

## Article

# Improving Polyp Segmentation with Boundary-Assisted Guidance and Cross-Scale Interaction Fusion Transformer Network

Lincen Jiang <sup>1,2,\*</sup>, Yan Hui <sup>1</sup> , Yuan Fei <sup>1</sup>, Yimu Ji <sup>1</sup> and Tao Zeng <sup>3,\*</sup>

<sup>1</sup> School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 1221045915@njupt.edu.cn (Y.H.); 1321048523@njupt.edu.cn (Y.F.); jjym@njupt.edu.cn (Y.J.)

<sup>2</sup> School of Computer and Software, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

<sup>3</sup> College of Electronic and Optical Engineering & College of Flexible Electronics (Future Technology), Nanjing University of Posts and Telecommunications, Nanjing 210023, China

\* Correspondence: 2022010301@njupt.edu.cn (L.J.); b21021229@njupt.edu.cn (T.Z.)

**Abstract:** Efficient and precise colorectal polyp segmentation has significant implications for screening colorectal polyps. Although network variants derived from the Transformer network have high accuracy in segmenting colorectal polyps with complex shapes, they have two main shortcomings: (1) multi-level semantic information at the output of the encoder may result in information loss during the fusion process and (2) failure to adequately suppress background noise during segmentation. To address these challenges, we propose a cross-scale interaction fusion transformer for polyp segmentation (CIFFormer). Firstly, a novel feature supplement module (FSM) supplements the missing details and explores potential features to enhance the feature representations. Additionally, to mitigate the interference of background noise, we designed a cross-scale interactive fusion module (CIFM) that combines feature information between different layers to obtain more multi-scale and discriminative representative features. Furthermore, a boundary-assisted guidance module (BGM) is proposed to help the segmentation network obtain boundary-enhanced details. Extensive experiments on five typical datasets have demonstrated that CIFFormer has an obvious advantage in segmenting polyps. Specifically, CIFFormer achieved an mDice of 0.925 and an mIoU of 0.875 on the Kvasir-SEG dataset, achieving superior segmentation accuracy to competing methods.

**Keywords:** polyp segmentation; boundary-assisted guidance; cross-scale interaction; transformer



**Citation:** Jiang, L.; Hui, Y.; Fei, Y.; Ji, Y.; Zeng, T. Improving Polyp Segmentation with Boundary-Assisted Guidance and Cross-Scale Interaction Fusion Transformer Network. *Processes* **2024**, *12*, 1030. <https://doi.org/10.3390/pr12051030>

Academic Editor: Chunhui Zhao

Received: 15 April 2024

Revised: 10 May 2024

Accepted: 17 May 2024

Published: 19 May 2024



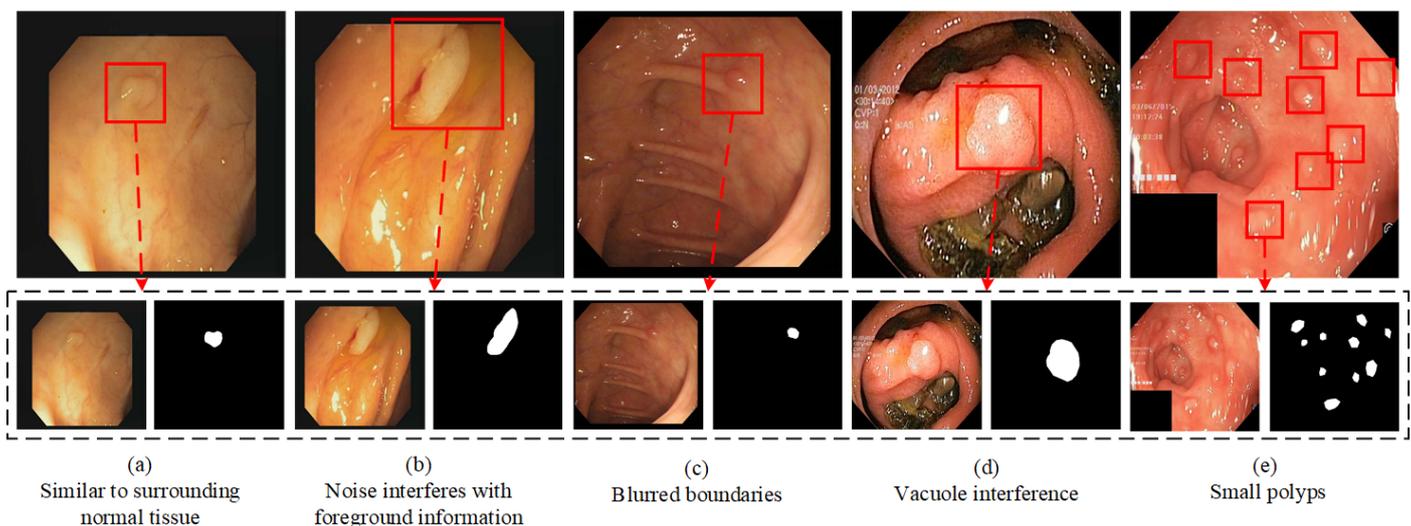
**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the changes in people's dietary habits, the incidence of colorectal cancer caused by the shedding of colorectal polyps is also increasing, threatening people's lives and health [1]. Regular colonoscopy by clinicians can help to detect polyps in time and remove them early, which can block polyps from growing and effectively prevent polyps from becoming cancerous [2]. With the help of medical imaging, automatic segmentation of polyps has become an important auxiliary means of clinical examination. But due to the varying sizes and shapes of polyp tissues, as well as the complex background of the intestinal environment, performing accurate segmentation of polyps very difficult and a daunting task in the field of medical imaging, as shown in Figure 1. Manual segmentation of accurate polyp images will have high resource costs, and it is difficult to accomplish manual segmentation of massive images, which adds a lot of burden to the work of healthcare professionals.

The great success of Transformer [3] has been explored in the field of computer vision. Recently, Vision Transformer (ViT) [4] has been proven to be a simple and extensible framework. Naseer et al. [5] shows that Transformer differs from convolutional neural networks in extracting information in the kernel for weight parameter training, as it instead

uses an attention mechanism to obtain similar features. It adaptively extracts features to train the weight parameters therein through dot product operations, which is able to globally model images and achieve performance comparable to traditional convolutional neural networks on multiple image-recognition tasks. As a result, Transformer has a more powerful generalization capability than CNNs. Subsequent studies have shown that an increasing number of models have been enhanced using Transformer, such as TransUNet [6], TransFuse [7], and Polyp-PVT [8]; these methods use pyramidal Transformer as the model encoder, which enables the models to obtain feature information in more dimensions and improves the segmentation accuracy of Transformer-like models. Transformer-based optimization models still have two problems concerning polyp segmentation: (1) The pyramid Transformer generates features at all levels with a significant distinction, and incorrect fusion methods will result in the loss of useful features. Redundancy and clutter of low-level information contribute to excessively smooth target predictions and the subsequent blurring of boundaries. (2) The image background creates more noise interference, leading to difficulties in polyp lesion feature extraction. Meanwhile, the uneven contrast of the light source during image capture results in some parts of the background being missegmented as lesions. Moreover, with the variants of Transformer, the textural features continue to be mixed and converged, which often leads to distractions for the attention.



**Figure 1.** Five typical types of polyp segmentation challenges.

Advances in deep learning are driving big improvements in polyp image segmentation accuracy. Researchers have used U-Net [9] for polyp segmentation to achieve better segmentation results. Subsequently, some researchers proposed U-Net++ [10], which introduces more upsampling operations and skip connections in the model, making the model able to extract multiple levels of features. Some researchers further proposed PraNet [11], which extracts high-layer features through a complexly structured convolutional module. However, although traditional CNN-based image segmentation models can achieve rough segmentation of polyp images, the segmentation results are not satisfactory. These models lack generalization ability to accurately segment images obtained from different colonoscopy devices.

To solve the above-mentioned difficulties in polyp segmentation, we propose an innovative cross-scale interaction fusion transformer for polyp segmentation (CIFFormer). We choose PVTv2 [12] as the backbone, and the hierarchical multi-scale model improves the expression of polyp features, thus extracting more powerful polyp features. A feature supplementation module (FSM) is introduced to exploit different scale-correlated information and address the loss of detailed information. To minimize disease diagnosis errors due to background noise interference and low contrast between polyps and normal tissue, the proposed CIFFormer leverages a cross-scale interactive fusion module (CIFM) to aggregates

the cross-layer features, which can better fuse the contextual information at each level to enhance feature representation. Additionally, a boundary-assisted guidance module (BGM) is introduced to address the polyps with blurred edges and discover boundary clues from high-level low-resolution and low-level high-resolution features. Finally, the performance of CIFFormer is evaluated on five typical polyp datasets, and the experimental results show that the proposed network is superior in segmentation robustness and generalization compared to other competing methods, and can segment various-sized polyps more accurately and completely, which can provide early pre-diagnostic information for patients with colonic polyps. In conclusion, the innovations of this paper consist of four main areas, including:

1. An FSM module feeds the encoder output features into four parallel paths for feature mapping of the dilated convolution. It employs scale-correlated convolutional layers to supplement and exploit low-level features, which can minimize the information loss in the process.
2. A CIFM module based on shuffle attention is proposed to effectively aggregate multi-scale contextual features and capture a wide range of feature information in multi-scale structures. To this end, it effectively suppresses the interference of background noise when extracting feature information.
3. A BGM module is designed to enhance polyp boundary features obtained from the encoding layer in order to pinpoint polyp areas.
4. The loss function adopts a hybrid supervision strategy in which a foreground mask is used to guide the network to correct mispredicted polyp tissue classifications. The boundary mask is used to correct the boundary information for unclear edges.

## 2. Related Work

### 2.1. Traditional Segmentation Methods

Traditional polyp segmentation methods deal with segmentation from an imaging perspective. Segmentation methods often based on shape, texture, color features or a combination of them. Hwang et al. [13] proposed a method that capitalized on the almost oval shape of small bowel rectal polyps for polyp detection. Ameling et al. [14] used contextual features for polyp segmentation by utilizing grayscale covariance matrices. Karkanis et al. [15] introduced new features based on texture measurement covariance to enhance polyp location information. Tajbakhsh et al. [16] utilized both polyp texture and shape features to propose a shape-contextual segmentation method. Traditional segmentation methods are artificially designed to extract features using some fixed information, which is only able to extract typical forms of polyps. However, because polyps vary widely in shape and size, the polyp segmentation methods of traditional methods have low robustness and limited generalization ability.

### 2.2. CNNs for Segmentation Methods

Benefiting from the rapid advancements in convolutional neural networks, the segmentation methods based on CNNs have obvious effects on polyp segmentation, and the segmentation accuracy has taken a great leap forward. After the appearance of symmetry-structured networks based on encoders and decoders, such as U-Net, UNet++, and ResUNet++, their outstanding performance has allowed them to gradually become the predominant force in the realm of medical imaging. Sun et al. [17] improved the encoder by extracting useful features using dilated convolution to enhance the feature reproduction of the network, which can learn advanced semantic information without degrading the resolution. Banik et al. [18] proposed Polyp-Net, which presents a dual-tree pooling with a local gradient-weighted embedding level set. Polyp-Net can effectively avoid the error information in high-signal regions, thus greatly reducing the probability of false positives. Tomar et al. [19] proposed a DDANet for polyp segmentation, which shares the same encoder. There are two parallel decoders, one of which performs an upsampling operation on the input to achieve pixel-level image recovery while the other predicts the segmentation

mask. This bidirectional decoding ensures the correct classification of image pixels and arrives at excellent segmentation results.

### 2.3. Transformer for Polyp Segmentation

TransFuse [20] introduced Transformer to polyp segmentation for the first time, using two parallel encoders, combining convolution and Transformer in a parallel fashion, and using the BiFusion module to fuse features with corresponding locations to achieve outstanding polyp segmentation. Then, Transformer-based segmentation models were demonstrated to have robust feature modeling capabilities and became the focus of research, as they are excellent at capturing long-range dependencies and effectively incorporating contextual information. The latest studies show that SSFormer [21] exploited the low-level features of Transformer to improve the effectiveness of the downstream tasks. Polyp-PVT, HSNNet [22], and PVT-CASCADE [23] solved the problem of the imbalance in the size ratio of foreground targets by augmenting complementary features, fully utilizing the advantages of Transformer-based polyp segmentation.

## 3. Methods

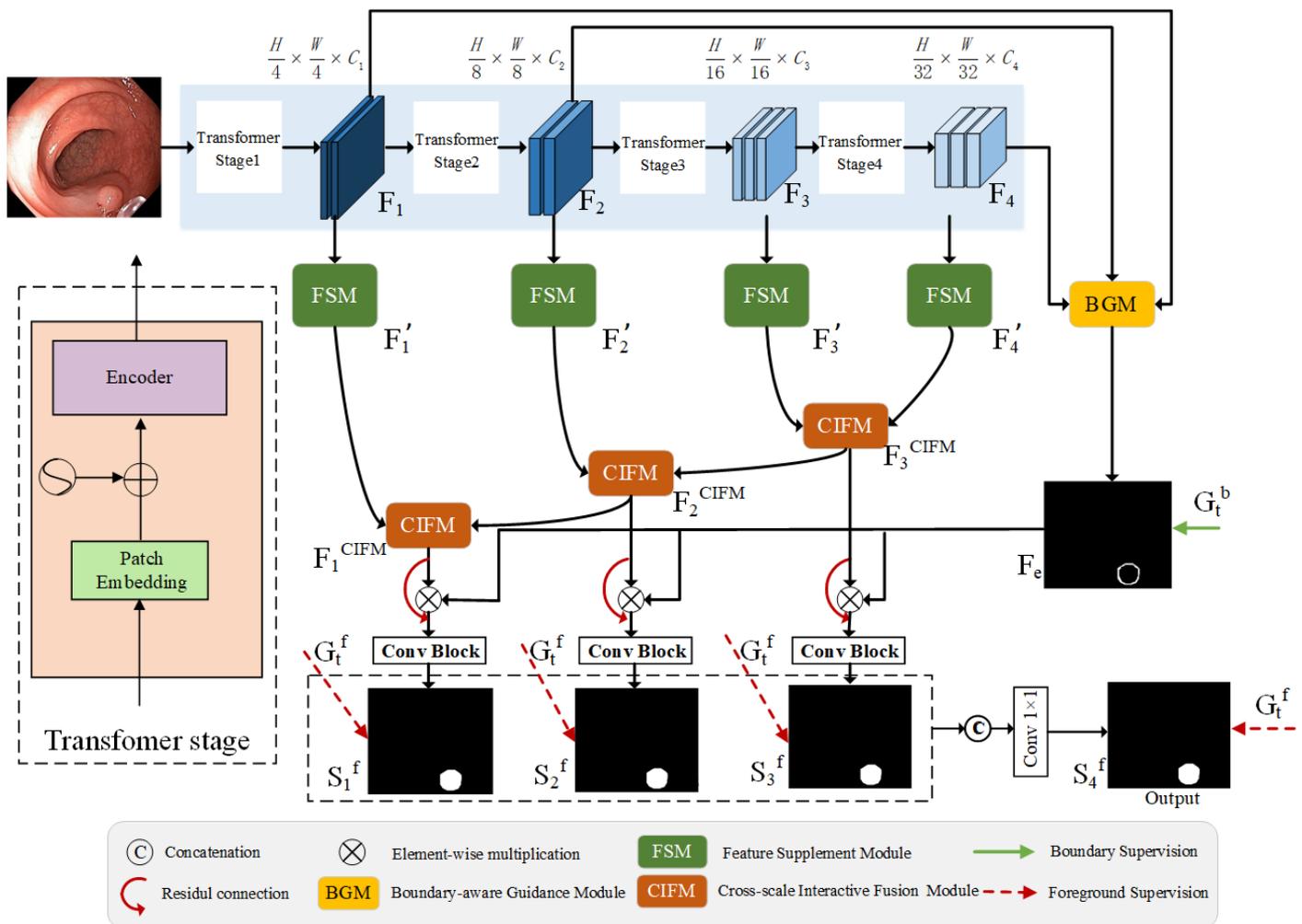
We first present the motivation for proposing the CIFFormer architecture and its components. Then, we describe the feature supplementation module (FSM) in detail, followed by the cross-scale interactive fusion module (CIFM) and boundary-assisted guidance module (BGM). Finally, the loss function used to train the network is elaborated.

### 3.1. The Architecture of CIFFormer

We aim to acquire expertise in an end-to-end segmentation network that can generate polyp masks directly from a given colonoscopy image without going through pre-processing and post-processing techniques. However, there are three difficulties in segmenting polyps. The first is the large-scale variation in lesion regions in colorectal polyp images. The second is the difficulty in distinguishing polyp boundaries from normal tissues in low-contrast images, which leads to the missegmentation of lesion regions and loss of edge detail information. Thirdly, the intestinal environment is complex and variable, and it is easy to introduce a large amount of background noise during polyp segmentation, which affects the segmentation results. To deal with these three challenges, we propose CIFFormer based on two motivations. Specifically, recent studies have shown that visual transformer is more likely to extract global information and show stronger feature representation compared to CNNs [24]. Motivated by this, we employ the widely used pyramidal Visual Transformer PVTv2 to extract multi-scale and robust feature representations for addressing polyps of variable size. Moreover, the model's ability to capture foreground features is compromised by the presence of background noise. To this end, we explore cross-scale interactive fusion to improve the sensitivity of foreground information. Meanwhile, we aim to establish an interaction between high and low semantic information that makes full use of boundary clues, which will improve the model's ability to capture polyp edge features and cope with boundary blurring difficulties.

Figure 2 presents the CIFFormer structure, which employs three novel and effectively validated components. Firstly, the input image  $X \in \mathbb{R}^{H \times W \times 3}$ , uses PVTv2 as the backbone, including four stages to generate feature maps to form pyramid features  $F_i \in \mathbb{R}^{H/2^{i+1} \times W/2^{i+1} \times C_i}$ , respectively, where  $i \in \{1, 2, 3, 4\}$  and  $C_i \in \{C_1, C_2, C_3, C_4\}$ ,  $C_1 = 64$ ,  $C_2 = 128$ ,  $C_3 = 320$ ,  $C_4 = 512$ .  $C_i$  is the channel dimension of the  $i$ -th layer. The stage-specific feature maps are then fed into the FSM to fill in missing details and explore potential feature  $F_i'$ . After that, the ability of the CIFM to aggregate the multi-scale parallel decoding module of the FSM and high-level features of the CIFM output  $F_i^{CIFM}$  is utilized. The CIFM is used to fill the semantic gaps between the cross-scale feature information of each layer. In addition, there is a lack of clear demarcation between polyp lesions and normal tissues. To effectively reduce or mitigate the impact of the background noise and capture the accurate segmentation edges, in the rightmost part of the network, a BGM

module is designed to fuse the low-level semantic features with the high-level semantic convolutions to obtain the contextual semantic information  $F_e \in \mathbb{R}^{H/2^{i+1} \times W/2^{i+1} \times C_i}$  of the edges. Next, the polyp foreground prediction  $S_i^f$  is generated by feeding the boundary clues  $F_e$  and feature map  $F_i^{CIFM}$  into the residual module. Finally, the output prediction map  $S_4^f$  is formed by fusing the three layers of foreground predictions generated with foreground supervision. Throughout the network training process, we employ a hybrid supervision approach that encompasses both foreground supervision and boundary supervision to optimize the network.



**Figure 2.** The architecture of the proposed CIFFormer, which consists of a PVTv2 backbone, the feature supplement module (FSM), cross-scale interactive fusion module (CIFM), and boundary-assisted guidance module (BGM).

### 3.2. Transformer Encoder

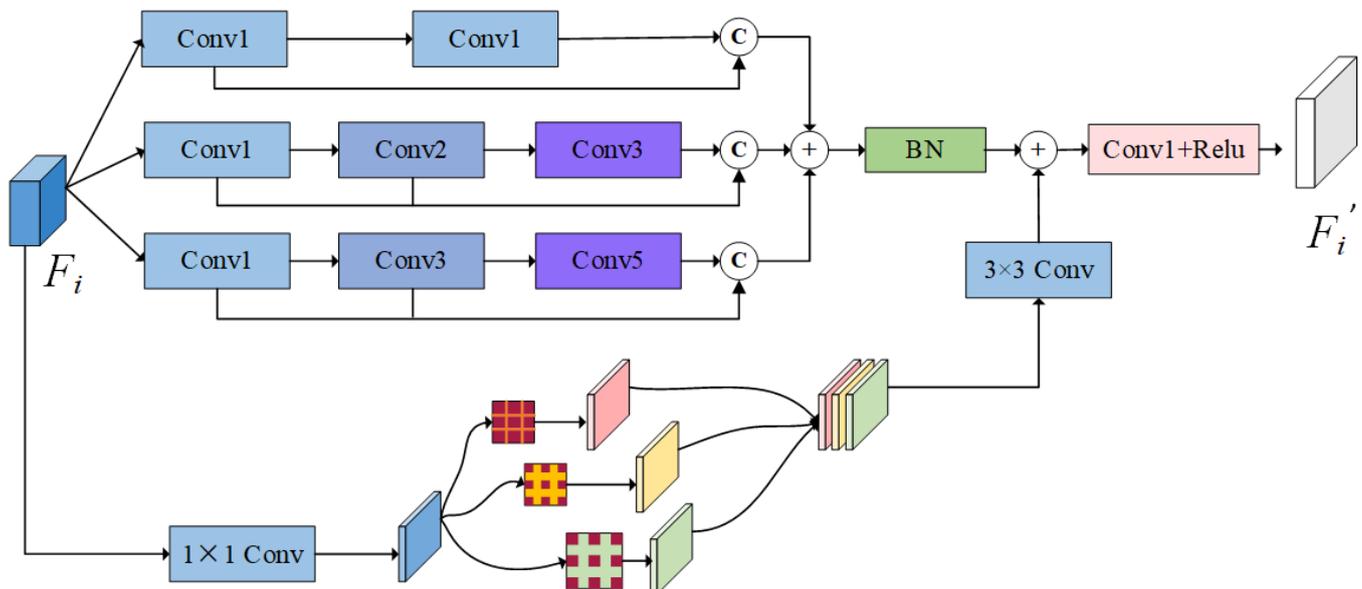
In the proposed CIFFormer, we employ PVTv2 as transformer encoder, which bears more semantic information. Specifically, PVTv2 is utilized as the backbone to extract spatial and channel features through four stages. Each stage follows a similar architecture consisting of a patch embedding layer and a Li-Transformer [12] encoder layer. The feature maps for each stage are denoted as  $\{F_i, i = 1, 2, 3, 4\}$ , respectively.  $F_1, F_2$  are low-level features, which contain more background and noise information. With the FSM module, it is possible to obtain fine-grained details to complement high-level features.  $F_3, F_4$  are high-level features, which have precise pixel information of the polyp boundary area.

### 3.3. Decoder

In order to utilize the boundary information efficiently to make the segmentation boundary more accurate, an edge-enhanced feature is incorporated into the decoder. BGM can learn edge-enhanced features to be incorporated into each decoder unit using a layer-by-layer strategy. Our decoder produces multiple side-out segmentation maps  $S_i^f$ ,  $i \in \{1, 2, 3, 4\}$ . Different from Densely Nested Top-Down Flows (DNTDF) [25], which innovates by enhancing the top-down flow of information to improve the detection of salient objects, or the Deep Unsupervised method, which introduces a Belief Capsule Network (BCNet) [26] to deal with the lack of annotations and to capture part-whole relations in salient object detection, our decoder requires supervision information from the foreground mask and the boundary mask, which is obtained by using the Sobel detection operator. Moreover, Reliable Mutual Distillation [27] is used to address the challenge of segmentation under noisy labels by using a distillation process between two models.

### 3.4. Feature Supplementation Module

In the encoder, as the features undergo successive convolution and downsampling operations, the dimensions of the feature space expand while the scale gradually diminishes. Consequently, there exists a risk of losing detailed characterizations of polyp textures and small lesion areas. To address this challenge, we employ the FSM to supplement features and explore potential features. The feature representation is complemented and enhanced by using dilated convolution to increase the number of receptive fields, which is a common strategy among segmentation tasks. A smaller receptive field can capture fine texture and detail information of the polyps to generate a detail-rich image, which gradually reduces the loss of detailed features and small objects during the downsampling process. The input features are mapped as four parallel perspectives features, and multiple convolutional layers with different dilation rates are utilized to complement and exploit the features, as shown in Figure 3. To extract features, we initially utilize two convolutional layers with a kernel size of  $3 \times 3$  and incorporate batch normalization (BN), as represented by the following equation:



**Figure 3.** Architecture of the proposed feature supplementation module, which consists of four parallel layers.

$$W_1 = \text{Concat}[\text{Conv1}(F_i), \text{Conv1}(\text{Conv1}(F_i))] \quad (1)$$

where  $F_i, i \in \{1, 2, 3, 4\}$  is the output of the four stages of the encoder, respectively.  $Conv_r()$  represents the convolutional layer with a kernel size of  $3 \times 3$  and BN, and  $r \in \{1, 2, 3, 5\}$  is the dilation rate.  $W_i, i \in \{1, 2, 3, 4\}$  is the result of the module's four parallel feature mappings.  $Concat[]$  is a concatenation operation.

$$W_{21} = Conv3(Conv2(Conv1(F_i))) \quad (2)$$

$$W_2 = Concat[Conv1(F_i), Conv2(Conv1(F_i)), W_{21}] \quad (3)$$

Next, the feature map  $F_i$  performs the same operation in two convolutional layers, which consists of three convolution operations with different dilation rates, correspondingly. A concatenation operation follows the third convolution operation, allowing the enhancement of fine-grained features and the exploration of features related to small objects. This process can be represented by:

$$W_{31} = Conv5(Conv3(Conv1(F_i))) \quad (4)$$

$$W_3 = Concat[Conv1(F_i), Conv3(Conv1(F_i)), W_{31}] \quad (5)$$

After performing the addition operation on  $W_1, W_2$ , and  $W_3$  already obtained above, the BN process is then performed. Then, the original feature  $F_i$  is passed through  $1 \times 1$  convolution to obtain  $W'_i$ . Then, to supplement this feature, we feed  $W'_i$  into three  $3 \times 3$  convolutions by using different dilated rates  $r_i \in \{2, 4, 8\}$ , followed by BN with a ReLU activation. After that, the three feature maps are as follows:

$$Y_m = C_{3 \times 3}^{r_m}(W'_i) \quad (6)$$

where  $Y_m, m = (1, 2, 3)$  are three scale features. Then, they are concatenated by passing through a  $3 \times 3$  convolutional layer for adaptive aggregation to obtain the concatenated features, as follows:

$$W_4 = Conv3 \times 3[Concat(Y_1, Y_2, Y_3)] \quad (7)$$

At last, feature maps  $W_1, W_2, W_3$ , and  $W_4$  are combined via an addition operation. We use a convolutional layer with a kernel size of  $3 \times 3$  to extract features, as shown in the following equation:

$$W = BN(W_1 + W_2 + W_3) \quad (8)$$

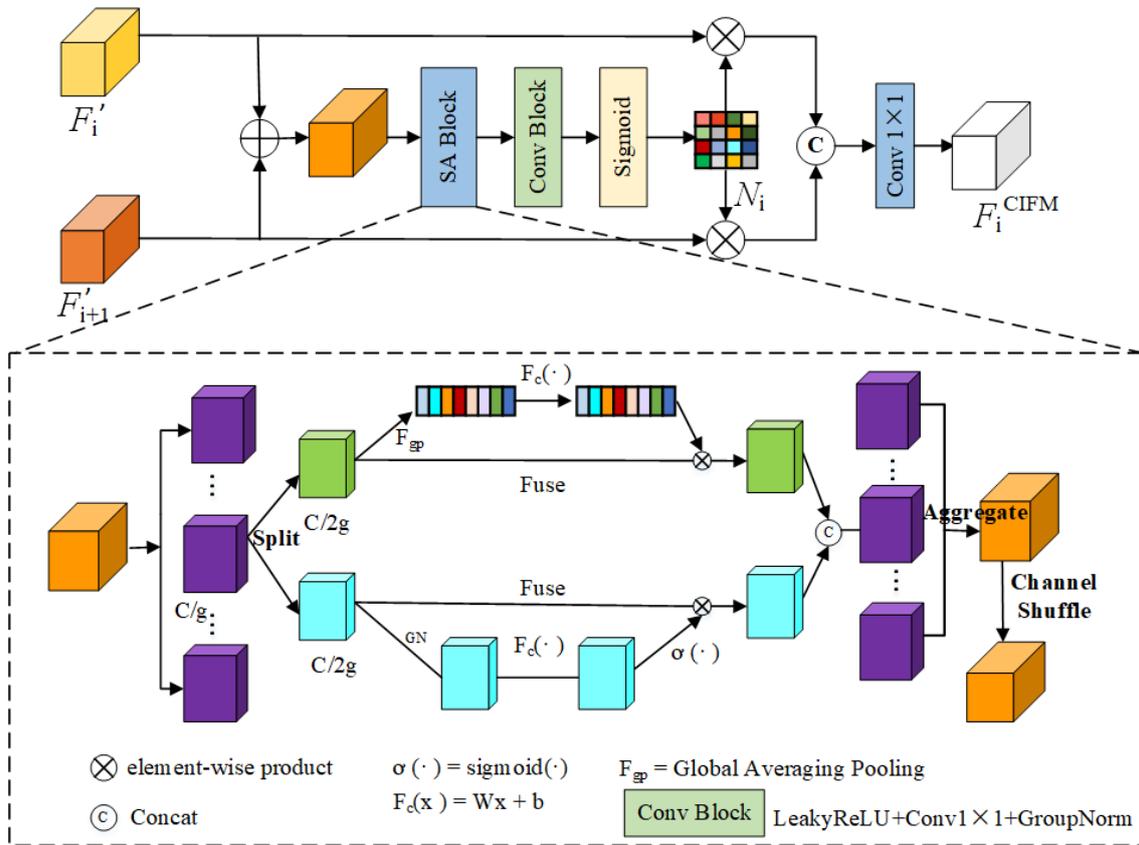
$$F'_i = Relu[Conv1(W + W_4)] \quad (9)$$

It is worth noting that our FSM can enhance multi-scale feature representations by utilizing different dilated convolutions, in which potential features are complemented and explored to reduce the loss of feature information in the training process.

### 3.5. Cross-Scale Interactive Fusion Module

In order to improve effective extraction of different branching features, inspired by ECA-Net [28] and PSPNet [29], we propose a novel CIFM module to fill the semantic gaps between cross-scale feature information at each layer and effectively aggregate multi-scale contextual features. CIFM effectively addresses the challenge of scarce contextual information in the lower layers, suppresses background noise, and enhances segmentation performance in polyp segmentation.

Firstly, the output features of the FSM, such as the  $i$  layer feature  $F'_i \in \mathbb{R}^{H/2^{i+1} \times W/2^{i+1} \times C_i}$  and the  $i+1$  layer feature  $F'_{i+1} \in \mathbb{R}^{H/2^{(i+1)+1} \times W/2^{(i+1)+1} \times C_{i+1}}, i \in \{1, 2, 3, 4\}$ , are used to perform cross-layer feature fusion. In order to fill semantic gaps between layers of different cross-scale feature channels and sizes, two-layer features are fed into the CIFM for further processing, as shown in Figure 4.



**Figure 4.** Architecture of the proposed cross-scale interactive fusion module.

Secondly, the  $F'_i$  and  $F'_{i+1}$  features are fused by elemental addition, and the fusion features are fed into the shuffle attention (SA) module for feature selection. The embedded SA module captures feature dependencies within channels and spaces by grouping the double-layered combined features using feature mapping. Then, it facilitates information interaction between submodules to enhance valuable features and suppress background noise information, ultimately enhancing the overall semantic representation of the feature mapping for subsequent processing. As a result, the SA module is introduced to identify important features for efficient feature learning. After that, the features are operated by LeakyReLU linear units, and then by  $1 \times 1$  convolution. Next, the feature maps are normalized using GroupNorm, which marks the scale difference between each different feature channel. Subsequently, the feature weight map  $N_i$  is generated using the sigmoid function. The processing is shown in the following equation:

$$N_i = \sigma \left[ GN \left( Conv1 \times 1 \left( LeakyReLU \left( SA \left( F'_i + F'_{i+1} \right) \right) \right) \right) \right] \quad (10)$$

where  $N_i$  is the output weight graph,  $SA()$  is the shuffle attention block, and  $GN()$  and  $LeakyReLU()$  are GroupNorm and LeakyReLU activation functions, respectively.

Finally, the feature weight map  $N_i$  obtained previously is utilized to assign weights to the inputs  $F'_i$  and  $F'_{i+1}$ , and the enhanced features are then concatenated for feature fusion, as illustrated in Algorithm 1. Subsequently, the number of channels is adjusted by a  $1 \times 1$  convolution operation in order to generate the output feature  $F_i^{CIFM}$  for cross-scale contextual feature fusion, where  $F_i^{CIFM}$  is the final output feature map of the CIFM module.

$$F_i^{CIFM} = Conv1 \times 1 \left( Concat \left[ \left( N_i \times F'_i \right), \left( N_i \times F'_{i+1} \right) \right] \right) \quad (11)$$

**Algorithm 1** Description of the CIFM

---

**Input:** Feature map  $X$ , Feature map  $X + 1$ , size =  $[B, L, C]$ ,  $B = \text{Batchsize}$ ,  $L = H * W$ ,  $C = \text{Channels}$  //  $X$  is low-level feature,  $X + 1$  is high-level feature

**Output:** Feature map  $Y$ , size =  $[B, L, C]$

- 1:  $input = [B, L, C] \rightarrow input = [B, H, W, C]$  //  $H = \text{height}$ ,  $W = \text{weight}$  as the height and weight of polyp image
- 2:  $x = [B, C, H, W]$ ,  $x + 1 = [B, C, H, W]$
- 3:  $x\_con = x + (x + 1)$
- 4:  $x\_con = x\_con.reshape(B * G, -1, H, W)$  // group into subfeatures,  $G$  is number of groups
- 5:  $x1, x2 = x\_con.chunk(2, dim = 1)$  // channel split
- 6:  $xn = avg\_pool(x1)$
- 7:  $xn = cw * xn + cb$  //  $cw, cb$ : parameters with shape  $[1, C // 2G]$
- 8:  $xn = x1 * \delta(xn)$
- 9:  $xs = GroupNorm(x2)$
- 10:  $xs = sw * xs + sb$  //  $sw, sb$ : parameters with shape  $[1, 1]$
- 11:  $xs = x2 * \delta(xs)$
- 12:  $out = torch.cat([xn, xs], dim = 1)$  // concatenate and aggregate
- 13:  $out = out.reshape(B, -1, H, W)$
- 14:  $out = channel\_shuffle(out, 2)$  // channel shuffle
- 15:  $Y = conv1 * 1[torch.cat(out * x, out * x + 1)]$
- 16: **return**  $Y$
- 17: *In the above formulas,  $\delta$  refers to ReLU function,  $*$  denotes convolution operation, AdaptiveAvgPool denotes AdaptiveAvgPooling.*

---

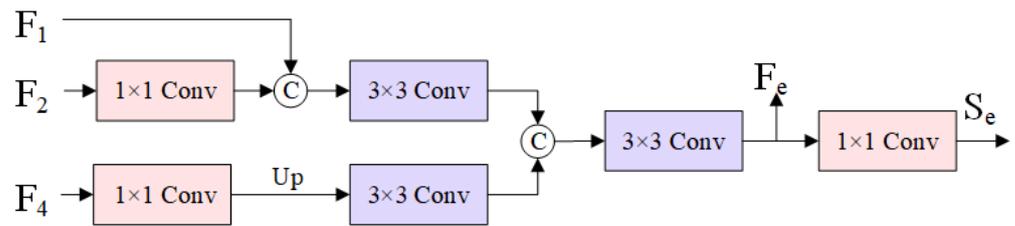
**3.6. Boundary-Assisted Guidance Module**

Boundary information is beneficial to extracting clear boundaries between polyps and normal tissue. Current approaches usually integrate low-level features to learn edge-enhanced representations because low-level features preserve sufficient boundary details. In order to reduce the information difference between low-level and high-level features, non-edge coarse details can also be introduced. We propose a BGM module to enhance the efficiency of obtaining boundary information for polyp lesions, and accurate classification of boundary pixels is achieved.

We combine two low-level features,  $F_1$  and  $F_2$ , with the high-level features  $F_4$  to construct the BGM. Specifically, as presented in Figure 5,  $F_2$  and  $F_4$  are first fed two  $1 \times 1$  convolutional layers, respectively, to obtain  $X_2$  and  $X_4$ . Then,  $F_1$  and  $X_2$  are cascaded and passed through a sequential operation to obtain  $F_{12} = Conv3 \times 3(Cat(F_1, X_2))$ , where  $Conv3 \times 3()$  is a sequential operation that consists of a  $3 \times 3$  convolutional layer, BN, and the ReLU activation function. In addition,  $Concat[]$  is a concatenation operation. Moreover, we conduct a concatenation operation on  $F_{12}$  and  $X_4$  to obtain the boundary-enhanced feature, which can be depicted by:

$$F_e = Conv3 \times 3(Concat[F_{12}, Conv3 \times 3(Up(X_4))]) \quad (12)$$

where  $Up()$  denotes an upsampling operation. Finally,  $F_e$  is passed through a  $1 \times 1$  convolutional layer to generate an edge map  $S_e$ , and the boundary map is upsampled to ensure that it has the same resolution as the original image. It is worth noting that  $F_e$  provides a boundary-assisted feature to weigh the features in the decoder and boost the segmentation performance.



**Figure 5.** Architecture of the proposed boundary-assisted guidance module.

### 3.7. Loss Function

In the polyp segmentation task, the small proportion of polyp lesions leads to an unbalanced pixel distribution of the classification results, and this unbalanced data distribution can cause the model to be heavily biased toward the background during training. To focus on foreground information, as described in Section 3, CIFFormer requires the supervision information from the foreground mask and the boundary mask. In order to optimize the network, we employ a hybrid supervision loss strategy, which is defined as:

$$L_{total} = \underbrace{\sum_{i=1}^4 L_f(G_t^f, S_i^f)}_{\text{foreground}} + \underbrace{\alpha L_b(G_t^b, S_e)}_{\text{boundary}} \quad (13)$$

where the left part denotes the foreground supervision and the right part denotes the boundary supervision, respectively.  $L_f()$  is the loss function for foreground learning.  $G_t^f$  is the ground-truth foreground mask of the polyp. Given the significant disparity between the foreground and background pixel points, it is crucial to select a loss function that enables the network to prioritize hard pixels. To achieve this, we form the loss function  $L_f()$  by linearly combining the two components:

$$L_f = L_{iou}^w + L_{bce}^w \quad (14)$$

where  $L_{iou}^w$  and  $L_{bce}^w$  are the weighted intersection-over-union (IoU) loss and the weighted binary cross-entropy (BCE) loss, respectively. Each pixel in the image will receive a weight reflecting its level of complexity in  $L_{bce}^w$ . This means that more emphasis will be placed on challenging pixels compared to those that are simpler. While the IoU loss excels at optimizing global structure rather than focusing on individual pixels, making it robust against unbalanced sample distributions, it treats all pixels equally regardless of their difficulty, unlike the BCE loss. Similarly to  $L_{bce}^w$ ,  $L_{iou}^w$  assigns greater importance to challenging pixels by using higher weights. Thus, the combination of  $L_{bce}^w$  and  $L_{iou}^w$  can function as a potent loss function that prioritizes difficult pixels while ensuring global structure learning.

$L_b()$  represents the loss function for boundary learning, for which the classical dice loss is utilized in this study, as shown in the following equation:

$$L_b(G_t^b, S_e) = 1 - \frac{2 \sum_i^N p_i y_i + 1}{\sum_i^N p_i^2 + \sum_i^N y_i^2 + 1} \quad (15)$$

where  $G_t^b$  is the ground-truth boundary mask, which is generated from each polyp mask by applying the Sobel detection operator using the default settings in the OpenCV library.  $N$  is the number of pixels of  $G_t^b$ .  $p_i$  and  $y_i$  are the  $i$ th pixel values of  $G_t^b$  and  $S_e$ , respectively. The utilization of weighted boundary loss in training allows for prioritizing boundary pixels by penalizing misclassified positive examples, resulting in precise and well-defined segmentation edges. In this study, we use hyper-parameter  $\alpha = 6$  in the training phase.

## 4. Dataset and Evaluation

### 4.1. Dataset

To compare the effectiveness of the proposed CIFFormer with other competing approaches, for the experiments, we used five typical polyp datasets, as shown in Table 1, including the following:

- CVC-ColonDB: This dataset, the first one used for polyp segmentation, comprises 380 images of colorectal polyps obtained by capturing frames from colonoscopy videos [13].
- ETIS-LaribPolypDB: This is an early collection of colorectal polyp images and was used to perform automated polyp segmentation tasks at the MICCAI 2015 conference [30].
- CVC-ClinicDB: This dataset comprises 612 images of colorectal polyps, which are obtained from a video taken at the time of a colonoscopy [31].
- Kvasir-SEG: The dataset was selected from 1000 images of colorectal polyps acquired from colonoscopy videos [32]. The dataset exhibits significant variability, encompassing images with distinct resolutions and polyp size ratios, and it underwent manual validation by a gastroenterologist with professional expertise.
- EndoTect: This dataset was specifically curated for the 2020 Endoscopy Challenge; it includes 200 polyp images sourced from diverse gastrointestinal tracts using various devices from HyperKvasir. As a result, the dataset presents a broad spectrum of image resolutions, ranging from  $720 \times 576$  to  $1280 \times 1024$ . Additionally, it showcases a diverse array of polyp lesion morphologies [33].

**Table 1.** Presentation of five polyp datasets.

Datasets	Year	Number	Raw Resolution
CVC-ColonDB	2012	380	$574 \times 500$
ETIS-LaribPolypDB	2014	196	$1225 \times 966$
CVC-ClinicDB	2015	612	$384 \times 288$
Kvasir-SEG	2020	1000	Range from $332 \times 487$ to $1920 \times 1072$
EndoTect	2020	60	Range from $720 \times 576$ to $1280 \times 1024$

The five datasets we selected contain typical types of polyps, taking into account the variations and diversity of polyp shapes. They are compared within the same experimental environment to demonstrate the effective segmentation performance of CIFFormer. To mitigate the problem of feature extraction being affected by resolution size, we firstly processed the images on five datasets with a uniform resolution of  $352 \times 352$ . To ensure the fairness of the results of the experiment, the same dataset partitioning setup as PraNet was used for all experiments. Specifically, the training set contains 900 samples from Kvasir-SEG and 550 samples from CVC-ClinicDB, while the test set includes the remaining samples from the above two datasets, as well as all the samples from the other three datasets, so the training and test sets cover most of the differential polyp features. It is worth mentioning that training the model with different feature image datasets not only improves the generalization of the model, but also makes the model more robust. Not only are the images from Kvasir-SEG and CVC-ClinicDB the same as the training set, which allows us to verify the effectiveness of CIFFormer, but also the differentiation of images from three different datasets can be used to verify the generalization ability of CIFFormer.

### 4.2. Evaluation Metrics

To evaluate the comparative performance of various segmentation networks on the same dataset, we have selected indicators for polyp segmentation from the generalized image segmentation evaluation metrics. These include the Mean Dice coefficient (mDice),

mean Intersection over Union (mIoU), Precision, and Recall, enabling us to gauge the semantic accuracy of polyp segmentation in relation to the ground truth, calculated as follows:

$$mDice = \frac{1}{k+1} \sum_{i=0}^k \frac{2TP}{2TP + FP + FN} \quad (16)$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{FN + TP + FP} \quad (17)$$

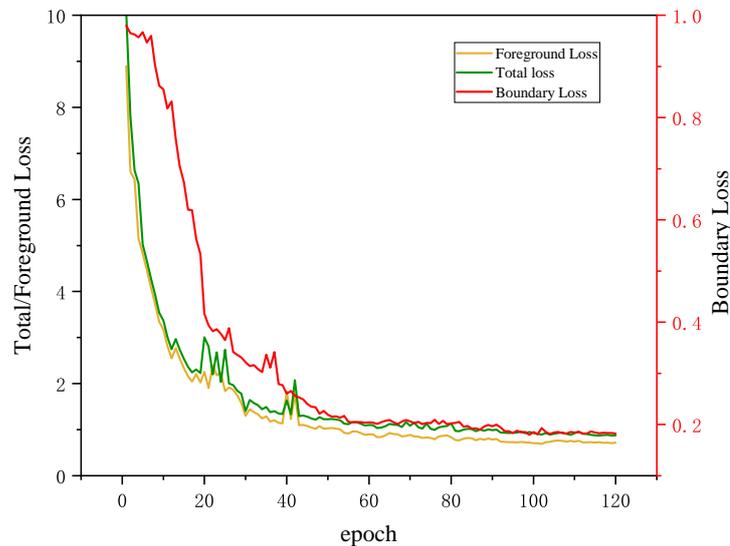
$$Precision = \frac{TP}{TP + FP} \quad (18)$$

$$Recall = \frac{TP}{TP + FN} \quad (19)$$

where  $k$  is the class of the segmentation result.  $TP$ ,  $FP$ ,  $TN$ , and  $FN$  refer to true positive, false positive, true negative, and false negative, respectively. The  $mDice$  and  $mIoU$  are used to measure the similarity between the network segmentation results and the labels.  $Recall$  is concerned with how well the model covers positive examples and  $Precision$  is concerned with how accurately the model predicts as a positive example.

#### 4.3. Implementation Details

In our comparative experiment, all the models were trained and tested on an NVIDIA RTX 3090 GPU with 24 GB of memory. The model was trained for 120 epochs and batch size set to 4. All input images were resized to a uniform  $352 \times 352$  size before model training. During training, we chose the AdamW optimizer with both a learning rate and weight decay of  $1 \times 10^{-4}$ . As the learning rate gradually decreases, our CIFFormer has the capability to be trained to achieve convergence towards the optimal model. Figure 6 shows the training loss curves, including  $L_f$  and  $L_b$ . It can be seen that the loss decreases as the number of training epochs increases and plateaus at epoch up to the 40th epoch for both  $L_f$  and  $L_b$ .



**Figure 6.** Training loss curves of foreground loss  $L_f$ , boundary loss  $L_b$  and total loss  $L_{total}$ .

## 5. Experimental Results

In this section, we perform comprehensive comparative experiments on five datasets using advanced segmentation networks. We validate the superior performance of CIFFormer through quantitative and qualitative comparisons. Additionally, we demonstrate the versatility of the model through cross-validation experiments. Finally, we conduct ablation experiments on the FSM, CIFM, and BGM modules proposed in our study.

### 5.1. Comparison with the Competitive Methods

To further validate the efficacy of the proposed networks in polyp segmentation, we conducted a comparison experiment of CIFFormer with the U-Net, U-Net++, CENet [34], PraNet, Polyp-PVT, TransUNet and SSFormer networks. Experiments were carried out on the same environment and parameter settings for all five datasets, encompassing both quantitative and qualitative comparisons.

**Quantitative Comparison:** We chose *mDice*, *mIoU*, *Recall*, and *Precision* as quantitative metrics to evaluate the segmentation performance of each network. It is clear that our CIFFormer achieves better performance than all competing methods on the two datasets.

The quantitative results of these evaluation metrics for each network on Kvasir-SEG are presented in Table 2. Our proposed CIFFormer has better performance than the other networks. Both *mDice* and *mIoU* have reached the optimal values of 0.925 and 0.875, respectively, which are 0.4% and 0.2% ahead of SSFormer. Recall and Precision are 0.942 and 0.921, respectively. Recall is 0.7% lower than CENet, and Precision is 0.4% lower than SSFormer. Although Recall and Precision do not reach the optimal metrics, they also obtain sub-optimal indicators. Comprehensively analyzing the four indices, CIFFormer can segment polyps more efficiently compared to Transformer-based polyp segmentation models, such as Polyp-PVT, TransUNet, and SSFormer.

**Table 2.** The experimental comparison of different network models on Kvasir-SEG dataset. (Bolded numbers in the table represent maximum values).

Method	Year	mDice	mIOU	Recall	Precision
U-Net	2015	0.818	0.746	0.948	0.862
U-Net++	2018	0.821	0.743	0.900	0.871
CENet	2019	0.873	0.866	<b>0.949</b>	0.906
PraNet	2020	0.898	0.840	0.914	0.922
Polyp-PVT	2021	0.917	0.864	0.925	0.911
TransUNet	2021	0.903	0.886	0.944	0.865
SSFormer	2022	0.921	0.873	0.938	<b>0.925</b>
CIFFormer (ours)	2024	<b>0.925</b>	<b>0.875</b>	0.942	0.921

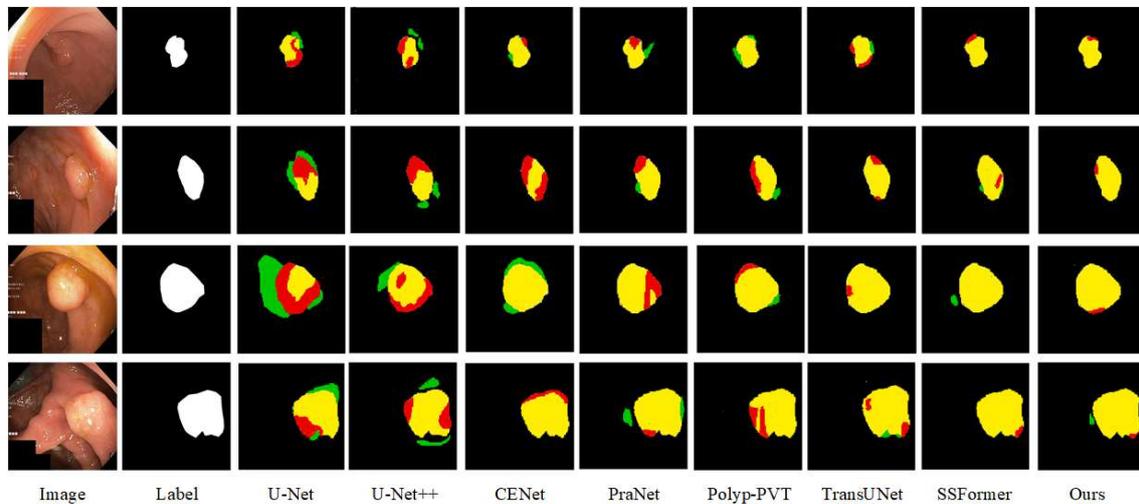
From Table 3, it is evident that CIFFormer demonstrates superior performance on the CVC-ClinicDB dataset. Specifically, the mIOU of our model achieves 0.2% and 2% improvement over TransUNet and Polyp-PVT, remaining relatively stable. Additionally, the mDice of our model is 0.934, which is not the highest value, but the second highest, and Polyp-PVT is only 0.3% higher than our model. The Recall and Precision are also as high as 0.946 and 0.945, compared with SSFormer, with 2% and 0.4% improvements, respectively. These results demonstrate that CIFFormer has an obvious segmentation advantage on the two datasets.

**Table 3.** The experimental comparison of different network models on CVC-ClinicDB dataset. (Bolded numbers in the table represent maximum values).

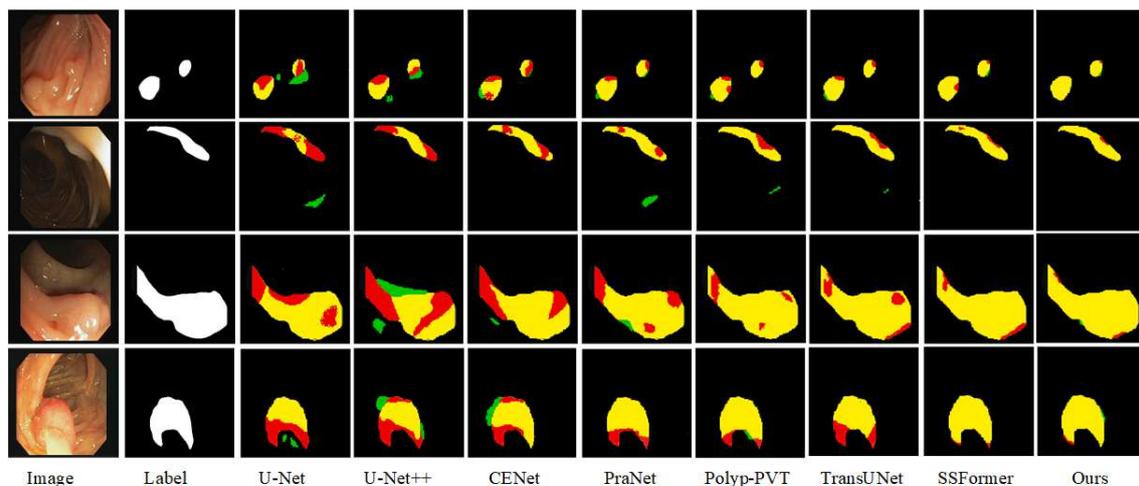
Method	Year	mDice	mIOU	Recall	Precision
U-Net	2015	0.823	0.755	0.943	0.881
U-Net++	2018	0.794	0.729	<b>0.954</b>	0.832
CENet	2019	0.901	0.845	0.915	<b>0.970</b>
PraNet	2020	0.899	0.849	0.935	0.948
Polyp-PVT	2021	<b>0.937</b>	0.889	0.901	0.935
TransUNet	2021	0.902	0.907	0.844	0.939
SSFormer	2022	0.913	0.868	0.926	0.941
CIFFormer (ours)	2024	0.934	<b>0.909</b>	0.946	0.945

**Qualitative Comparison:** In order to more intuitively assess the segmentation ability of the various advanced networks on the Kvasir-SEG and CVC-ClinicDB datasets, we

selected four samples with large differences in polyp appearance from each dataset, and analyzed the segmentation ability of the networks based on the visualized segmentation results, as shown in Figures 7 and 8.



**Figure 7.** Results of the polyp segmentation images visualized on the dataset Kvasir-SEG. The yellow, green, and red colors in each prediction map represent the true positive, false positive, and false negative regions, respectively.



**Figure 8.** Results of the polyp segmentation images visualized on the CVC-ClinicDB datasets.

Four images from the Kvasir-SEG dataset were selected with different polyp sizes; due to the low light of the images, the segmented polyp target areas have poor contrast with the background areas. The polyps in the first and second groups are similar to the surrounding tissues, and we can see that the polyp contours could not be prepared for localization in U-Net and U-Net++. Furthermore, some of the polyps were incorrectly predicted as part of the background. The segmentation effect was improved in Polyp-PVT, TransUNet, and SSFormer. The third and fourth groups of polyp organization are more obvious, due to increased background noise interference; the CENet and PraNet network segmentation results show that background noise interferes with the models' sensitivity to foreground information, and they misclassify the background noise as the target. In contrast, while the segmentation results of CIFFormer are closest to the labeled images, there remains some ambiguity in segmenting the edge details.

Sample images of polyp tissues that are hidden and difficult to be seen in the CVC-ClinicDB dataset are selected, and from the qualitative analysis, we can see that the four

irregular polyp shapes are more difficult to segment. The second group of polyps has a large chromatic aberration with the background, thus affecting the segmentation boundaries, and the U-Net and PraNet networks are unable to correctly segment the polyps and normal tissues. The third and fourth original images are irregular polyp shapes, and the segmentation results in CE-Net, PraNet, Polyp-PVT, and TransUNet have the problem of misclassifying the lesion area as the background. Precise segmentation accuracy has been achieved by SSFormer, but it cannot perform segmentation well when dealing with convex sharp corners and critical regions and does not reach the segmentation advantage of the CIFFormer network proposed in this paper. Compared with competing networks, CIFFormer can suppress complex backgrounds, has coherent boundaries, and is more suitable for polyp segmentation.

### 5.2. Cross-Validation Experiment

**Model generalization ability:** Each dataset has its own proprietary features used to achieve good results when performing the polyp segmentation task. In addition to performing well on a specific dataset, an excellent segmentation method needs to maintain a stable performance on unknown data. Therefore, the model needs to have excellent generalization capabilities to obtain the best segmentation results. We performed a series of cross-validation experiments to test the generalization ability of CIFFormer on three unseen datasets, namely the CVC-ColonDB, ETIS-LaribPolypDB, and EndoTect datasets.

Table 4 presents the comparison results of competitive segmentation methods. Our proposed CIFFormer network has clear advantages on all unseen datasets, and has obtained the highest mean Dice scores of 0.824 on CVC-ColonDB and 0.793 on ETIS-LaribPolypDB. Compared to the second-ranked SSFormer, the mDice scores of CIFFormer are 0.9% and 0.7% higher in the CVC-ColonDB and ETIS-LaribPolypDB datasets, respectively. On the EndoTect dataset, mDice, mIOU, Recall, and Precision reached 0.729, 0.633, 0.756, and 0.759, respectively. mDice and Recall of our proposed method did not have the highest values, but mDice was 7% higher than U-Net, Recall was 3.2% higher than CENet, and higher than the highest value, obtained by PraNet, by only a small difference of 0.2%. In conclusion, from the results of the cross-validation experiments on the three datasets, CIFFormer demonstrates robust performance across all three datasets, proving its effectiveness in polyp segmentation and demonstrating strong generalization ability.

**Table 4.** The results of cross-validation experiments with competitive methods (Bolded numbers in the table represent maximum values).

Method	CVC-ColonDB				ETIS-LaribPolypDB				EndoTect			
	mDice	mIOU	Recall	Precision	mDice	mIOU	Recall	Precision	mDice	mIOU	Recall	Precision
U-Net	0.584	0.625	0.781	0.854	0.714	0.622	0.802	0.785	0.659	0.565	0.733	0.715
U-Net++	0.726	0.670	0.766	0.825	0.756	0.631	0.756	0.791	0.662	0.571	0.741	0.718
CENet	0.753	0.733	0.810	0.833	0.732	0.658	0.783	0.795	0.674	0.569	0.726	0.735
PraNet	0.784	0.716	0.832	0.848	0.744	0.647	0.791	0.806	0.651	0.584	<b>0.758</b>	0.746
Polyp-PVT	0.774	0.685	0.825	0.857	0.751	0.695	0.814	<b>0.869</b>	0.714	0.595	0.737	0.738
TransUNet	0.806	0.704	0.834	<b>0.861</b>	0.777	0.682	0.824	0.827	0.706	0.598	0.741	0.742
SSFormer	0.815	<b>0.721</b>	0.841	0.860	0.786	0.701	<b>0.837</b>	0.835	<b>0.733</b>	0.601	0.748	0.755
CIFFormer (ours)	<b>0.824</b>	0.719	<b>0.857</b>	0.858	<b>0.793</b>	<b>0.716</b>	0.826	0.844	0.729	<b>0.633</b>	0.756	<b>0.759</b>

The visualization results on three datasets are shown in Figures 9–11. The ETIS dataset was selected for its images with high segmentation difficulty and low polyp and background discrimination. The segmentation results are rough around the edged of the polyps under the influence of light, and there are large areas misclassified as lesions by the U-Net and U-Net++ networks. Especially for irregular polyps, the background noise is misclassified as polyp tissue. The SSFormer network is more sensitive to the polyp region, and enhances the effectiveness of segmentation. Compared with the real label map, CIFFormer is closest to the label, although there is an unclear problem at the

boundary. EndoTect selects polyp images with relatively regular shapes and different sizes, and most of the networks can segment the contours correctly. In the third row of images, the background noise is too large and interferes with the foreground information feature extraction, misclassifying the noise as a polyp region. This problem is improved in the TransUNet, SSFormer, and CIFFormer models, which are better able to prepare the polyp region for imaging. The problem of polyp organization and background blurring exists in CVC-colonDB for small target segmentation, which is solved in PraNet. In the fifth row, the polyp lesion has a piece of protruding tissue, and it can be seen that CIFFormer can enhance the information interaction between the polyp foreground and the background through the CIFM module, capturing the protruding features that can be easily overlooked and segmenting them effectively at the edges. The cross-validation results of competing methods on CVC-ColonDB, ETIS-LaribPolypDB, and EndoTect datasets above prove that our model demonstrates superior generalization ability.

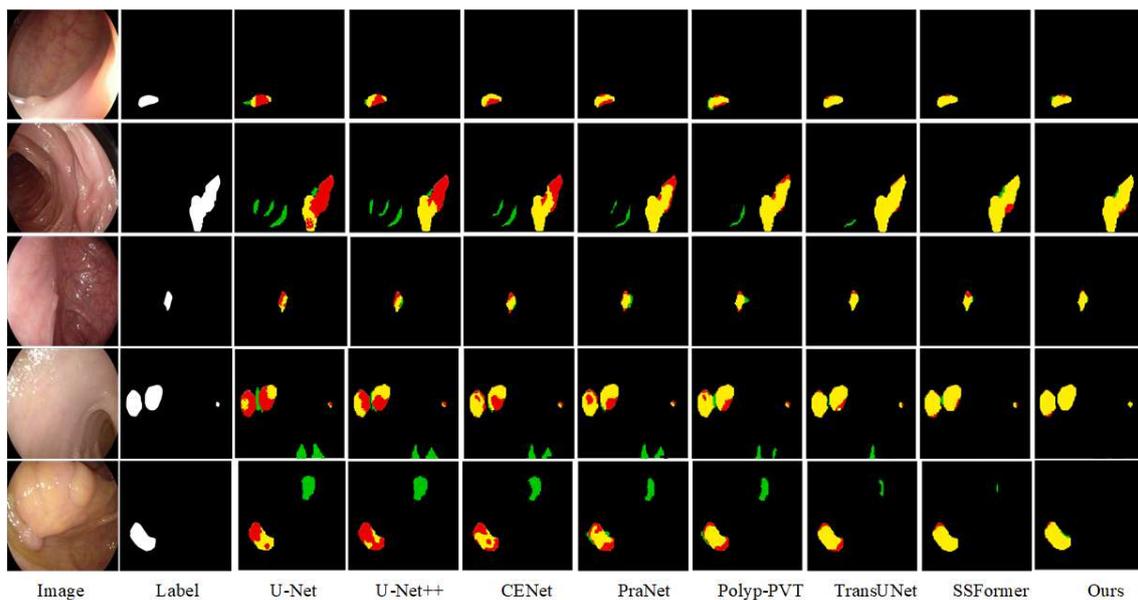


Figure 9. Results of the polyp segmentation images visualized on the dataset ETIS-LaribPolypDB.

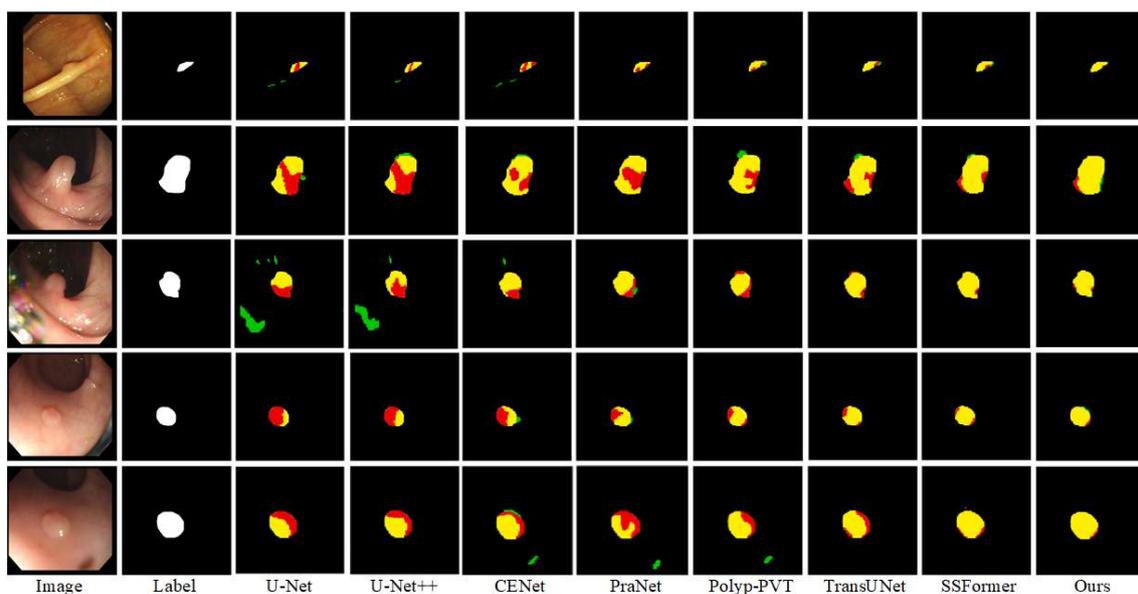
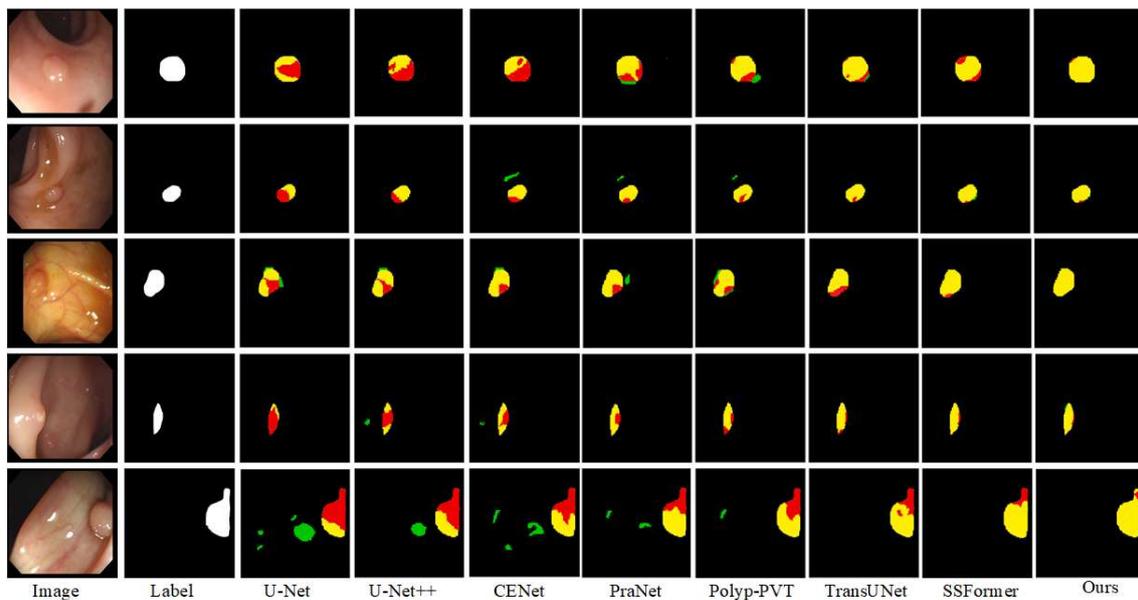


Figure 10. Results of the polyp segmentation images visualized on the dataset EndoTect.



**Figure 11.** Results of the polyp segmentation images visualized on the dataset CVC-colonDB.

### 5.3. Ablation Study

To verify the effect of each module in CIFFormer on the segmentation results, quantitative analyses of ablation experiments are performed in this section, as shown in Table 5. CIFFormer first selects the baseline model and adds FSM, CIFM, and BGM as well as a combination of modules to the baseline model individually. Then, the ablation study proves the effectiveness of each module on CVC-ColonDB, ETIS-LaribPolypDB, CVC-ClinicDB, Kvasir-SEG, and EndoTect datasets. It can be seen that CIFFormer, with the addition of three modules, achieves the best performance for polyp segmentation.

**Table 5.** Comparison table of segmentation results for ablation experiments of our model.

Method				CVC-ColonDB		ETIS-Larib		CVC-ClinicDB		Kvasir-SEG		EndoTect	
Baseline	FSM	CIFM	BGM	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
✓	-	-	-	0.786	0.702	0.749	0.662	0.767	0.704	0.865	0.785	0.836	0.783
✓	✓	-	-	0.793	0.716	0.751	0.678	0.781	0.715	0.873	0.791	0.845	0.795
✓	✓	✓	-	0.817	0.722	0.754	0.683	0.796	0.736	0.882	0.817	0.857	0.801
✓	-	✓	✓	0.806	0.726	0.752	0.679	0.787	0.744	0.879	0.824	0.861	0.814
✓	✓	-	✓	0.815	0.721	0.761	0.681	0.795	0.741	0.877	0.819	0.854	0.806
✓	✓	✓	✓	0.826	0.734	0.773	0.702	0.804	0.758	0.889	0.833	0.865	0.828

#### (1) Effectiveness of FSM

By incorporating the FSM module into the baseline, we aimed to enhance potential detailed features while minimizing the loss of valid information. The results, as depicted in row 4 of Table 5, demonstrate notable improvements in metrics across all five datasets. Specifically, the mDice improves by 0.7%, 0.2%, and 1.4% in CVC-ColonDB, ETIS-Larib, and CVC-ClinicDB, respectively. Thus, the module significantly enhanced the model's capture of polyp features, suppressed the interference of noise, and improved the segmentation ability.

#### (2) Effectiveness of CIFM

The CIFM module is used for cross-scale interaction feature fusion. From Table 5, it can be seen that overlaying CIFM on top of the baseline model and the FSM increases mDice and mIoU by a certain percentage. On the Kvasir-SEG and EndoTect datasets, mDice improves by 0.9% and 1.2%, respectively, and the mIoU improves from 0.791 to 0.817. The results indicate that incorporating the CIFM module into the baseline network can be a beneficial

strategy for obtaining more refined foreground and edge features, suppressing irrelevant background information, and ultimately enhancing polyp segmentation performance.

### (3) Effectiveness of BGM

As shown in Table 5, there is a small decrease in mDice and mIoU compared to the third and sixth rows when BGM is missing across the datasets. The mDice decreases from 0.773 to 0.754 on the ETIS-Larib dataset and from 0.889 to 0.882 on the Kvasir-SEG dataset, so the BGM has a contributing role in the segmentation effect of the polyps and helps to achieve good polyp margins in terms of the segmentation performance.

### (4) Loss Function

In order to enhance the mDice and mIoU performance, as well as facilitate faster convergence during the training phase, we employ a loss function that combines the foreground loss  $L_f$  and the boundary loss  $L_b$  using a linear combination approach. We train two CIFFormer on CVC-ColonDB dataset, one with only foreground loss (CIFFormer +  $L_f$ ) and the other with both foreground loss and weighted boundary loss (CIFFormer +  $L_f + L_b$ ). According to the results presented in Table 6, the introduction of the boundary loss leads to a significant improvement in both mDice and mIoU metrics, with increases of 1.7 % and 3.2%, respectively, from 0.735 to 0.752 and 0.662 to 0.694.

**Table 6.** Ablation study on the impact of the loss function on CVC-ColonDB dataset.

Method	mDice	mIoU
CIFFormer + $L_f$	0.735	0.662
CIFFormer + $L_f + L_b$	0.752	0.694

## 6. Conclusions

In this paper, we propose a novel cross-scale interaction fusion transformer for polyp segmentation, termed CIFFormer, for accurate and robust segmentation of polyps in colonoscopy images. Rather than continuing to explore different variants of the Transformer model like in previous studies, we focus on preserving detailed features and reducing background noise from interfering with foreground features. Specifically, considering that the Transformer encoder has a large divide at all levels, we first propose an FSM to minimize the information loss in the encoder process. Then, we introduce a CIFM to effectively suppress the interference of background noise in extracting foreground information. Moreover, a BGM module is designed to enhance boundary features. Extensive experiments are conducted on five challenging polyp segmentation datasets. In order to evaluate the generalizability of CIFFormer, we performed cross-validation experiments. The results support the notion that CIFFormer exhibits promising potential for further advancement and utilization within the medical image segmentation domain.

**Author Contributions:** Conceptualization, L.J., Y.H. and T.Z.; methodology, L.J., Y.F. and Y.J.; software, Y.F.; validation, Y.H.; writing—original draft preparation, L.J.; writing—review and editing, T.Z., Y.H. and Y.F.; supervision, Y.J. and T.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Key R&D Program of China (2023YFB2904000, 2023YFB2904004), Jiangsu Key Development Planning Project (BE2023004-2), Natural Science Foundation of Jiangsu Province (Higher Education Institutions) (20KJA520001), The 14th Five-Year Plan project of Equipment Development Department (315107402), Jiangsu Hongxin Information Technology Co., Ltd. Project (JSSGS2301022EGN00), Future Network Scientific Research Fund Project (No. FNSRFP-2021-YB-15), 2021 Jiangsu Higher Education Teaching Reform Research General Project (No. 2021JSJG519).

**Data Availability Statement:** The data and code used to support the findings of this study are available from the author upon request (2022010301@njupt.edu.cn).

**Conflicts of Interest:** The authors declare that this study received funding from Jiangsu Hongxin Information Technology Co., Ltd. The funder had the following involvement with the study: Provide confirmation of the research significance of the topic and the value of directing future in-depth research on the topic for application to industry.

## References

- Kim, T.; Lee, H.; Kim, D. UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021.
- Gross, S.; Kennel, M.; Stehle, T.; Wulff, J.; Aach, T. Polyp Segmentation in NBI Colonoscopy. *DBLP* **2009**, *22*, 252–256.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Houshy, N. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
- Naseer, M.; Ranasinghe, K.; Khan, S.; Hayat, M.; Yang, M.H. Intriguing Properties of Vision Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 23296–23308.
- Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
- Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021.
- Dong, B.; Wang, W.; Fan, D.P.; Li, J.; Fu, H.; Shao, L. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. *arXiv* **2021**, arXiv:2108.06932.
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
- Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [[CrossRef](#)]
- Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020.
- Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Shao, L. PVTv2: Improved Baselines with Pyramid Vision Transformer. *Comput. Vis. Media* **2021**, *8*, 415–424. [[CrossRef](#)]
- Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *IEEE Trans. Med. Imaging* **2016**, *35*, 630–644. [[CrossRef](#)]
- Ameling, S.; Wirth, S.; Paulus, D.; Lacey, G.; Vilaro, F. Texture-Based Polyp Detection in Colonoscopy. *DBLP* **2009**, *22*, 346–350.
- Jensen, T.R.; Schmainda, K.M. Computer-aided detection of brain tumor invasion using multiparametric MRI. *J. Magn. Reson. Imaging* **2010**, *30*, 481–489. [[CrossRef](#)] [[PubMed](#)]
- Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; De Lange, T.; Halvorsen, P.; Johansen, H.D. Resunet++: An advanced architecture for medical image segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; pp. 225–2255.
- Sun, X.; Zhang, P.; Wang, D.; Cao, Y.; Liu, B. Colorectal Polyp Segmentation by U-Net with Dilation Convolution. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019.
- Banik, D.; Roy, K.; Bhattacharjee, D.; Nasipuri, M.; Krejcar, O. Polyp-Net: A Multi-model Fusion Network for Polyp Segmentation. *IEEE Trans. Instrum. Meas.* **2020**, *70*, 4000512.
- Tomar, N.K.; Jha, D.; Ali, S.; Johansen, H.D.; Halvorsen, P. DDANet: Dual Decoder Attention Network for Automatic Polyp Segmentation. In Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges, Virtual, 10–15 January 2021.
- Qu, L.; Liu, S.; Wang, M.; Li, S.; Yin, S.; Qiao, Q.; Song, Z. TransFuse: A Unified Transformer-based Image Fusion Framework using Self-supervised Learning. *arXiv* **2022**, arXiv:2201.07451.
- Wang, J.; Huang, Q.; Tang, F.; Meng, J.; Su, J.; Song, S. Stepwise Feature Fusion: Local Guides Global. *arXiv* **2022**, arXiv:2203.03635.
- HSNet: A hybrid semantic network for polyp segmentation. *Comput. Biol. Med.* **2022**, *150*, 106173. [[CrossRef](#)]
- Monaco, J.; Raess, P.; Chawla, R.; Bagg, A.; Madabhushi, A. Image segmentation with implicit color standardization using cascaded EM: Detection of myelodysplastic syndromes. In Proceedings of the IEEE International Symposium on Biomedical Imaging, Barcelona, Spain, 2–5 May 2012; pp. 740–743.
- Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Veit, A. Understanding Robustness of Transformers for Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
- Fang, C.; Tian, H.; Zhang, D.; Zhang, Q.; Han, J. Densely Nested Top-Down Flows for Salient Object Detection. *Sci. China Inf. Sci.* **2021**, *65*, 182103. [[CrossRef](#)]

26. Liu, Y.; Dong, X.; Xu, S.; Zhang, D. Deep unsupervised part-whole relational visual saliency. *Neurocomputing* **2024**, *563*, 126916. [[CrossRef](#)]
27. Fang, C.; Wang, Q.; Cheng, L.; Gao, Z.; Pan, C.; Cao, Z.; Zheng, Z.; Zhang, D. Reliable mutual distillation for medical image segmentation under imperfect annotations. *IEEE Trans. Med. Imaging* **2023**, *42*, 1720–1734. [[CrossRef](#)] [[PubMed](#)]
28. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
29. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
30. Silva, J.; Histace, A.; Romain, O.; Dray, X.; Granado, B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surgery* **2013**, *9*, 283–293. [[CrossRef](#)] [[PubMed](#)]
31. Bernal.; Jorge.; Vilarino.; Fernando.; Fernandez-Esparrach.; Gloria.; Gil.; Debora.; Sanchez, J.; F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [[CrossRef](#)]
32. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-seg: A segmented polyp dataset. In Proceedings of the MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, Republic of Korea, 5–8 January 2020; Proceedings, Part II 26; Springer: Berlin/Heidelberg, Germany, 2020; pp. 451–462.
33. Hicks, S.A.; Jha, D.; Thambawita, V.; Halvorsen, P.; Hammer, H.L.; Riegler, M.A. The EndoTect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges, Virtual Event, 10–15 January 2021; Proceedings, Part VIII; Springer: Cham, Switzerland, 2021; pp. 263–274.
34. Hao, H.; Fu, H.; Liu, J.; Cheng, J.; Zhou, K.; Gao, S.; Zhang, T.; Zhao, Y.; Gu, Z. CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2019**, *38*, 2281–2292.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.