Journal of Digital Imaging (2023) 36:2427–2440
https://doi.org/10.1007/s10278-023-00885-y

ECTransNet: An Automatic Polyp Segmentation Network Based on Multi-scale Edge Complementary

Weikang Liu[1] · Zhigang Li[1] · Chunyang Li[1] · Hongyan Gao[1]

Received: 13 April 2023 / Revised: 13 July 2023 / Accepted: 14 July 2023 / Published online: 25 July 2023

Abstract

Colonoscopy is acknowledged as the foremost technique for detecting polyps and facilitating early screening and prevention of colorectal cancer. In clinical settings, the segmentation of polyps from colonoscopy images holds paramount importance as it furnishes critical diagnostic and surgical information. Nevertheless, the precise segmentation of colon polyp images is still a challenging task owing to the varied sizes and morphological features of colon polyps and the indistinct boundary between polyps and mucosa. In this study, we present a novel network architecture named ECTransNet to address the challenges in polyp segmentation. Specifically, we propose an edge complementary module that effectively fuses the differences between features with multiple resolutions. This enables the network to exchange features across different levels and results in a substantial improvement in the edge fineness of the polyp segmentation. Additionally, we utilize a feature aggregation decoder that leverages residual blocks to adaptively fuse high-order to low-order features. This strategy restores local edges in low-order features while preserving the spatial information of targets in high-order features, ultimately enhancing the segmentation accuracy. According to extensive experiments conducted on ECTransNet, the results demonstrate that this method outperforms most state-of-the-art approaches on five publicly available datasets. Specifically, our method achieved mDice scores of 0.901 and 0.923 on the Kvasir-SEG and CVC-ClinicDB datasets, respectively. On the Endoscene, CVC-ColonDB, and ETIS datasets, we obtained mDice scores of 0.907, 0.766, and 0.728, respectively.

Keywords Colonoscopy · Polyp segmentation · Multi-scale features · ECTransNet

Introduction

Medical image segmentation has become a prevalent approach for the classification of distinct anatomical structures within a given tomographic image. It is utilized in various applications, such as liver segmentation [1], gastric lesion segmentation [2], and pancreas segmentation [3] or polyp segmentation, which enables the differentiation between abnormal regions and normal regions. Polyps are abnormal growths on the surfaces of organs (such as the colon, rectum, stomach, and throat) that can potentially lead to cancer. Colorectal cancer is ranked as the third most common cancer in males and the second most common in females, according to the World Health Organization's global human database [4]. Polyps can be classified into two main types: non-neoplastic (including inflammatory, juvenile, and hyperplastic polyps) and neoplastic (including adenomatous and serrated polyps). The clinical characteristics of polyps depend on their type, location, size, and number. Some polyps are asymptomatic, while others may cause bleeding, pain, changes in bowel habits, or other complications. Timely detection and removal of precancerous polyps can effectively reduce the incidence of colorectal cancer.

Currently, colonoscopy examination represents the most conventional approach for polyp detection. However, recent reports have highlighted that approximately 26% of colonoscopy examinations result in missed lesions due to its reliance on manual manipulation and the clinical acuity

✉ Zhigang Li
li7275@163.com

Weikang Liu
lwk1112@foxmail.com

Chunyang Li
justlcy1@163.com

Hongyan Gao
3140039278@163.com

[1] School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114051, China

and experience of the endoscopist [5]. Colonic polyp image segmentation is a technique used to identify and separate polyp regions from the background in colonoscopic images or videos. It provides accurate and reliable information regarding the position, size, shape, and quantity of polyps. Subsequently, through the utilization of computerized image analysis and identification of detailed features, it enables pathological classification, thereby achieving the objective of rapid screening. Consequently, this technique has the potential to reduce oversights and omissions by endoscopists, thereby lowering the miss rate of polyps or other abnormal tissues. This holds significant importance in the prevention of colorectal cancer.

Nevertheless, polyp image segmentation still faces several challenges and issues [6]. Firstly, the quality of colonoscopy images is influenced by various factors such as lighting conditions, occlusions, blurriness, artifacts, blood, mucus, and more. These factors can decrease image contrast and clarity, thereby increasing the difficulty of segmentation. Secondly, polyps exhibit diversity and uncertainty in their appearance, including variations in size, shape, color, texture, and other features. Such variations may exist not only among different types of polyps but also within different regions of the same polyp, leading to segmentation instability and inconsistency.

Conventional techniques for polyp segmentation primarily rely on low-level attributes such as texture [7], geometric features [8], and simple linear iterative clustering of super-pixels [9]. Regrettably, these methods tend to produce inferior segmentation accuracy and limited generalization capability. With the development of deep learning in medical image analysis, deep learning techniques provide an efficient and accurate solution for polyp segmentation. In recent years, several image segmentation models based on convolutional neural networks (CNNs) have exhibited remarkable performance [10–12].

The U-shaped encoder–decoder architecture has gained increasing attention due to its potential to generate high-resolution predictions by leveraging multi-level features for reconstruction. The utilization of skip connections enables the effective transmission of complex and information-rich feature maps from the encoder network to the decoder sub-networks in deep neural networks. This technique facilitates the propagation of high-level semantic features while retaining fine-grained details throughout the network architecture [13, 14]. PraNet [15] utilized reverse attention mechanisms to effectively localize polyp boundary regions. This is achieved by incorporating advanced feature aggregation and boundary attention blocks within the network architecture, which aid in aligning misaligned predictions and improving overall segmentation accuracy. Hardnet-mseg [16] built upon PraNet and proposed a simplified encoder–decoder architecture. Specifically, they replaced the original Res2Net backbone network with Hardnet [17] and removed the attention mechanism to achieve faster and more accurate polyp segmentation. HRENet [18] introduced the information context enhancement (ICE) technique and the adaptive feature aggregation (AFA) module, along with training using edge and structure consistency perception loss (ES-CLoss), resulting in outstanding model performance.

PolypSeg [19] employed two distinct modules, the adaptive scale context module (ASCM) and the semantic global context module (SGCM), to address the inherent dimensional variations between polyps and enhance the fusion of high- and low-level features. ASCM is specifically designed to improve feature representation by adapting to the varying scales of polyps, while SGCM facilitates global context modeling to effectively fuse features at different scales.

To address the challenges of scale imbalance and color distribution caused by target size, SANet [20] employed probability correction and color migration strategies. To address redundancies and complementarities within multi-scale features, MSNet [21] introduced multi-scale subtractive networks. In a similar vein, MSRFNet [22] incorporated a cross-scale fusion mechanism to disseminate high- and low-level features while also introducing a shape flow network to refine polyp boundaries. TGANet [23] utilized a text-guided approach to assimilate the distinctive features of polyps of varying dimensions, with the ultimate goal of enhancing the network's capacity to generalize across different polyp sizes.

The majority of the aforementioned research methods are based on CNNs. While they have shown certain improvements in segmentation accuracy compared to traditional approaches based on superpixels and textures, there is still room for enhancement in terms of polyp boundary localization and segmentation. PraNet utilizes reverse attention to enhance the model's ability to locate polyp boundaries, but its segmentation accuracy remains unsatisfactory. HRENet and PolypSeg have employed approaches that enhance global contextual semantic relationships, resulting in improved model performance. However, CNNs often lose some crucial information during the downsampling process and have limited receptive fields, which weakens their ability to build global contextual semantic relationships. As a result, traditional CNN methods often exhibit limited generalization ability when applied to polyp images from different patients [24]. MSNet and MSRFNet address this limitation by processing multi-scale feature maps, but they lack attention to semantic information among pixels. TGANet employs text-guided attention to focus on features of polyps at different sizes, enhancing generalization ability. However, there is still a need for further improvement in the segmentation of local details in TGANet.

To enhance the focus on semantic information among pixels, we attempted to incorporate the transformer [25] encoder into the task of polyp image segmentation. Unlike traditional convolutional neural networks, the transformer utilizes self-attention mechanisms, enabling the model to capture a larger effective receptive field. For long sequence data, the transformer can learn dependencies between different positions in the sequence through self-attention, thereby better capturing global contextual semantic information. However, the transformer may lack the ability to extract local feature information effectively. Therefore, we decided to adopt a hybrid structure combining convolutional neural networks and transformer. We used the convolutional neural networks to extract local polyp features and then employed the self-attention mechanism of the transformer to build global semantic relationships among pixels in the polyp feature maps. This approach overcomes the weak ability of convolutional neural networks to construct global contextual semantic relationships, thereby improving the accuracy and generalization of the model for polyp segmentation. Furthermore, to address the issue of polyp boundary localization, we designed an edge complementary module (ECM). The ECM fuses and complements polyp feature maps at multiple resolutions, reducing the loss of crucial information and effectively capturing boundary cues of the polyps.

In this study, we propose a novel polyp image segmentation model called ECTransNet, aimed at improving the accuracy and generalization of polyp image segmentation tasks. ECTransNet combines convolutional neural networks, transformer, and ECM, integrating the local feature extraction capability of convolutional neural networks with the global semantic relationship construction capability of transformer. In the feature map decoding process, to better utilize the feature maps outputted by ECM, we introduce a residual-based feature aggregation decoder (FAD) that adaptively fuses high-level and low-level features. This approach preserves accurate spatial position information while enhancing the model's capability to segment local details. The objective of this research is to validate this deep learning model and test the feasibility of this approach.

Materials and Methods

Datasets

This study utilized five publicly available polyp datasets, namely, Kvasir-SEG [26], CVC-ClinicDB [27], Endoscene [28], CVC-ColonDB [29], and ETIS [30]. A detailed description of each dataset will be provided below.

Kvasir-SEG The Kvasir-SEG dataset consists of 1000 polyp images with their corresponding annotations. What sets this dataset apart from others is that the size and shape of the polyps vary within images. The image sizes range from 332 × 487 to 1920 × 1072. In the dataset, there are 700 large polyps, greater than 160 × 160, 48 small polyps, less than 64 × 64, and 323 medium-sized polyps. Of these, 900 images were used for training and validation, and 100 images were used for testing.

CVC-ClinicDB CVC-ClinicDB contains 612 images from 25 colonoscopy examination videos, from which 29 sequences were selected. The image size is 384 × 288. 550 images were used for training and validation, and 62 images were used for testing.

Endoscene Endoscene contains 912 images from 44 colonoscopy examination sequences of 36 patients. As the Endoscene dataset is a combination of CVC-ClinicDB and CVC-300, we used CVC-300 as a testing dataset, which consists of 60 images in total.

CVC-ColonDB CVC-ColonDB comes from 15 different colonoscopy examination sequences, from which 380 images were sampled. All 380 images were used for testing.

ETIS ETIS contains 196 images collected from 34 colonoscopy examination videos. The image size is 1225 × 966. The polyp shapes in this dataset have a larger variation than other datasets, and they are mostly small and difficult to detect, making this dataset very challenging. All 196 images in the dataset were used for testing.

We followed the same training setting as PraNet and randomly split the images of Kvasir-SEG and CVC-ClinicDB into 80% for training (800 images from Kvasir and 488 images from CVC-ClinicDB were mixed together) and 10% for validation (100 images from Kvasir and 62 images from CVC-ClinicDB), and the remaining images were used for testing. Furthermore, due to the correlation between a set of colonoscopy images from the same patient, to minimize errors, we ensured that during the dataset partitioning, 1 to 4 images were randomly extracted from each patient's colonoscopy images as the test set. The specific data partitioning is illustrated in Table 1.

Table 1 Datasets used in our experiments

Dataset	Images	Input size	Train	Valid	Test
Kvasir-SEG	1000	Variable	800	100	100
CVC-ClinicDB	612	384×288	488	62	62
Endoscene	912	574×500	-	-	60
CVC-ColonDB	380	574×500	-	-	380
ETIS	196	1225×966	-	-	196

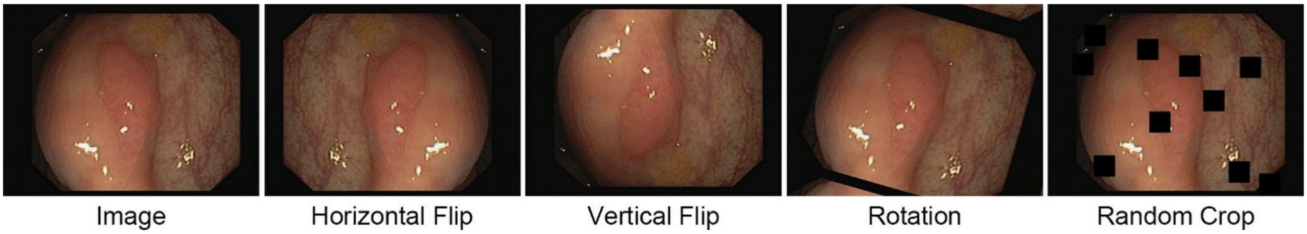

Fig. 1 Image data enhancement

Data Preprocessing

In this study, the number of polyp images available for training the deep learning model was only 1288, which is significantly insufficient for training a robust model and prone to overfitting. Therefore, during the training phase, we employed data augmentation techniques to enhance the dataset by applying various operations, including random horizontal flipping, random vertical flipping, random rotation of 35 degrees, and random cropping. These augmentation methods were utilized to mitigate the limitations caused by the limited training data and enhance the generalization capability of the model.

Specifically, for data augmentation, we applied the horizontal flip function from the “albumentations” library in Python to perform horizontal flipping on a batch of data. The probability parameter, p, of the horizontal flip function was set to 0.5, meaning that during each training iteration, 50% of the images were randomly selected for horizontal flipping. The same settings were applied to random vertical flipping and random rotation.

Regarding random cropping, this operation involved randomly erasing regions within the image. In our study, we utilized random erasing with 10 rectangular regions of size 32×32 pixels each. The probability parameter, p, for this operation was also set to 0.5. The specific visual effects of these augmentation techniques can be observed in Fig. 1.

Methods

The proposed automatic polyp image segmentation model is named ECTransNet. This model effectively integrates multi-scale polyp boundary information and enhances feature representations to achieve more precise and accurate polyp segmentation maps. Figure 2 illustrates the overall architecture of ECTransNet, which includes Res2Net [31],

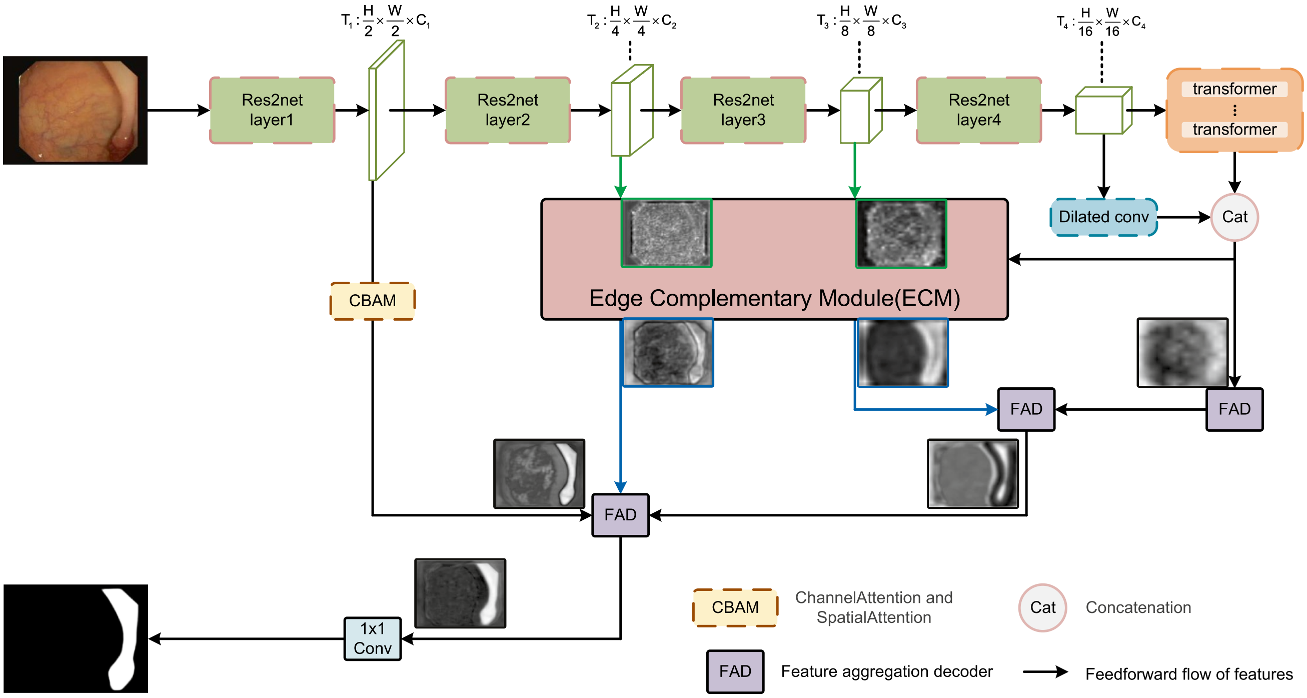
$T_1 : \frac{H}{2} \times \frac{W}{2} \times C_1$
$T_2 : \frac{H}{4} \times \frac{W}{4} \times C_2$
$T_3 : \frac{H}{8} \times \frac{W}{8} \times C_3$
$T_4 : \frac{H}{16} \times \frac{W}{16} \times C_4$
Res2net layer1
Res2net layer2
Res2net layer3
Res2net layer4
transformer
transformer
Dilated conv
Cat
Edge Complementary Module(ECM)
CBAM
FAD
FAD
FAD
1x1 Conv
CBAM
ChannelAttention and SpatialAttention
Cat
Concatenation
FAD
Feature aggregation decoder
Feedforward flow of features

Fig. 2 The proposed ECTransNet architectures

transformer encoder, edge complementary module (ECM), convolutional block attention module (CBAM) [32], and feature aggregation module (FAD). In the following sections, we will provide detailed descriptions of these components and discuss how they contribute to the effectiveness of the ECTransNet.

Encoder Module

ECTransNet is a network based on the Res2Net encoder for an input polyp image of size h×w. The backbone network of Res2Net extracts four levels of resolution features $\{f_i, i = 1, 2, 3, 4\}$, resolution $[h/2^i, w/2^i]$ of the image. We divide f_i into low-level feature $\{f_i, i = 1\}$, edge features $\{f_i, i = 2, 3\}$, and position feature $\{f_i, i = 4\}$. We introduce the transformer encoder and dilated convolution layer as the bottleneck layer of the model to reduce the dimensionality of the features and increase the nonlinear expression capability of the model. The position feature $\{f_i, i = 4\}$ is fed into the bottleneck layer to obtain a global feature map of the polyp F_g.

Res2Net is an improved multi-scale feature extraction network built upon the foundation of ResNet [33]. It introduces hierarchical residual connections within a single residual block, enabling each network layer to capture features from multiple scales and receptive fields. Res2Net has demonstrated superior performance compared to ResNet in tasks such as image classification, semantic segmentation, and object detection. Moreover, it can be seamlessly integrated with other neural network models, offering convenience and flexibility in various applications. As shown in Fig. 3b, the input image matrix is evenly divided into s parts at the channel level, denoted as x_i, where $1 \leq i \leq s$. In the figure, 1×1 Conv represents a 1×1convolutional block, while 3×3 represents a 3×3 convolutional block. The feature formula is represented as follows:

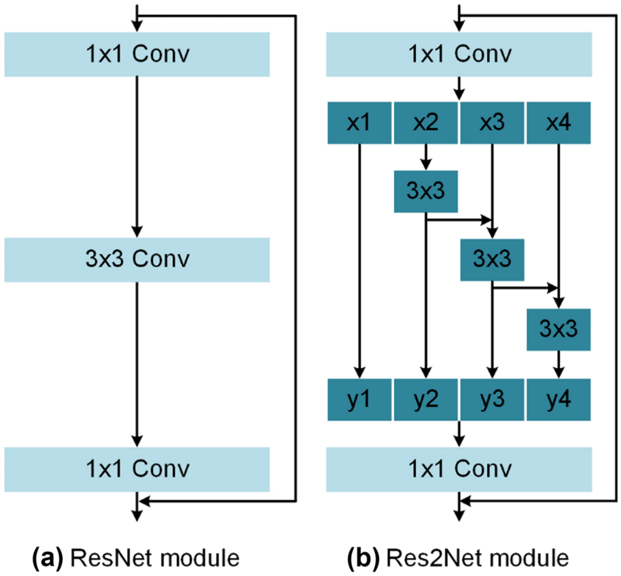

Fig. 3 Differences between ResNet and Res2Net

$$y_i = \begin{cases} x_i i = 1; \\ \text{conv}(x_i) i = 2; \\ \text{conv}(x_i + y_{i-1}) 2 < i \leq s. \end{cases}$$

In the formula, conv represents the 3×3 convolution operation. Finally, all y_i are concatenated together to obtain multi-scale features that capture finer-grained representations.

Transformer is a neural network model based on self-attention mechanism that has made significant contributions in the field of natural language processing (NLP). Inspired by its success in NLP, researchers have applied the transformer to visual data such as images and videos, achieving impressive results and making it a prominent research topic in the field of computer vision. Transformer mainly consists of an encoder and a decoder, which rely entirely on self-attention mechanisms to capture the global dependencies in the sequence. Figure 4 shows the structure of the transformer encoder, which is composed of stacked layers as shown in Fig. 4. Each layer has two sub-layers: the first sub-layer includes a multi-head attention, a residual connection, and a layer normalization, and the second sub-layer includes a feed-forward neural network layer, a residual connection, and a layer normalization.

The essence of the multi-head attention layer lies in performing multiple independent attention computations, enabling the model to capture different features. Multiple heads may focus on different aspects, enhancing the model's expressive capacity. Residual connections address the issues of gradient vanishing and network degradation by adding the input and output of the network. Layer normalization accelerates model convergence and acts as a regularization technique. The feed-forward neural network consists of multiple linear transformations and nonlinear functions, allowing nonlinear transformations of features within sequences and improving the model's expressive power and generalization ability.

Edge Complementary Module and Feature Aggregation Decoder

The ECM plays a crucial role in extracting polyp boundary information, allowing for the complementary fusion of information from different scales. It effectively captures the boundary information cues of the polyps, enabling the generation of more detailed polyp segmentation maps.

As illustrated in Fig. 5a, the ECM consists of two parallel sets of dilated convolutions [34] with dilation rates of 1, 3, and 6, respectively.

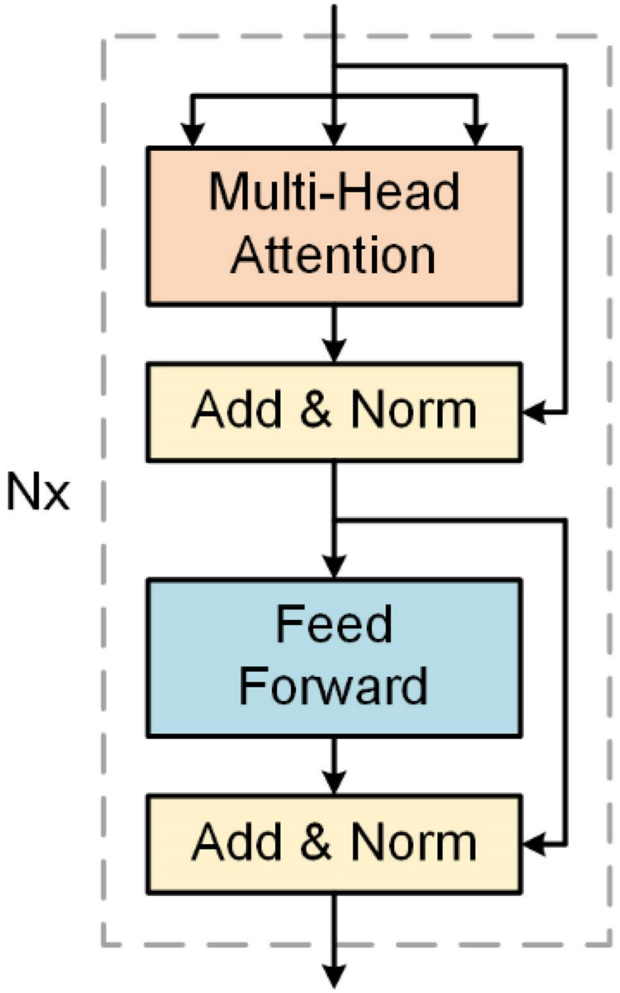

Fig. 4 Transformer encoder

If we represent the operation of a 3 × 3 dilated convolution, followed by batch normalization and ReLU, as CBR(·), then the feature map can be expressed as follows:

$$f_{2,a} = \mathrm{CBR}\left(f_{2,r=1}\right) \oplus \mathrm{CBR}\left(f_{2,r=3}\right) \oplus \mathrm{CBR}\left(f_{2,r=6}\right)$$

$$f_{3,a} = \mathrm{CBR}\left(f_{3,r=1}\right) \oplus \mathrm{CBR}\left(f_{3,r=3}\right) \oplus \mathrm{CBR}\left(f_{3,r=6}\right)$$

Here, $\oplus$ represents the concatenation operation. $f_{2,a}$ and $f_{3,a}$ are feature maps that have been fused with different dimensions, providing richer feature information. Next, the difference feature matrix between the global feature map F_g and $f_{3,a}$ is computed, as well as the difference feature matrix between $f_{2,a}$ and $f_{3,a}$:

$$f_{2,b} = |f_{3,a} - f_{2,a}|$$

$$f_{3,b} = |F_g - f_{3,a}|$$

Specifically, the global feature map F_g is subtracted element-wise from the local feature map $f_{3,a}$, and the absolute value of the resulting matrix is computed to obtain the difference feature $f_{3,b}$. Similarly, the local feature maps $f_{3,a}$ and $f_{2,a}$ undergo element-wise subtraction, and the absolute value of the resulting matrix is computed to obtain the difference feature $f_{2,b}$.

The difference feature maps $f_{2,b}$ and $f_{3,b}$ contain information about the feature disparities between images of different resolutions. Shallow-level features extracted by convolutional neural networks often contain abundant local details such as color, texture, edges, and corners. As the network goes deeper, the receptive field of the network expands, enabling the extraction of more abstract features such as objects, scenes, and positional information in the images. However, the perception of fine details diminishes. Therefore, by performing difference calculations on feature maps extracted by shallow-level networks, we can extract boundary information between polyps and the background, leading to more precise polyp localization and reducing over-segmentation and under-segmentation issues. Further difference calculations are performed on $f_{2,b}$ and $f_{3,b}$:

$$f_{2,c} = |f_{3,b} - f_{2,b}|$$

In order to effectively utilize the boundary information of polyps among different resolutions, we perform convolutional extraction of useful information on $f_{2,b}$ and $f_{2,c}$, followed by matrix element addition operation. This process reduces the loss of critical information and yields more accurate detailed features of the polyps. The feature map obtained after f_2 and f_3 undergo the ECM is illustrated as follows:

$$f_{e2} = \mathrm{CBR}\left(f_{2,b}\right) + \mathrm{CBR}\left(f_{2,c}\right)$$

$$f_{e3} = \mathrm{CBR}\left(f_{3,b}\right)$$

The FAD enables more effective utilization of the extracted feature maps from the encoder and edge complementation modules. Its structure is depicted in Fig. 5b. It performs adaptive fusion of features from different levels, enabling the restoration of polyp location information while preserving more precise local detail information. Starting from the global feature map F_g, both F_g and f_3 are fed into the FAD, resulting in the generation of the decoded map D_1:

$$D_1 = RB(F_g \oplus f_3)$$

Next, D_1, f_{e3}, and f_2 enter the FAD to progressively restore the boundaries and detailed information of the polyp, resulting in D_2:

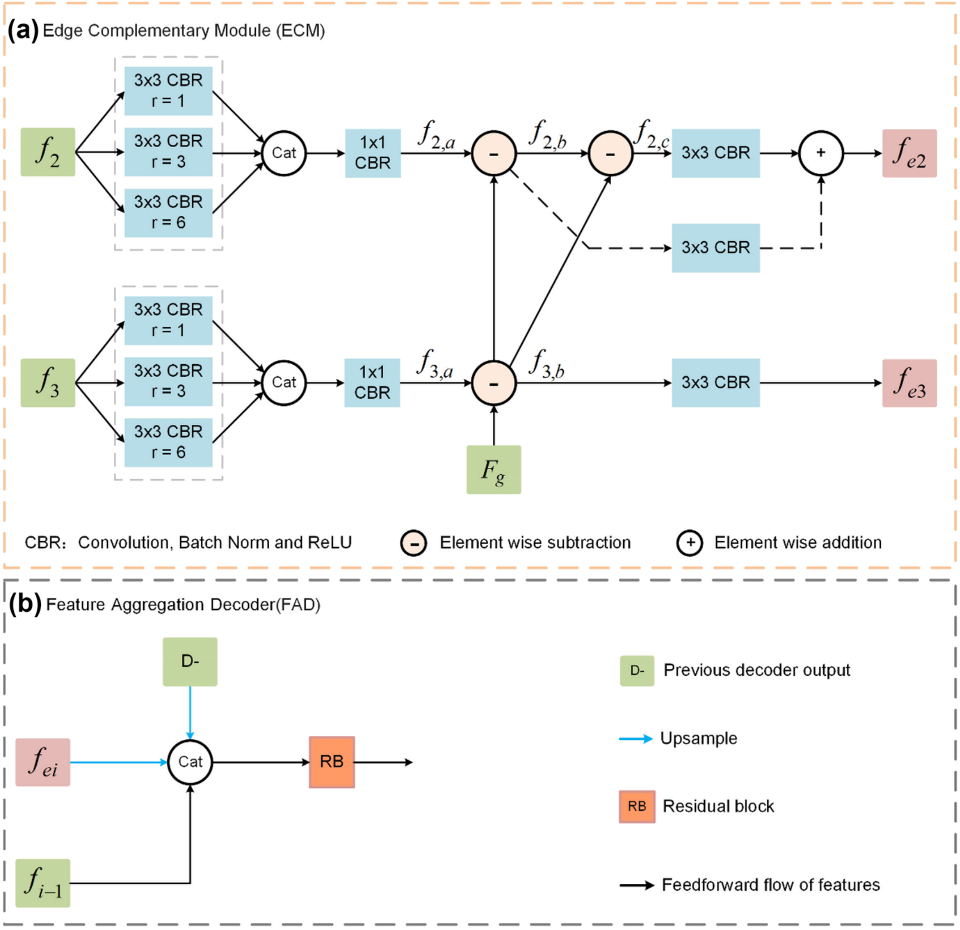

Fig. 5 The details of ECM and FAD

$$D_2 = RB(D_1 \oplus f_{e3} \oplus f_2)$$

Subsequently, the fusion process is performed by combining D_2, f_{e2}, and f_1:

$$D_3 = RB(D_2 \oplus f_{e2} \oplus f_1)$$

Finally, D_3 is employed to generate an accurate segmentation map of the polyp. Here, $\oplus$ denotes the concatenation operation, and RB represents the residual module. Additionally, prior to entering the FAD, f_1 undergoes CBAM, which is an attention module composed of channel attention and spatial attention. It enables the inference of attention maps in both channel and spatial dimensions, which are then fused with the feature maps to adaptively optimize the importance of features. As f_1 corresponds to the shallow-level features of the image, it possesses a high resolution and contains abundant local details of the polyp. However, it also carries a significant amount of noise, which can affect the final polyp segmentation map.

Therefore, CBAM is employed to filter out irrelevant feature information from f_1, enabling the utilization of local details while simultaneously reducing the impact of background noise and other disturbances on the final segmentation map.

Loss Function

Our approach for model supervision involved the implementation of binary cross entropy (BCE) loss and dice loss [35]. BCEloss is a loss function applicable to binary classification tasks, designed to quantify the disparity between target values and predicted values. Its mathematical expression is presented as follows:

$$L_{BCE} = -w * (Y * \log(X) + (1 - Y) * \log(1 - X))$$

where X is the model prediction value, Y is the label truth value, w is the weight value, and the default is 1.

The dice loss function was proposed to address the issue of imbalanced positive and negative samples within a dataset. Given the limited acquisition methods for colonoscopy polyp datasets and the significant disparities often present between positive and negative samples, employing only the BCE loss function would lead to a model that predominantly predicts the majority class. The dice loss function serves as a measure of similarity between two samples, with values ranging from 0 to 1. Larger values indicate higher similarity between samples. By utilizing both BCE loss and dice loss

Table 2 Quantitative results on Kvasir-SEG datasets

Method	Backbone	mDice	mIoU	Recall	Precision	*P*-values
Dataset: Kvasir-SEG						
U-Net [10]	-	0.818 (0.805–0.834)	0.743 (0.715–0.763)	0.807 (0.781–0.835)	0.902 (0.887–0.915)	7.59e-05
U-Net++ [11]	-	0.839 (0.827–0.851)	0.772 (0.751–0.784)	0.861 (0.853–0.873)	0.895 (0.885–0.902)	9.74e-05
DeepLabV3+ [36]	Xception	0.890 (0.885–0.896)	0.831 (0.828–0.834)	0.885 (0.881–0.889)	0.920 (0.918–0.923)	5.31e-03
PraNet [15]	Res2Net50	0.896 (0.891–0.898)	0.838 (0.835–0.841)	**0.894** (0.891–0.897)	0.921 (0.918–0.923)	1.53e-02
TGANet [23]	ResNet50	0.896 (0.895–0.900)	0.839 (0.836–0.843)	0.890 (0.887–0.897)	0.932 (0.930–0.934)	2.19e-02
ECTransNet (ours)	Res2Net50	**0.901** (0.898–0.902)	**0.847** (0.846–0.849)	0.890 (0.888–0.891)	**0.946** (0.945–0.949)	-

together, they complement each other and enhance the accuracy and robustness of segmentation. The specific formula for dice loss is as follows:

$$L_{Dice} = 1 - \frac{2|X \cap Y|}{|X| + |Y|}$$

Thus, the final loss function used for the decoder output takes the form of the following:

$$Loss = L_{BCE} + L_{Dice}$$

Results

Implementation Details and Evaluation Metrics

We built our model with PyTorch and accelerated it with an NVIDIA RTX3090 GPU. We use AdamW optimizer with initial learning rate set to 0.0001 and batch size of 16. Training is performed for a total of 200 epochs. Our loss function is a combination of dice loss and BCE loss. During training, we scaled the images to 256×256 and enhanced the data with random horizontal and vertical flips, rotations, and random cropping operations. For quantitative evaluation, we use some common metrics: average dice, average IoU, recall, and precision.

Experiments on the Public Polyp Benchmarks

We compared ECTransNet with several state-of-the-art methods, including U-Net, U-Net++, DeepLabV3+, PraNet, and TGANet. These algorithms are widely used for polyp segmentation and general medical image segmentation. The results of these methods in Tables 2, 3, 4, 5, and 6 were obtained by re-executing their publicly released code or models, while the partitioning of their datasets was identical to our approach. The quantitative results are presented in Tables 2, 3, 4, 5, and 6. For convenience, we have highlighted the best results for each evaluation metric in bold text.

Table 3 Quantitative results on CVC-ClinicDB datasets

Method	Backbone	mDice	mIoU	Recall	Precision	*P*-values
Dataset: CVC-ClinicDB						
U-Net [10]	-	0.831 (0.825–0.837)	0.783 (0.779–0.785)	0.867 (0.863–0.872)	0.885 (0.882–0.889)	7.65e-07
U-Net++ [11]	-	0.852 (0.849–0.855)	0.809 (0.807–0.811)	0.907 (0.904–0.912)	0.887 (0.884–0.891)	1.91e-07
DeepLabV3+ [36]	Xception	0.893 (0.891–0.895)	0.843 (0.841–0.845)	0.894 (0.892–0.896)	0.919 (0.915–0.923)	6.65e-07
PraNet [15]	Res2Net50	0.895 (0.892–0.898)	0.853 (0.849–0.857)	0.912 (0.907–0.915)	0.899 (0.890–0.903)	2.43e-06
TGANet [23]	ResNet50	0.922 (0.915–0.928)	0.871 (0.865–0.877)	**0.934** (0.930–0.938)	0.924 (0.917–0.930)	7.83e-01
ECTransNet (ours)	Res2Net50	**0.923** (0.921–0.924)	**0.878** (0.875–0.880)	0.931 (0.928–0.933)	**0.933** (0.931–0.935)	-

Table 4 Quantitative results on Endoscene datasets

Method	Backbone	mDice	mIoU	Recall	Precision	*P*-values
Dataset: Endoscene						
U-Net [10]	-	0.710 (0.705–0.713)	0.630 (0.625–0.634)	0.706 (0.705–0.709)	0.864 (0.861–0.868)	1.49e-07
U-Net++ [11]	-	0.762 (0.758–0.765)	0.690 (0.687–0.692)	0.756 (0.753–0.759)	0.861 (0.859–0.862)	4.07e-07
DeepLabV3+ [36]	Xception	0.864 (0.861–0.867)	0.788 (0.785–0.791)	0.923 (0.921–0.925)	0.849 (0.847–0.851)	2.25e-05
PraNet [15]	Res2Net50	0.866 (0.864–0.868)	0.795 (0.793–0.798)	0.903 (0.901–0.905)	0.871 (0.868–0.875)	3.31e-05
TGANet [23]	ResNet50	0.878 (0.873–0.882)	0.806 (0.799–0.810)	**0.965** (0.961–0.967)	0.828 (0.821–0.832)	1.71e-04
ECTransNet (ours)	Res2Net50	**0.907** (0.901–0.910)	**0.840** (0.831–0.842)	0.954 (0.950–0.955)	**0.877** (0.866–0.882)	-

In addition, we performed a paired *t*-test between the dice scores obtained by ECTransNet and those obtained by other methods. The table below reports the *P*-values of the paired *t*-test. When the *P*-value is less than 0.05, it indicates significant differences in the dice scores between ECTransNet and the corresponding methods.

Results on Kvasir-SEG In Table 2, we present the quantitative evaluation of ECTransNet, while the qualitative results are reported in Fig. 6. From the quantitative results, it is evident that our method outperforms all other state-of-the-art (SOTA) methods in metrics other than recall. Compared to PraNet, our method demonstrates an improvement of 0.5% in mDice, 0.9% in mIoU, and 2.5% in precision, while experiencing a decrease of 0.4% in recall. The actual segmentation effect is shown in Fig. 6. It can be observed that our method accurately captures the boundary information of polyps in images containing multiple polyps from the Kvasir-SEG dataset, yielding precise segmentation of the polyp count. In contrast, other methods often exhibit instances of missed detections.

Results on CVC-ClinicDB Quantitative results of CVC-ClinicDB are presented in Table 3. Compared to the highly competitive TGANet, our method demonstrates improvements of 0.1% in mDIce, 0.7% in mIoU, and 0.9% in Precision, while experiencing a decrease of 0.3% in recall score. The segmentation results of ClinicDB, as shown in Fig. 6, reveal that ECTransNet accurately segments polyps, while TGANet exhibits over-segmentation. Although PraNet and DeepLabV3+ accurately locate the polyps, their handling of details is insufficient. The ECM within our ECTransNet effectively leverages the differences between features at various resolutions, capturing the edge information of polyps, facilitating multi-scale feature complementarity, and obtaining more precise local features of polyps.

Results on Endoscene Quantitative results on the Endoscene dataset are reported in Table 4. It is evident that our ECTransNet outperforms other SOTA methods in terms of mDice and mIoU, exhibiting significant improvements of 2.9% and 3.4%, respectively, compared to TGANet. Our method

Table 5 Quantitative results on CVC-ColonDB datasets

Method	Backbone	mDice	mIoU	Recall	Precision	*P*-values
Dataset: CVC- ColonDB						
U-Net [10]	-	0.627 (0.622–0.629)	0.562 (0.560–0.565)	0.654 (0.651–0.659)	0.812 (0.809–0.814)	3.86e-08
U-Net++ [11]	-	0.628 (0.625–0.632)	0.578 (0.572–0.581)	0.721 (0.718–0.723)	0.785 (0.781–0.787)	5.03e-08
DeepLabV3+ [36]	Xception	0.727 (0.721–0.732)	0.647 (0.643–0.651)	0.752 (0.749–0.757)	0.824 (0.822–0.828)	2.96e-05
PraNet [15]	Res2Net50	0.713 (0.710–0.716)	0.642 (0.639–0.645)	0.710 (0.708–0.712)	0.808 (0.805–0.811)	1.7e-06
TGANet [23]	ResNet50	0.752 (0.747–0.757)	0.674 (0.670–0.677)	**0.787** (0.783–0.793)	0.798 (0.796–0.800)	3.05e-03
ECTransNet (ours)	Res2Net50	**0.766** (0.764–0.768)	**0.687** (0.683–0.689)	0.765 (0.760–0.769)	**0.873** (0.870–0.879)	-

Table 6 Quantitative results on ETIS datasets

Method	Backbone	mDice	mIoU	Recall	Precision	*P*-values
Dataset: ETIS						
U-Net [10]	-	0.414 (0.411–0.419)	0.368 (0.359–0.372)	0.467 (0.458–0.471)	0.717 (0.706–0.721)	1.74e-06
U-Net++ [11]	-	0.423 (0.418–0.427)	0.352 (0.341–0.355)	0.406 (0.394–0.413)	0.685 (0.681–0.688)	1.38e-06
DeepLabV3+ [36]	Xception	0.627 (0.621–0.634)	0.567 (0.554–0.569)	0.701 (0.695–0.709)	0.704 (0.697–0.712)	1.40e-04
PraNet [15]	Res2Net50	0.626 (0.623–0.631)	0.570 (0.562–0.573)	0.698 (0.695–0.703)	0.752 (0.748–0.757)	1.32e-04
TGANet [23]	ResNet50	0.703 (0.696–0.709)	**0.674** (0.670–0.677)	**0.788** (0.784–0.793)	0.798 (0.796–0.800)	2.36e-02
ECTransNet (ours)	Res2Net50	**0.728** (0.719–0.733)	0.655 (0.649–0.660)	0.769 (0.758–0.772)	**0.800** (0.790–0.812)	-

achieves a recall score 1.1% lower than TGANet, while surpassing TGANet by 4.9% in precision. Figure 6 demonstrates that for small-sized polyp images in the Endoscene dataset, our ECTransNet performs well in segmentation. The FAD module in ECTransNet effectively integrates high-level and low-level features, benefiting from the accurate polyp-specific local detail feature maps generated by ECM. By fusing and decoding feature maps of different scales, FAD generates precise segmentation prediction maps.

Results on CVC-ColonDB In this study, we exclusively employed CVC-ColonDB as our test dataset, which was not involved in the model training process and thus remains invisible to the model. As depicted in Table 5, our approach demonstrates superior generalization capabilities compared to other SOTA methods. Specifically, when compared to the highly competitive TGANet, our method exhibits respective improvements of 1.4% and 1.3% in mDice and mIoU metrics. As illustrated in Fig. 6, our method accurately identifies

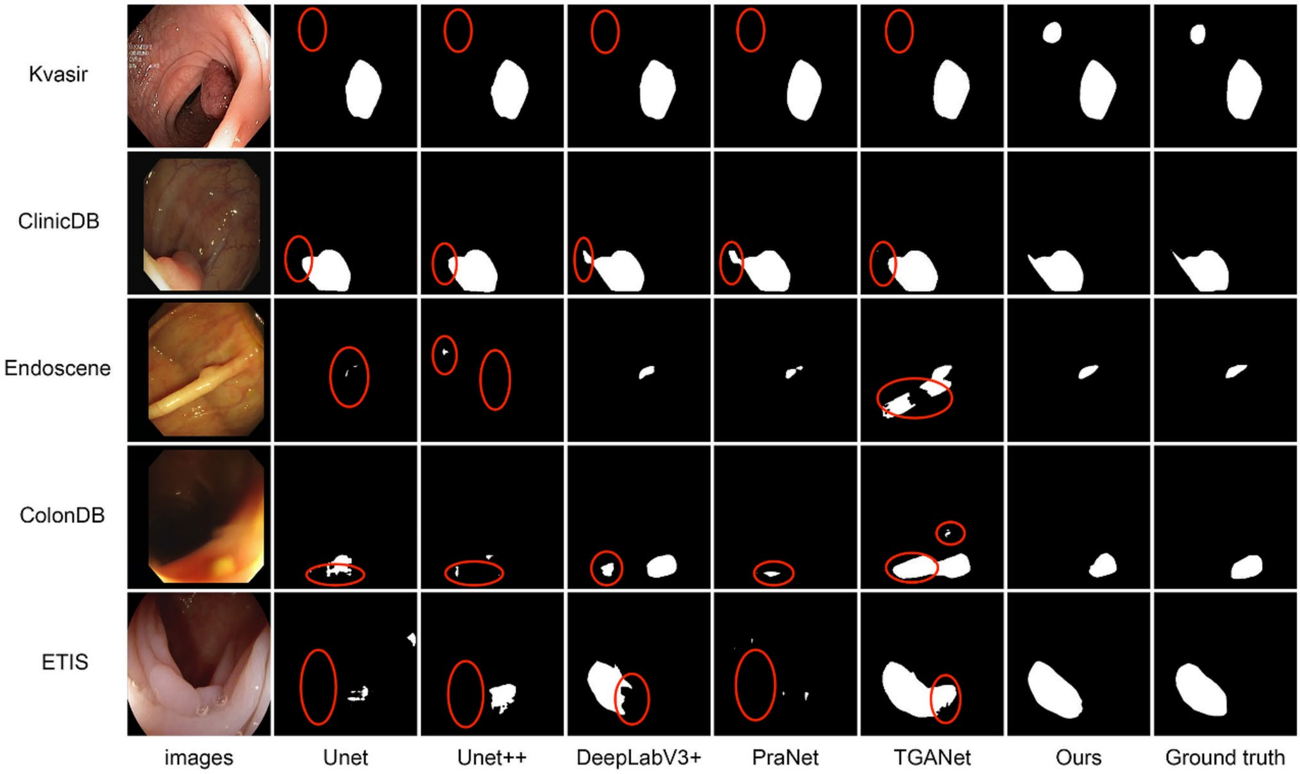

Fig. 6 Visual comparison of polyp segmentation results

polyps and background in the ColonDB dataset, whereas the segmentation performance of alternative methods is considerably inferior. This discrepancy arises from the utilization of transformer encoder in ECTransNet, which encode global semantic relationships among pixels in the polyp's global feature map, enabling enhanced discrimination between polyps and the background.

Results on ETIS The ETIS dataset is also invisible to the model, and it is a challenging dataset because most of its polyp images vary greatly in morphology. As shown in Table 6, our ECTransNet achieves the highest mDice and precision scores, while TGANet achieves the highest mIoU and recall scores. Figure 6 illustrates the segmentation prediction maps for the ETIS dataset, highlighting the performance of different methods. It is evident that our ECTransNet maintains robustness when confronted with polyp images from various sources.

Ablation Study

In order to investigate the roles of each component within ECTransNet, we conducted ablation experiments on the Kvasir-SEG dataset to validate the effectiveness of the transformer encoder, ECM, FAD, and CBAM. By sequentially removing these modules from ECTransNet while preserving the integrity of other modules, we trained the models and observed their impact on model performance. The quantitative results of our ablation experiments are presented in Table 7. Firstly, we eliminated the transformer encoder, resulting in a decrease of 0.7% in mDice score and 0.6% in mIoU score. This indicates that the pixel-wise global semantic relationships constructed by the transformer encoder contribute to the segmentation performance of the model.

Subsequently, we removed the ECM, which had a notable impact on the model. The mDice and mIoU scores were 0.878 and 0.838, respectively, representing a decrease of 2.3% and 0.9% compared to ECTransNet. This suggests that the complementary fusion of feature maps across different scales by ECM is beneficial for the final segmentation prediction. It allows for a more effective utilization of the differences in features at different resolutions, resulting in feature maps that are richer and more accurate in information. Then, we removed the FAD and utilized only upsampling for segmentation map restoration. It was observed that the performance of the model was significantly influenced by FAD. After removing FAD, the mDice and mIoU scores of ECTransNet decreased by 2% and 1.5%, respectively. FAD facilitates the fusion of high-level and low-level features and utilizes feature maps of different scales to restore the segmentation prediction map. Additionally, the residual modules within FAD enhance the accuracy of feature extraction and improve the model's ability to restore boundary information of polyps.

To fuse the low-level features, we employed the CBAM for information filtering. The CBAM consists of both channel attention and spatial attention mechanisms. Since low-level feature maps often contain abundant local details and irrelevant noise, we utilized channel attention to focus on the channels containing useful information in the feature maps. Subsequently, spatial attention was employed to highlight the local features of polyps. Therefore, the CBAM aids the model in effectively predicting the local detailed features of polyps and obtaining more precise segmentation prediction maps. As shown in Table 7, after removing CBAM, the model experienced a decrease of 0.5% and 0.2% in mDice and mIoU scores, respectively, on the Kvasir-SEG dataset.

Discussion

The challenges in polyp image segmentation primarily stem from the similarity between polyps and the background color, substantial morphological variations, and the blurred boundaries between polyps and the mucosa. Deep learning models often result in varying degrees of over-segmentation or under-segmentation. U-Net addresses these challenges by utilizing skip connections to integrate high- and low-level features, thereby recovering lost information during the

Table 7 Ablation study for ECTransNet on the Kvasir-SEG

Experiment description	mDice	mIoU	Recall	Precision
ECTransNet (ours)	0.901 (0.898–0.902)	0.847 (0.846–0.849)	0.890 (0.887–0.891)	0.946 (0.945–0.949)
Without transformer encoder	0.893 (0.890–0.895)	0.841 (0.837–0.843)	0.875 (0.872–0.881)	0.931 (0.928–0.934)
Without ECM	0.878 (0.874–0.880)	0.838 (0.835–0.841)	0.879 (0.877–0.883)	0.941 (0.938–0.943)
Without FAD	0.881 (0.876–0.885)	0.832 (0.830–0.835)	0.884 (0.881–0.886)	0.927 (0.925–0.929)
Without CBAM	0.896 (0.892–0.898)	0.845 (0.842–0.847)	0.889 (0.886–0.892)	0.939 (0.937–0.942)

downsampling process. However, a semantic gap commonly exists between high- and low-level features. The presence of a substantial amount of noise in the low-level features of polyp images often leads to a notable impact on the final segmentation map. U-Net++, an extension of U-Net, removes the long connections and introduces a multitude of short connections. While it integrates features from various hierarchical levels, its capacity to establish semantic relationships among pixels is limited, resulting in the omission of multiple polyps.

DeeplabV3+ employs a feature pyramid to integrate the global feature maps of polyps, resulting in improved segmentation performance for polyp images. However, due to its reliance on only one skip connection during the decoding stage to incorporate local feature maps, it suffers from the loss of critical fine-grained details, leading to a deterioration in its ability to accurately segment the edges of polyps. The aforementioned shortcomings lead to inferior segmentation results for both U-Net and DeepLabV3+, as depicted in Tables 2, 3, 4, 5, and 6.

PraNet leverages a reverse attention module to indirectly extract contour cues. However, it still falls short in adequately capturing segmentation details and effectively handling the segmentation of multiple polyps. While it improves the accuracy of polyp segmentation, it tends to exhibit omission when dealing with multiple target polyps, as observed in Fig. 6 (Kvasir-SEG dataset). On the other hand, TGANet employs text-guided attention to address the challenge of varying sizes and quantities of polyps. This approach enables the network to learn additional feature representations for distinguishing polyps of different sizes. Nevertheless, TGANet also lacks attention towards the semantic information among pixels, which limits its ability to capture pixel-level semantic details.

We introduce transformer encoder into ECTransNet, leveraging a global self-attention mechanism to extract the semantic relationship between polyp pixels and background pixels, thereby obtaining precise spatial information about the polyps' locations. Similarly, our ECTransNet also adopts skip connection structures. Additionally, we construct an ECM to capture feature differences among different levels of local feature maps, highlighting the local detailed information of polyps in the feature maps. We incorporate these features into the decoding process to generate more accurate segmentation maps. According to Tables 2 and 3, it can be observed that ECTransNet achieves the best mDice and mIoU scores on Kvasir-SEG and CVC-ClinicDB datasets, with values of 0.901 and 0.847 and 0.923 and 0.878, respectively. These results outperform the highly competitive TGANet. In order to better fuse high- and low-level features, we construct a residual-based feature aggregation decoder (FAD). This decoder takes polyp feature maps at multiple scales as input, allowing for the restoration of accurate polyp locations while preserving more local details. As shown in Fig. 6, it is evident that our ECTransNet retains more local details compared to other methods, particularly on the ClinicDB dataset.

In real clinical environments, the performance of deep learning-based segmentation methods can be influenced by imaging protocols and variations among different patients. In such cases, models that can adapt to different source datasets are considered more generalizable. In Tables 4, 5, and 6, the Endoscene, CVC-ColonDB, and ETIS datasets are considered invisible to the model since they were not included in the model's training process. In the Endoscene and CVC-ColonDB datasets, our method achieves the best mDice and mIoU scores of 0.907 and 0.840 and 0.766 and 0.687, respectively. In the ETIS dataset, ECTransNet achieves the best mDice score of 0.728, while the most competitive TGANet obtains the best mIoU score of 0.674. These results indicate that our ECTransNet is more generalizable. This can be attributed to the effectiveness of our ECM in complementing the differences in polyp feature information at various resolutions, preserving the representative information of polyps. Furthermore, the precise restoration of polyps is achieved through the utilization of FAD, contributing to the improved performance of our approach.

Additionally, we conducted ablation experiments to validate the effectiveness of each component in ECTransNet. We removed the corresponding modules from the model while keeping the remaining components intact and trained the modified model. Table 7 demonstrates that when the transformer encoder is removed, there is a decrease of 0.8% in mDice and 0.6% in mIoU. This indicates that the transformer encoder plays a crucial role in extracting the semantic relationships for feature extraction in the model. It enables more accurate acquisition of spatial information regarding the polyps and establishes global semantic relationships among polyp pixels. Next, we removed the ECM, resulting in a decrease of 2.3% and 0.9% in mDice and mIoU, respectively. The model's performance experienced a significant decline. The ECM module performs differential computations on feature maps of varying resolutions and integrates complementary information to emphasize the boundaries of polyps, facilitating better extraction of local features. This is crucial for enhancing the accuracy of polyp segmentation.

Similarly, Table 7 demonstrates that the removal of FAD had a considerable impact on the model's performance, resulting in a noteworthy decrease of 2% and 1.5% in mDice and mIoU, respectively. Additionally, recall and precision also exhibited varying degrees of decline. The FAD module integrates feature maps from both ECM and Res2Net. The feature maps extracted by ECM contain essential information related to polyp boundaries, while Res2Net's feature maps encompass detailed polyp information. FAD utilizes feature maps of different scales to minimize the loss of critical polyp information and performs fusion of high- and low-level features, thereby generating precise polyp segmentation maps.

Furthermore, in the integration of low-level features, we employ CBAM to perform feature filtering on the low-level features. Although low-level features contain rich detailed characteristics, they also exhibit a considerable amount of irrelevant noise. Hence, we utilize CBAM to filter these features. The channel attention and spatial attention mechanisms within CBAM enable the inference of attention maps in both the channel and spatial dimensions. Subsequently, these attention maps are fused with the feature maps, allowing for the adaptive optimization of feature importance. In Table 7, it can be observed that after removing the CBAM module, the mDice and mIoU metrics decreased by 0.5% and 0.2%, respectively. Throughout the ablation experiment, the removal of ECM and FAD led to a significant decline in the model's performance. Furthermore, the other modules also exerted varying degrees of influence on the model.

One limitation of our study is that although our method accurately localizes the majority of polyp positions for segmentation, it performs poorly on images with low contrast and severe artifacts. This issue has become a focal point for our future research endeavors. We aim to contribute novel insights to the task of polyp image segmentation. Moving forward, we plan to address the aforementioned challenges by refining the design of our network architecture to handle more complex scenarios.

Author Contribution Weikang Liu (first author): conceptualization, methodology, software, data curation, and writing, original draft; Zhigang Li (corresponding author): conceptualization, supervision, funding acquisition, resources, and writing, review and editing; Chunyang Li: data curation, investigation, validation, and formal analysis; Hongyan Gao: visualization and writing, review and editing.

Funding This work was supported in part by the Liaoning Provincial Education Department's Service Local Project, grant/award number 2020FWDF04, and the Scientific Research Fund of Liaoning Provincial Education Department of China, grant/award number: 2020LNZD05.

Data Availability The codes are available online at https://github.com/liuweikang1112/polyp_seg.

Declarations

Ethics Approval This research study was conducted retrospectively from data obtained for clinical purposes. We used only data coming from publicly available datasets.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

1. Wang J, Zhang X, Lv P et al. Automatic Liver Segmentation Using EfficientNet and Attention-Based Residual U-Net in CT, Journal of Digital Imaging; 35, pp. 1479–1493, 2022.https://doi.org/10.1007/s10278-022-00668-x.
2. Sun Y, Li Y, Wang P et al. Lesion Segmentation in Gastroscopic Images Using Generative Adversarial Networks, Journal of Digital Imaging; 35, pp. 459–468, 2022.https://doi.org/10.1007/s10278-022-00591-1.
3. Li M, Lian F, Guo S. Multi-scale Selection and Multi-channel Fusion Model for Pancreas Segmentation Using Adversarial Deep Convolutional Nets, Journal of Digital Imaging; 35, pp. 47–55, 2022.https://doi.org/10.1007/s10278-021-00563-x.
4. Sung H, Ferlay J, Siegel R L et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries, CA: A Cancer Journal for Clinicians; 71, pp. 209–249, 2021.https://doi.org/10.3322/caac.21660.
5. Zhao S, Wang S, Pan P et al. Magnitude, Risk Factors, and Factors Associated With Adenoma Miss Rate of Tandem Colonoscopy: A Systematic Review and Meta-analysis, Gastroenterology; 156, pp. 1661–1674.e1611, 2019.https://doi.org/10.1053/j.gastro.2019.01.260.
6. Favoriti P, Carbone G, Greco M et al. Worldwide burden of colorectal cancer: a review, Updates in Surgery; 68, pp. 7–11, 2016.https://doi.org/10.1007/s13304-016-0359-y.
7. FIORI M, MUSÉ P, SAPIRO G. A COMPLETE SYSTEM FOR CANDIDATE POLYPS DETECTION IN VIRTUAL COLONOSCOPY; 28, pp. 1460014, 2014.https://doi.org/10.1142/s0218001414600143.
8. Mamonov A V, Figueiredo I N, Figueiredo P N et al. Automated Polyp Detection in Colon Capsule Endoscopy, IEEE Transactions on Medical Imaging; 33, pp. 1488–1502, 2014.https://doi.org/10.1109/TMI.2014.2314959.
9. Maghsoudi O H. Superpixel based segmentation and classification of polyps in wireless capsule endoscopy. In: 2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). pp. 1–4, 2017.https://doi.org/10.1109/SPMB.2017.8257027.
10. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham, pp. 234–241, 2015.https://doi.org/10.1007/978-3-319-24574-4_28.
11. Zhou Z, Siddiquee M M R, Tajbakhsh N et al. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation, IEEE Transactions on Medical Imaging; 39, pp. 1856–1867, 2018.https://doi.org/10.1109/TMI.2019.2959609.
12. Jha D, Smedsrud P H, Riegler M A et al. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM). pp. 225–2255, 2019.https://doi.org/10.1109/ISM46123.2019.00049.
13. Drozdzal M, Vorontsov E, Chartrand G et al. The Importance of Skip Connections in Biomedical Image Segmentation. In: Deep Learning and Data Labeling for Medical Applications. Cham, pp. 179–187, 2016.https://doi.org/10.1007/978-3-319-46976-8_19.
14. Jha D, Riegler M A, Johansen D et al. DoubleU-Net: A Deep Convolutional Neural Network for Medical Image Segmentation. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). pp. 558–564, 2020.https://doi.org/10.1109/CBMS49503.2020.00111.
15. Fan D-P, Ji G-P, Zhou T et al. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Cham, pp. 263–273, 2020.https://doi.org/10.1007/978-3-030-59725-2_26.
16. Huang C-H, Wu H-Y, Lin Y-L J a p a. Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves

over 0.9 mean dice and 86 fps, 2021.https://doi.org/10.48550/arXiv.2101.07172.

17. Chao P, Kao C Y, Ruan Y et al. HarDNet: A Low Memory Traffic Network. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3551–3560, 2019.https://doi.org/10.1109/ICCV.2019.00365.
18. Shen Y, Jia X, Meng M Q H. HRENet: A Hard Region Enhancement Network for Polyp Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham, pp. 559–568, 2021.https://doi.org/10.1007/978-3-030-87193-2_53.
19. Zhong J, Wang W, Wu H et al. PolypSeg: An Efficient Context-Aware Network for Polyp Segmentation from Colonoscopy Videos. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Cham, pp. 285–294, 2020.https://doi.org/10.1007/978-3-030-59725-2_28.
20. Wei J, Hu Y, Zhang R et al. Shallow Attention Network for Polyp Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham, pp. 699–708, 2021.https://doi.org/10.1007/978-3-030-87193-2_66.
21. Zhao X, Zhang L, Lu H. Automatic Polyp Segmentation via Multi-scale Subtraction Network. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Cham, pp. 120–130, 2021.https://doi.org/10.1007/978-3-030-87193-2_12.
22. Srivastava A, Jha D, Chanda S et al. MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation, IEEE Journal of Biomedical and Health Informatics; 26, pp. 2252–2263, 2022.https://doi.org/10.1109/JBHI.2021.3138024.
23. Tomar N K, Jha D, Bagci U et al. TGANet: Text-Guided Attention for Improved Polyp Segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. Cham, pp. 151–160, 2022.https://doi.org/10.1007/978-3-031-16437-8_15.
24. Hosseini H, Xiao B, Jaiswal M et al. On the Limitation of Convolutional Neural Networks in Recognizing Negative Images. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 352–358, 2017.https://doi.org/10.1109/ICMLA.2017.0-136.
25. Vaswani A, Shazeer N, Parmar N et al. Attention is All you Need. 2017.https://doi.org/10.48550/arXiv.1706.03762
26. Jha D, Smedsrud P H, Riegler M A et al. Kvasir-SEG: A Segmented Polyp Dataset. In: MultiMedia Modeling. Cham, pp. 451–462, 2020.https://doi.org/10.1007/978-3-030-37734-2_37.
27. Bernal J, Sánchez F J, Fernández-Esparrach G et al. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Computerized Medical Imaging and Graphics; 43, pp. 99–111, 2015.https://doi.org/10.1016/j.compmedimag.2015.02.007.
28. Vázquez D, Bernal J, Sánchez F J et al. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images, Journal of Healthcare Engineering; 2017, pp. 4037190, 2017.https://doi.org/10.1155/2017/4037190.
29. Bernal J, Sánchez J, Vilariño F. Towards automatic polyp detection with a polyp appearance model, Pattern Recognition; 45, pp. 3166–3182, 2012.https://doi.org/10.1016/j.patcog.2012.03.002.
30. Silva J, Histace A, Romain O et al. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer, International Journal of Computer Assisted Radiology and Surgery; 9, pp. 283–293, 2014.https://doi.org/10.1007/s11548-013-0926-3.
31. Gao S H, Cheng M M, Zhao K et al. Res2Net: A New Multi-Scale Backbone Architecture, IEEE Transactions on Pattern Analysis and Machine Intelligence; 43, pp. 652–662, 2021.https://doi.org/10.1109/TPAMI.2019.2938758.
32. Woo S, Park J, Lee J-Y et al. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19, 2018.https://doi.org/10.48550/arXiv.1807.06521.
33. He K, Zhang X, Ren S et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778, 2016.https://doi.org/10.1109/CVPR.2016.90
34. Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions. In: ICLR. 2016.https://doi.org/10.48550/arXiv.1511.07122.
35. Milletari F, Navab N, Ahmadi S A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571, 2016.https://doi.org/10.1109/3DV.2016.79.
36. Chen L-C, Zhu Y, Papandreou G et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818, 2018.https://doi.org/10.1007/978-3-030-01234-2_49.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.