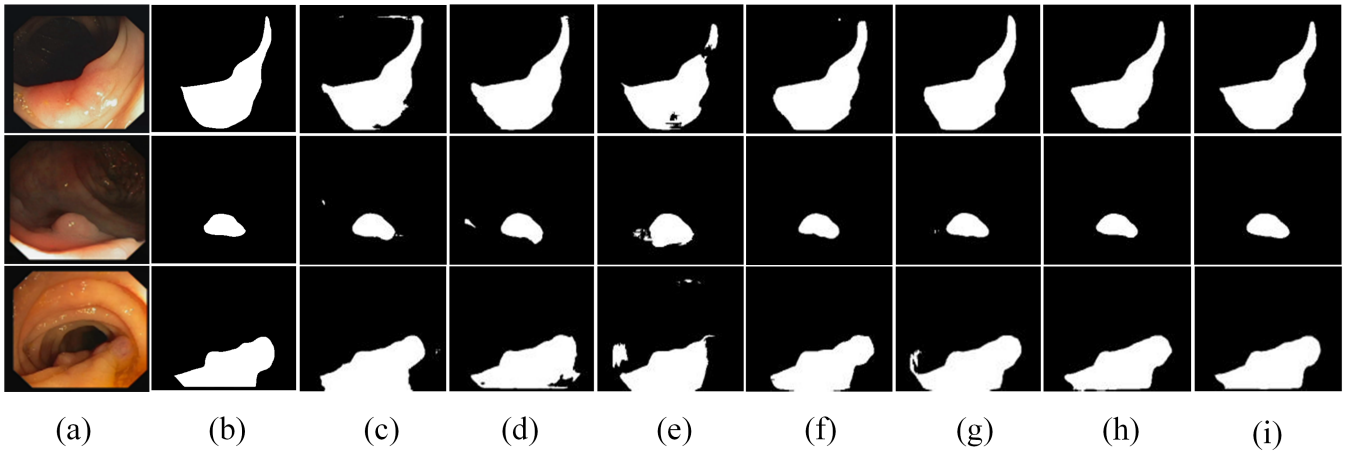

Fig. 9. Segmentation effect diagram of CVC-ClinicDB. Column(a) shows the original images, column(b) shows the ground truth masks, and columns(c) to (i) show the segmentation results from U-Net, U-Net++, ResUNet++, Attention U-Net, TransUNet, UCTransNet, and the ACMS-TransNet, respectively.

outperform ACB when used alone. However, the combination of MHCIA and SEMS-FFN complements each other better, demonstrating their synergistic effect. In summary, each module plays an important role in improving the model's performance. Removing any module leads to a decline in performance, while the joint use of all three modules effectively leverages their respective advantages, maximizing the model's segmentation ability.

We proposed ACMS-TransNet realizes the synergistic effect of multiple modules and enhances the fusion of local and global information by effectively combining ACB, MHCIA and SEMS-FFN modules. In all evaluation indicators, the model was superior to the baseline model, which fully verified that the ACMS-TransNet significantly improved the segmentation accuracy of polyp data sets, while maintaining a high accuracy.

V. Conclusion

In this study, we propose the ACMS-TransNet for polyp segmentation tasks. The model significantly improves the performance of polyp segmentation by introducing several innovative modules, including the ACB module, the MHCIA module of the MSCF-SETrans, and the SEMS-FFN module. Through in-depth analysis and verification, the results show that the ACMS-TransNet is superior to these modules alone and some mainstream traditional methods based on U-Net in several key evaluation indicators, which verifies the effectiveness and superiority of the model in colon polyp segmentation task. In future research, we plan to further optimize the model structure by incorporating more advanced deep learning techniques to improve segmentation accuracy.

References

[1] H. Brenner, J. Chang-Claude, C. M. Seiler, A. Rickert, and M. Hoffmeister, "Protection from colorectal cancer after colonoscopy: a population-based, case-control study," *Annals of internal medicine*, vol. 154, no. 1, pp. 22–30, 2011.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention*. Munich, Germany: Springer, 2015, pp. 234–241.

[4] P. K. Sahoo, S. Soltani, and A. K. Wong, "A survey of thresholding techniques," *Computer vision, graphics, and image processing*, vol. 41, no. 2, pp. 233–260, 1988.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, 2018, pp. 3–11.

[10] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, pp. 749–753, 2018.

[11] O. Oktay, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[12] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural networks*, vol. 121, pp. 74–87, 2020.

[13] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[14] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Medical image computing and computer assisted intervention*, 2021, pp. 36–46.

[15] 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," *arXiv preprint arXiv:2102.04306*, 2021.

[16] 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 *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2441–2449.

[17] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 096–10 105.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[19] M. Lou, H.-Y. Zhou, S. Yang, and Y. Yu, "Transxnet: learning both global and local dynamics with a dual dynamic token mixer for visual recognition," *arXiv preprint arXiv:2310.19380*, 2023.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[21] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*. Springer, 2020, pp. 451–462.

[22] 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," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[23] J. Bertels, T. Eelbode, M. Berman, D. Vandermeulen, F. Maes, R. Bisschops, and M. B. Blaschko, "Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 92–100.

[24] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "Doubleu-net: A deep convolutional neural network for medical image segmentation," in *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*. IEEE, 2020, pp. 558–564.

[25] G. Sun, Y. Pan, W. Kong, Z. Xu, J. Ma, T. Racharak, L.-M. Nguyen, and J. Xin, "Da-transunet: integrating spatial and channel dual attention with transformer u-net for medical image segmentation," *Frontiers in Bioengineering and Biotechnology*, vol. 12, p. 1398237, 2024.