Pattern Recognition 140 (2023) 109555

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Cross-level Feature Aggregation Network for Polyp Segmentation

Tao Zhou[a,b], Yi Zhou[c], Kelei He[d], Chen Gong[a,*], Jian Yang[a], Huazhu Fu[e], Dinggang Shen[f,g,*]

[a] PCA Lab, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200240
[c] School of Computer Science and Engineering, Southeast University, Nanjing 211189, China
[d] Medical School, National Institute of Healthcare Data Science, Nanjing University, Nanjing 210023, China
[e] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore
[f] School of Biomedical Engineering, ShanghaiTech University, and Shanghai Clinical Research and Trial Center, Shanghai 201210, China
[g] Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200232, China

ARTICLE INFO

Article history:
Received 13 April 2022
Revised 12 November 2022
Accepted 24 March 2023
Available online 26 March 2023

Keywords:
Polyp segmentation
boundary-aware features
cross-level feature fusion
boundary aggregated module

ABSTRACT

Accurate segmentation of polyps from colonoscopy images plays a critical role in the diagnosis and cure of colorectal cancer. Although effectiveness has been achieved in the field of polyp segmentation, there are still several challenges. Polyps often have a diversity of size and shape and have no sharp boundary between polyps and their surrounding. To address these challenges, we propose a novel Cross-level Feature Aggregation Network (CFA-Net) for polyp segmentation. Specifically, we first propose a boundary prediction network to generate boundary-aware features, which are incorporated into the segmentation network using a layer-wise strategy. In particular, we design a two-stream structure based segmentation network, to exploit hierarchical semantic information from cross-level features. Furthermore, a Cross-level Feature Fusion (CFF) module is proposed to integrate the adjacent features from different levels, which can characterize the cross-level and multi-scale information to handle scale variations of polyps. Further, a Boundary Aggregated Module (BAM) is proposed to incorporate boundary information into the segmentation network, which enhances these hierarchical features to generate finer segmentation maps. Quantitative and qualitative experiments on five public datasets demonstrate the effectiveness of our CFA-Net against other state-of-the-art polyp segmentation methods. The source code and segmentation maps will be released at https://github.com/taozh2017/CFANet.

1. Introduction

Colorectal cancer (CRC) is the third most frequently diagnosed cancer around the world [1,2]. CRC usually arises from adenomatous polyps, and if left untreated, a polyp usually takes 10 – 15 years to develop into cancer. Therefore, effective detection and removal of polyps before they become malignant can prevent the occurrence of CRC and significantly reduce mortality rates. To decrease mortality, early detection and assessment of polyps are highly critical. For an initial evaluation, a popular procedure for clinicians is to identify the adenomatous polyps, and then polyps are delineated in colonoscopy images manually by highly trained clinicians. However, manual detection and segmentation of polyps are time-consuming and subjective. Thus, an effective solution is to develop automatic polyp segmentation algorithms to help clinicians accurately locate and segment polyp regions for further diagnosis [3,4].

Polyps vary over time at different development stages with a diversity of sizes and shapes, making their accurate segmentation challenging (see Fig. 1). Moreover, it is difficult to segment polyps due to the high intrinsic similarities between a polyp and its surrounding mucosa. To handle these challenges, various deep learning models have developed and demonstrated promising performance for polyp segmentation. For example, Akbari et al. [6] adopted a fully convolutional network (FCN) and Otsu thresholding to extract the largest connected regions for polyp segmentation. Sun et al. [7] proposed an FCN-based polyp segmentation framework, in which a dilated convolution is introduced to learn high-level semantic features without resolution reduction. Moreover, UNet-based methods with an encoder-decoder structure have shown promising performance. Among these methods, high-level features in the decoder are gradually up-sampled

* Corresponding authors.
E-mail addresses: taozhou.ai@gmail.com (T. Zhou), yizhou.szcn@gmail.com (Y. Zhou), hkl@nju.edu.cn (K. He), chen.gong@njust.edu.cn (C. Gong), csjyang@njust.edu.cn (J. Yang), hzfu@ieee.org (H. Fu), Dinggang.Shen@gmail.com (D. Shen).

https://doi.org/10.1016/j.patcog.2023.109555
0031-3203/

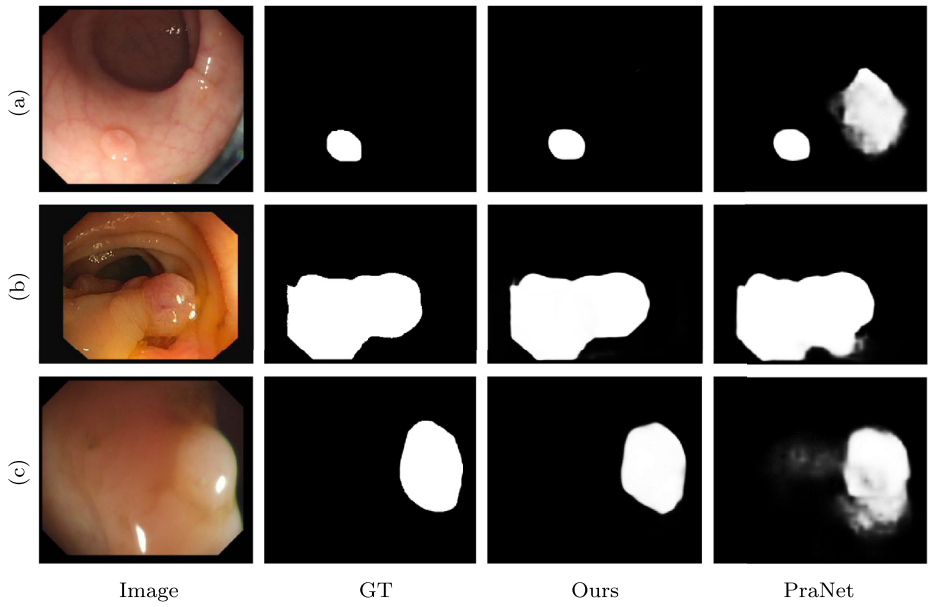

Fig. 1. Different challenging scenarios, including (a) small polyp, (b) big polyp, and (c) non-sharp boundary. Our model outperforms a cutting-edge model (*i.e.*, PraNet [5]).

and fused with corresponding low-level ones in the encoder via skip connections, in which high-level semantic and low-level context information can be integrated effectively [8]. Two variants of the UNet architecture, ResUNet++ [9] and UNet++ [10] have been developed for polyp segmentation and achieved satisfactory performance. However, the above-mentioned methods often focus on segmenting the whole region of the polyp, while neglecting some valuable boundary information. To overcome this issue, some works introduced area-boundary constraints [11] or constructed a multi-task framework to extract contour information [12] for improving the segmentation performance. Additionally, Fan et al. [5] utilized a reverse attention (RA) module to exploit boundary cues, which is helpful for accurately segmenting polyps.

Although effectiveness has been achieved in the field of automatic polyp segmentation, there are several challenges remaining for existing methods. First, in the case of flat lesions or unclean bowel preparation, the boundaries between polyps and their background are not sharp, leading to inaccurate segmentation results. Therefore, it is critical to exploit boundary information that provides boundary-aware guidance to establish the correlation between polyp regions and boundary cues. Second, scale variation is one of the major challenges in polyp segmentation, how to effectively characterize the multi-scale information from a convolutional layer deserves further exploration. Third, the convolutional neural network (CNN) consists of a series of multi-scale convolutional layers. The shallower layers retain the structure details (*e.g.*, boundaries), while deeper layers encode high-level semantic information to locate the polyp regions. Accordingly, it is challenging to effectively integrate deep semantic and structure features for generating the final segmentation map.

To this end, a Cross-level Feature Aggregated Network (CFA-Net) is proposed for polyp segmentation, consisting of a boundary prediction network and a polyp segmentation network. The boundary prediction network is specifically designed to generate boundary-aware features, which are incorporated into the polyp segmentation network in a layer-wise strategy for boosting the segmentation performance. In the polyp segmentation network, a two-stream structure is presented to capture the hierarchical semantic information. In addition, a Cross-level Feature Fusion (CFF) module is proposed to integrate the adjacent features from different levels, in which multi-scale context information can be also captured to deal with the scale variations of polyps. Moreover, a Boundary Aggregated Module (BAM) is presented to effectively incorporate boundary-aware features into the segmentation network. Finally, a unified framework is formulated to simultaneously conduct boundary prediction and polyp segmentation, and the boundary information can be fully captured to enhance the hierarchical features in the segmentation network, leading to finer segmentation results.

The main contributions of this paper are summarized as follows:

- We propose a novel *Cross-level Feature Aggregated Network*, which simultaneously exploits boundary information and captures hierarchical semantic information for accurate segmenting polyps.
- A *Cross-level Feature Fusion* module is proposed to fully utilize the features from adjacent layers, which also conducts cross-level feature fusion at different scales to deal with scale variations.
- We propose a *Boundary Aggregated Module* to capture the boundary context information and then incorporate them into the polyp segmentation network, which can overcome inaccurate boundary prediction to boost the segmentation performance.
- Extensive experiments are conducted on five public colonoscopy datasets, and the results demonstrate that the proposed CFA-Net outperforms the other state-of-the-art polyp segmentation methods. Meanwhile, a comprehensive ablation study validates the effectiveness of all key components in the proposed model.

The rest of this paper is organized as follows. We discuss three types of works related to our model in Section 2. We describe the framework of our proposed CFA-Net for the polyp segmentation in Section 3. In Section 4, we provide the experimental settings, comparison results, and ablation study. Finally, we conclude the paper in Section 5.

2. Related Work

We present a brief overview of the three types of works that are most related to the proposed polyp segmentation method, including medical image segmentation, polyp segmentation, and multi-scale and multi-level fusion.

2.1. Medical Image Segmentation

Medical image segmentation [13,14] plays an important role in identifying interested and affected regions in the computer-aided diagnosis system. Currently, CNN-based methods have presented promising performance in the medical image segmentation field [14–16]. Among these methods, a representative architecture, namely, UNet [17], has gained significant success for biomedical image segmentation, and several variants based on the UNet architecture have been developed to obtain more precise segmentation. For example, Jha et al. [9] proposed a novel framework for medical image segmentation (namely ResUNet++), which is an extended version of ResUNet by integrating additional layers (*e.g.*, squeeze-and-excitation and attention blocks) into the UNet structure. Li et al. [18] presented a hybrid densely connected UNet framework (namely H-DenseUNet), which includes two key components, *i.e.*, a 2-D DenseUNet for extracting intra-slice features, and a 3-D network for hierarchically aggregating volumetric contexts for the follow-up segmentation. To reduce the semantic gap between the encoder and decoder, Zhou et al. [10] proposed UNet++ for biomedical image segmentation, which can effectively alleviate the unknown network depth and design a new skip connection strategy for improving the segmentation performance.

2.2. Polyp Segmentation

Early polyp segmentation methods mainly rely on hand-crafted features [19–22], *e.g.*, color, shape, texture, appearance, or a combination of the above features [5]. After extracting hand-crafted features, these models often train a classifier to detect/segment a polyp from its surroundings. However, they still suffer from unsatisfactory results due to the limited representation capability of hand-crafted features. For example, Ameling et al. [23] adopted texture features, including grey-level-co-occurrence and local binary patterns, to achieve polyp segmentation. Further, the covariances of texture measurements are used to represent different polyp regions [24]. Tajbakhsh et al. [22] proposed an automated polyp detection method from colonoscopy videos, which fully utilizes context and shape to remove non-polyp structures and accurately locate polyps. However, the texture and shape of polyps highly differ in real-world applications, making the traditional methods suffer from unsatisfactory segmentation performance due to the limited-expression ability of hand-crafted features. Recently, the FCN has been widely applied for polyp detection and segmentation tasks. For instance, Akbari et al. [6] proposed a polyp segmentation framework based on a fully CNN and adopted Otsu thresholding to extract the largest connected regions for segmenting polyp regions. Sun et al. [7] proposed an FCN-based polyp segmentation framework, in which a dilated convolution is introduced to learn high-level semantic features without resolution reduction. Moreover, two variants of the UNet architecture, including ResUNet++ [9], and UNet++ [10], have been proposed for polyp segmentation which led to a promising performance. However, the above-mentioned methods often focus on segmenting the whole region of the polyp while neglecting some valuable boundary information. To overcome this problem, Fang et al. [11] designed a boundary-sensitive loss to introduce area-boundary constraints for producing more precise predictions. PsiNet [12] was presented with three parallel decoders, which are designed for three tasks, *i.e.*, contour extraction, mask prediction, and distance map estimation. Nonetheless, the contour information has been captured, which cannot be effectively incorporated into the mask prediction decoder.

2.3. Multi-scale and Multi-level Fusion

Multi-scale feature representation provides an effective solution to deal with the scale variations of objects in detection and segmentation tasks. For instance, Li et al. [25] proposed to utilize different sizes of convolution kernels to adaptively detect multi-scale image features for image super-resolution. Jiang et al. [26] utilized a multi-scale progressive fusion module to fully exploit the inherent correlations among multi-scale rain streaks. He et al. [27] proposed to adaptively capture multi-scale contents for dealing with the scale variations of objects. In addition, multi-level fusion strategies have been developed in several fields of computer vision. For example, feature maps from different levels are adopted with shortcut connections to provide multiple granularities for semantic segmentation [28,29]. In the visual recognition task, deep features from different levels were integrated to boost the fused layer representation [30]. Several works have also been developed to study the integration of multi-level features in the field of saliency detection and camouflaged object detection [31–34]. Moreover, multi-scale features have been captured and validated effectively in the field of medical imaging [35–37]. For example, Sinha et al. [38] adopted a multi-scale strategy to incorporate semantic information at different levels for aggregating the relevant contextual features. Fang et al. [36] designed a pyramid-output network to fully utilize multi-scale features for reducing the gaps between features at different scales. In addition, several works focus on multi-scale feature fusion and aggregation for polyp segmentation [39–41].

3. Methodology

In this section, we first provide an overview of the proposed cross-level aggregation network for polyp segmentation in Sec. 3.1. Then we present the two key components, including the boundary prediction network (Sec. 3.2) and polyp segmentation network (Sec. 3.3). Finally, we present the overall loss function in Sec. 3.4.

3.1. Overview

Fig. 2 illustrates the architecture of the proposed cross-level aggregation network for polyp segmentation, which involves three key parts, *i.e.*, encoder network, boundary prediction network, and two-stream polyp segmentation network. Specifically, a colonoscopy image is first fed into an encoder network (Res2Net-50 [42] as the backbone), to extract multi-level features, which are denoted as F_i $(i = 1, 2, \ldots, 5)$. The feature resolution of $\frac{W}{8} \times \frac{H}{8}$ for the first level, and a general resolution of $\frac{W}{2^m} \times \frac{H}{2^m}$ (when $i > 1$) are obtained in this case. To accurately segment polyps, it is critical to exploit boundary-aware features for boosting the segmentation performance. Therefore, low-level features (*i.e.*, F_1 and F_2) are integrated, followed by constructing a boundary prediction decoder. Then, a two-stream segmentation network is constructed to effectively exploit hierarchical semantic information from cross-level features, in which two adjacent features are fused by using the proposed CFF module. Next, the boundary-aware features can be incorporated into the two segmentation decoders in a layer-wise strategy using the proposed boundary feature aggregation module. The final segmentation results are obtained by combining the outputs of the two segmentation decoders. The details of each key component are provided in the following sections.

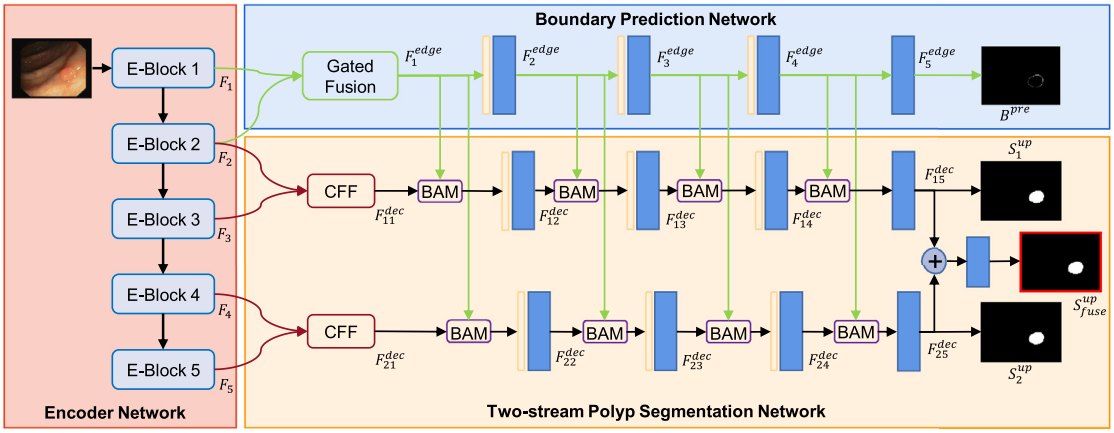

Fig. 2. Overview of the proposed CFA-Net for polyp segmentation, consisting of three key parts: encoder network, boundary prediction network, and two-stream polyp segmentation network. An input image is first passed through the encoder with five E-blocks to extract multi-scale convolutional features. Then, the low-level features are fused via a gated fusion strategy, and then the fused feature is fed into the boundary prediction network, which learns the boundary-aware features and generates the boundary prediction map. Then, the adjacent cross-level features are fused via the proposed cross-level feature fusion module, which is fed into the two-stream segmentation network. A boundary aggregated module is proposed to fully incorporate the boundary-aware features into the segmentation decoders. The final segmentation results are obtained by combing the outputs of two decoders and passing over a convolutional block. Here the orange rectangle denotes upsampled operation, and the blue rectangle represents a 3×3 convolutional layer followed by batch normalization and ReLU activation function.

3.2. Boundary Prediction Network

Polyps are visually embedded in their background, thus the boundary between a camouflaged object and its surrounding background is not sharp. Therefore, effective extracting boundary information can boost the polyp segmentation performance. To this end, a boundary prediction network is proposed to generate boundary maps, in which boundary-aware features can be incorporated into the polyp segmentation network for improving the segmentation performance. Specifically, according to several previous works [43,44], only low-level features preserve sufficient boundary information, thus the two low-level layers are integrated, including F_1 and F_2. Considering that noise could exist in the low-level features, a gated fusion strategy [45] is adopted to integrate the two features. Specifically, the two features are fed into a 3×3 convolutional layer followed by batch normalization and an activation function, respectively, so that to obtain F_1' and F_2'. Subsequently, the two features are concatenated and two separate 1×1 convolutional layers are applied to compute the combined weights. The above processing steps can be depicted as follows:

$$\begin{cases} G_1 = \mathcal{C}_{1\times1}\left(\mathcal{CAT}(F_1', F_2')\right), \\ G_2 = \mathcal{C}_{1\times1}\left(\mathcal{CAT}(F_1', F_2')\right), \end{cases} \tag{1}$$

where $\mathcal{C}_{1\times1}$ and $\mathcal{CAT}$ represent a 1×1 convolutional layer and a concatenation operation, respectively. G_1 and G_2 are spatial-wise gates for the two feature maps. Next, the two gates are further concatenated, and a softmax layer is applied to obtain $W_1^{(i,j)} = e^{G_1^{(i,j)}} / (e^{G_1^{(i,j)}} + e^{G_2^{(i,j)}})$ and $W_2^{(i,j)} = e^{G_2^{(i,j)}} / (e^{G_1^{(i,j)}} + e^{G_2^{(i,j)}})$, where $W_1^{(i,j)} + W_2^{(i,j)} = 1$. Therefore, the fused feature can be obtained by using the gated weight strategy, which is

$$F_1^{edge(i,j)} = W_1^{(i,j)} \cdot F_1^{(i,j)} + W_2^{(i,j)} \cdot F_2^{(i,j)}. \tag{2}$$

Then, the fused low-level feature F_1^{edge} is obtained and further fed into the three convolutional blocks, each consisting of a 3×3 convolutional layer followed by batch normalization and a ReLU activation function. For convenience, the outputs of the three convolutional blocks are denoted as F_2^{edge}, F_3^{edge}, and F_4^{edge}. Subsequently, F_4^{edge} is fed into a 3×3 convolutional layer to generate the boundary map, which is up-sampled to the same resolution as the original image. Thus, the generated boundary map and its detection edge map can be measured using the binary cross-entropy (BCE) loss, which is expressed as

$$\mathcal{L}_{\text{boundary}} = -\sum_i [B_i^{det} \log(B_i^{pre}) + (1 - B_i^{det}) \log(1 - B_i^{pre})], \tag{3}$$

where B_i^{pre} and B_i^{det} imply the predicted and detected boundary maps of the i-th image, respectively. In this study, the Canny edge detection algorithm is used to extract the boundary map of each image. It is worth noting that the boundary prediction network can provide boundary-aware features to enhance polyp segmentation.

3.3. Polyp Segmentation Network

The polyp segmentation network is designed using a two-stream structure with two decoders, in which different scale features can be integrated to capture the hierarchical semantic information. Specifically, F_2 and F_3 are combined and then fed into the first segmentation decoder. Furthermore, F_4 and F_5 are fused and then fed into the second segmentation decoder. Additionally, boundary-aware features (*i.e.*, F_1^{edge}, F_2^{edge}, F_3^{edge}, and F_4^{edge}) are incorporated into the segmentation decoders to boost the segmentation performance. Finally, the outputs of the two decoders are fused to obtain the final segmentation results. To achieve this goal, a CFF module is proposed to effectively fuse the two adjacent features (*e.g.*, F_2 and F_3, as well as F_4 and F_5), and a BAM is presented to incorporate the boundary-aware features into the segmentation decoders. The following sections present the details of the two key components.

3.3.1. Cross-level Feature Fusion Module

Multi-level features at different solutions can be obtained using the feature extraction network. Therefore, it is important to effectively integrate multi-level features, which can boost the representation ability of different scale features. Thus, a CFF module is proposed to fuse the two adjacent features and then feed them into the segmentation network. Specifically, as shown in Fig. 3, the two adjacent features F_i and F_{i+1} are fed into a 1×1 convolutional layer to reduce the channel size, and obtain $S_i \in \mathbb{R}^{W_i * H_i * L}$ and $S_{i+1} \in \mathbb{R}^{W_i * H_i * L}$. Then, the two features from the adjacent layers are cascaded and then fed into a two-bypass network, of which each stream has a different convolutional kernel. In this way, the information between the two-stream network can be shared for

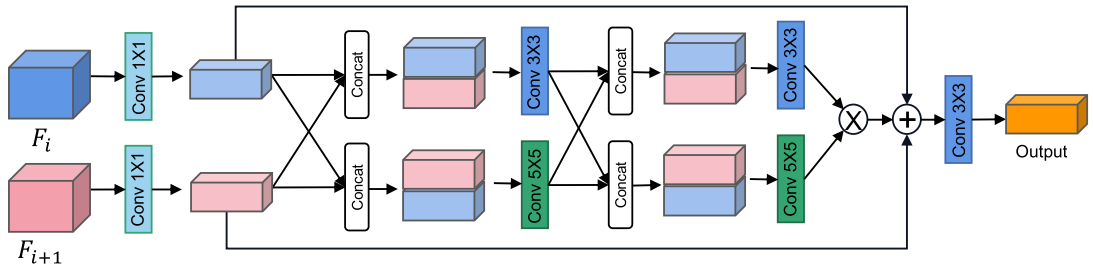

Fig. 3. The flowchart of the proposed cross-level feature fusion module.

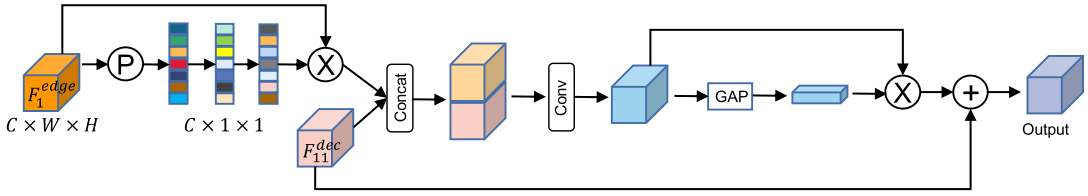

Fig. 4. The flowchart of the proposed boundary aggregated module ("P" denotes a GAP operation).

capturing features from cross-level and multiple scales. The above-mentioned process can be described as follows:

$$\begin{cases} S_{12} = \mathcal{B}_{conv3\times3}\Big(\mathcal{CAT}(S_{11}, S_{21})\Big), \\ S_{22} = \mathcal{B}_{conv5\times5}\Big(\mathcal{CAT}(S_{21}, S_{11})\Big), \end{cases} \tag{4}$$

where $\mathcal{B}_{conv3\times3}(\cdot)$ is a sequential operation that consists of a 3×3 convolutional layer followed by batch normalization and an activation function, and the same settings by using a 5×5 convolutional layer for $\mathcal{B}_{conv5\times5}(\cdot)$. Next, the two multi-scale features S_{12} and S_{22} are further cascaded to be fed into two different convolutional layers, which can be presented as follows:

$$\begin{cases} S_{13} = \mathcal{B}_{conv3\times3}\Big(\mathcal{CAT}(S_{12}, S_{22})\Big), \\ S_{23} = \mathcal{B}_{conv5\times5}\Big(\mathcal{CAT}(S_{22}, S_{12})\Big), \end{cases} \tag{5}$$

To fully fuse the multi-scale and original cross-level features, the two features S_{13} and S_{23} are combined using an element-wise multiplication operation, and then the original cross-levels can be further combined by an addition operation, thus the fused feature can be obtained as

$$F_{fuse} = S_{13} \otimes S_{23} \oplus S_{11} \oplus S_{21}, \tag{6}$$

where $\otimes$ and $\oplus$ represent element-wise product and addition, respectively.

Subsequently, to further smooth the fused feature, it is fed into a sequential operation to obtain the final cross-level fusion feature, namely, $F_{11}^{dec} = \mathcal{B}_{conv3\times3}(F_{fuse})$.

3.3.2. Boundary Aggregated Module

To fully make use of boundary-aware features, two key problems need to be taken into consideration. The redundancy and noise in boundary-aware features should be reduced, and it is important to effectively incorporate boundary-aware features into the segmentation decoder. Therefore, a BAM is presented to address the above-mentioned problems by excavating useful information from boundary-aware features and obtaining the aggregated features to boost the segmentation performance. Specifically, as shown in Fig. 4, the boundary-aware feature is first fed into a channel attention operation, and this process is depicted by

$$F_{1,att}^{edge} = \mathcal{C}_{att}\Big(F_1^{edge}\Big), \tag{7}$$

where $\mathcal{C}_{att}(\cdot)$ denotes the channel attention operation. More specifically, it is implemented by

$$\mathcal{C}_{att}(F) = \mathcal{MLP}(\mathcal{P}_{max}(F)) \otimes F, \tag{8}$$

where $\mathcal{MLP}(\cdot)$ and $\mathcal{P}_{max}$ present a two-layer perceptron and global max pooling (GMP) operation, respectively. Additionally, F denotes an input feature map.

Next, the feature at each layer from the segmentation decoder and the attention enhanced boundary-aware feature are combined using a simple concatenation operation. The concatenated feature (*i.e.*, $F_{11}^{cat} = \mathcal{CAT}(F_{1,att}^{edge}, F_{11}^{dec})$) is processed through a 3×3 convolutional layer and then fed into a global average pooling (GAP) layer. Moreover, the output of the GAP layer is adopted to enhance the concatenated feature. Furthermore, to make the network more efficient and preserve the original information, a residual connection is adopted to combine the previous feature in the current decoder. Therefore, the process can be depicted as follows:

$$F_{11}^{agg} = \mathcal{P}_{ave}\Big(F_{11}^{cat}\Big) \otimes F_{11}^{cat} \oplus F_{11}^{dec}, \tag{9}$$

where F_{11}^{agg} represents the aggregated feature, and it is further processed by an up-sampled operation and fed into a sequential operation $\mathcal{B}_{conv3\times3}$. $\mathcal{P}_{ave}(\cdot)$ denotes a GAP operation. Then, the output is regarded as the input of the following BAM. It is worth noting that the boundary-aware features are well incorporated into the segmentation decoder, thus some useful boundary information can boost the segmentation performance.

The proposed polyp segmentation network involves two decoders, which are supervised by using the ground truth segmentation results. Further, the output features (*i.e.*, F_{15}^{dec} and F_{25}^{dec}) of the two decoders are cascaded and then fed into a sequential operation $\mathcal{B}_{conv3\times3}$, which produces the final segmentation results.

3.4. Loss Function

The binary cross-entropy (BCE) loss is widely used in several segmentation tasks, however, it ignores the global structure of an image when computing the loss for each pixel independently. To overcome these issues, our polyp segmentation loss function is defined as $\mathcal{L}_{seg} = \mathcal{L}_{wIoU} + \mathcal{L}_{wBCE}$, where $\mathcal{L}_{wIoU}$ and $\mathcal{L}_{wBCE}$ indicate the weighted IoU (wIoU) loss and BCE (wBCE) loss for the global and local restrictions [46], respectively. Specifically, the wIoU loss is defined by

$$\mathcal{L}_{wIoU} = 1 - \frac{\sum_{i=1}^{W}\sum_{j=1}^{H}(g_{i,j} * p_{i,j}) * (1+5\mu_{i,j})}{\sum_{i=1}^{W}\sum_{j=1}^{H}(g_{i,j} + p_{i,j} - g_{i,j} * p_{i,j}) * (1+5\mu_{i,j})}, \tag{10}$$

where $g_{i,j}$ and $p_{i,j}$ are the values at pixel (i, j) of the ground truth and predicted segmentation maps, respectively. $\mu_{i,j}$ denotes the pixel importance, which can be calculated by a pixel and its surrounding pixels [46].

The wBCE loss is defined by

$$\mathcal{L}_{\text{wBCE}} = -\frac{\sum_{i=1}^{W}\sum_{j=1}^{H}(1+5\mu_{i,j})\sum_{l=0}^{l}\mathbf{1}(g_{i,j}=l)\log\mathbf{Pr}(p_{i,j}=l|\Phi)}{\sum_{i=1}^{W}\sum_{j=1}^{H}5\mu_{i,j}}, \tag{11}$$

where $\mathbf{1}$ is an indicator function, and the notation $l \in \{0, 1\}$ indicates two kinds of labels. Φ denotes all the parameters of the model and $\mathbf{Pr}(p_{i,j} = l|\Phi)$ represents the predicted probability.

It is noteworthy that $\mathcal{L}_{\text{wIoU}}$ can increase the weights of hard pixels to highlight their importance, and $\mathcal{L}_{\text{wBCE}}$ pays more attention to hard pixels rather than treating all pixels equally. Moreover, three segmentation maps (*i.e.*, S_1^{up}, S_2^{up}, and S_{fuse}^{up}) are up-sampled to have the same size as the ground truth map (*i.e.*, G).

Finally, the overall loss function can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\mathcal{L}_{\text{boundary}} + \\ &\mathcal{L}_{\text{seg}}(G, S_1^{up}) + \mathcal{L}_{\text{seg}}(G, S_2^{up}) + \mathcal{L}_{\text{seg}}(G, S_{fuse}^{up}).\end{aligned} \tag{12}$$

4. Experiments

We provide the details of datasets and evaluation metrics (Sec. 4.1), as well as implementation details (Sec. 4.2). Then, we compare the proposed model with the state-of-the-art polyp segmentation methods in Sec. 4.3. To clarify the validity of the key components in our model, we conduct ablation experiments in Sec. 4.4. Further, we present failure cases and provide limitation discussions in Sec. 4.5.

4.1. Datasets and Evaluation Metrics

4.1.1. Datasets

To validate the effectiveness of the proposed segmentation model, we conduct comparison experiments on five benchmark datasets. The details of each dataset are provided below. • **CVC-ClinicDB** [47]: This dataset contains 612 images collected from colonoscopy video sequences, whose resolutions are 288×384. • **ETIS** [48]: This dataset includes 196 polyp images with a size of 966 × 1225. • **CVC-ColonDB** [22]: This dataset consists of 380 images with a size of 500 × 570. • **Kvasir** [49]: This dataset includes 1,000 polyp images, which are collected from several colonoscopy video sequences. • **CVC-300** [50]: This dataset contains 60 polyp images with a size of 500×574.

4.1.2. Evaluation Metrics

We adopt four widely used metrics [51], including mean dice score (Dice), mean intersection over union (IoU), specificity (SPE), and sensitivity (SEN). Additionally, four metrics are introduced, which are widely used in the field of object detection [32,52,53], including S-measure (S_α) [54], F-measure [55] (F_β^w), E_ϕ [56], and mean absolute error ($\mathcal{M}$) [57]. The details of the four evaluation metrics are provided below.

- S_α is used to evaluate the structural similarity between the regional perception (S_r) and object perception (S_o), which is defined by $S_\alpha = \alpha \times S_o + (1-\alpha) \times S_r$ (α is a trade-off parameter and it is set to 0.5 as default [54]).
- F_β is defined by $F_\beta = \left(1+\beta^2\right)\frac{P\times R}{\beta^2 P+R}$, where β is set to 1. In our experiments, we utilize the improved version of F_β, namely, weighted F-measure (F_β^w), which can be proven to overcome the interpolation, dependency, and equal-importance flaws of F_β.
- E_ϕ captures image-level statistics and their local pixel matching information. It can be defined by $E_\phi = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{i=1}^{H}\phi_{FM}(i, j)$, where ϕ_{FM} denotes the enhanced-alignment matrix [56].
- $\mathcal{M}$ [57] is adopted to evaluate the difference between the ground truth and the normalized prediction, and it is defined by $\mathcal{M} = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{i=1}^{H}|S(i, j) - G(i, j)|$, where G and S present the ground truth and normalized prediction (they are normalized to [0,1]).

4.2. Implementation Details

The proposed framework is implemented in PyTorch, which is trained using one NVIDIA Tesla P40 GPU with 24 GB memory. In our model, Res2Net-50 [42] is used as the backbone network, which has been pre-trained on ImageNet [58]. In addition, the Adam algorithm is used to optimize the proposed model. The initial learning rate is set to $1e-4$ and is divided by 10 every 30 epoch. Different data augmentation strategies are adopted, including random flipping, crop, and rotation. The input images are resized to 352×352, and we also train our model using different scaling ratios, *i.e.*, $\{0.75, 1, 1.25\}$. The batch size is set to 10 and the model has trained over 100 epochs. To train the proposed model, we follow the same experimental settings in [5], where 900 images from the Kvasir and 550 images from the CVC-ClinicDB are collected to form the training set. The remaining images from the two datasets (*i.e.*, Kvasir and CVC-ClinicDB) and other three datasets (*i.e.*, ETIS, CVC-ColonDB, and CVC-300) are adopted for testing. During the testing stage, the test images are resized to 352×352 and then fed into the model to obtain the segmentation maps. The segmentation maps are rescaled to the original size to conduct the final evaluation.

4.3. Comparison to State-of-the-art Methods

4.3.1. Comparison Methods

To evaluate the effectiveness of the proposed polyp segmentation method, we compare it with six state-of-the-art methods, including UNet [10], UNet++ [10], SFA [11], PraNet [11], MSNet [59], and C2FNet [60]. The results of UNet [10], UNet++ [10], SFA [11], and PraNet [11] are collected from https://github.com/DengPingFan/PraNet. The results of MSNet [59] are collected from the original paper. For C2FNet, we retrained and tested based on the released codes using the recommended parameters.

4.3.2. Quantitative Comparison

Table 1 provides the quantitative comparison between our model and six state-of-the-art methods in terms of eight evaluation metrics on the CVC-ClinicDB [47] and Kvasir [49] datasets. On the CVC-ClinicDB dataset, it can be observed that our model outperforms all compared methods. MSNet and C2FNet achieve relatively better segmentation performance than the other comparison methods. Furthermore, SFA takes into account the dependency between the region and boundary, but it still fails to segment polyps, while our model can effectively segment them and achieve the best performance. This is because our model can fully capture the multi-scale information to deal with the scale variations of polyps, and the boundary-aware features provide boundary cues to boost the segmentation performance. On the Kvasir dataset, the proposed polyp segmentation method consistently obtains the best performance. For instance, in terms of mDice, mIou, S_α, F_β^w, and E_ϕ, our model achieves 3.3%, 3.6%, 2.1%, 3.8%, and 2.9% improvements over C2FNet.

Table 2 shows the quantitative comparison between our model and six state-of-the-art methods on the CVC-300 [50] and ColonDB [22] datasets. From the results, it can be observed that our method performs better than other segmentation approaches. For example, compared with PraNet, our method achieves the improvements are 2.5%, 3.8%, 3.8% in the terms of mDice, mIoU, and F_β^w on the CVC-300 dataset. On the ColonDB dataset, our method

Table 1
Quantitative polyp segmentation results on the CVC-ClinicDB and Kvasir datasets using eight metrics. "↑" & "↓" indicate that larger or smaller is better. The best two results are shown in red and blue fonts.

Methods	CVC-ClinicDB [47]								Kvasir [49]							
	mDice ↑	mIou ↑	SPE ↑	SEN ↑	S_α ↑	F_β^w ↑	E_ϕ ↑	$\mathcal{M}$ ↓	mDice ↑	mIou ↑	SPE ↑	SEN ↑	S_α ↑	F_β^w ↑	E_ϕ ↑	$\mathcal{M}$ ↓
UNet [17]	0.823	0.755	0.947	0.835	0.889	0.811	0.954	0.019	0.818	0.746	0.950	0.857	0.858	0.794	0.893	0.055
UNet+ [10]	0.794	0.729	0.927	0.795	0.873	0.785	0.931	0.022	0.821	0.744	0.986	0.807	0.862	0.808	0.910	0.048
SFA [11]	0.700	0.607	0.919	0.802	0.793	0.647	0.885	0.042	0.723	0.611	0.965	0.799	0.782	0.670	0.849	0.075
PraNet [5]	0.899	0.849	0.990	0.911	0.936	0.896	**0.979**	0.009	0.898	0.840	0.978	**0.912**	0.915	0.885	0.948	0.030
MSNet [59]	0.918	0.869	**0.975**	0.933	**0.946**	**0.913**	**0.979**	**0.008**	**0.905**	**0.849**	**0.981**	0.911	**0.923**	**0.892**	**0.954**	**0.028**
C2FNet [60]	**0.919**	**0.872**	0.974	**0.941**	0.941	0.906	0.976	0.009	0.886	0.831	0.974	0.904	0.905	0.870	0.935	0.036
CFA-Net (Ours)	**0.933**	**0.883**	**0.991**	**0.960**	**0.950**	**0.924**	**0.989**	**0.007**								
	0.915	**0.861**	**0.985**	**0.926**	**0.924**	**0.903**	**0.962**	**0.023**								

Table 2
Quantitative polyp segmentation results on the CVC-300 and ColonDB datasets using eight metrics. "↑" & "↓" indicate that larger or smaller is better. The best two results are shown in red and blue fonts.

Methods	CVC-300 [50]								ColonDB [22]							
	mDice ↑	mIou ↑	SPE ↑	SEN ↑	S_α ↑	F_β^w ↑	E_ϕ ↑	$\mathcal{M}$ ↓	mDice ↑	mIou ↑	SPE ↑	SEN ↑	S_α ↑	F_β^w ↑	E_ϕ ↑	$\mathcal{M}$ ↓
UNet [17]	0.710	0.627	0.966	0.768	0.843	0.684	0.876	0.022	0.504	0.436	0.798	0.525	0.710	0.491	0.781	0.059
UNet+ [10]	0.707	0.624	0.957	0.738	0.839	0.687	0.898	0.018	0.482	0.408	0.828	0.497	0.693	0.467	0.764	0.061
SFA [11]	0.467	0.329	0.935	0.889	0.640	0.341	0.817	0.065	0.456	0.337	0.861	0.703	0.629	0.366	0.754	0.094
PraNet [5]	0.871	0.797	0.988	**0.941**	0.925	0.843	**0.972**	0.010	0.712	0.640	0.874	0.740	0.820	0.699	0.872	0.043
MSNet [59]	0.865	0.799	0.988	0.931	0.926	**0.848**	0.953	0.010	**0.751**	**0.671**	0.931	**0.775**	**0.838**	**0.736**	**0.883**	**0.041**
C2FNet [60]	**0.874**	**0.801**	0.988	**0.952**	**0.927**	0.844	0.968	**0.009**	0.724	0.650	0.894	0.752	0.826	0.705	0.868	0.044
CFA-Net (Ours)	**0.893**	**0.827**	**0.990**	**0.952**	**0.938**	**0.875**	**0.978**	**0.008**	**0.743**	**0.665**	**0.953**	**0.762**	**0.835**	**0.728**	**0.898**	**0.039**

Table 3
Quantitative polyp segmentation results on the ETIS dataset using eight metrics. "↑" & "↓" indicate that larger or smaller is better. The best two results are shown in red and blue fonts.

Methods	ETIS [48]							
	mDice ↑	mIou ↑	SPE ↑	SEN ↑	S_α ↑	F_β^w ↑	E_ϕ ↑	$\mathcal{M}$ ↓
UNet [17]	0.398	0.335	0.703	0.484	0.684	0.366	0.740	0.036
UNet+ [10]	0.401	0.344	0.727	0.415	0.683	0.390	0.776	0.035
SFA [11]	0.297	0.217	0.781	0.633	0.557	0.231	0.633	0.109
PraNet [5]	0.628	0.567	0.805	0.688	0.794	0.600	0.841	0.031
MSNet [59]	**0.723**	**0.652**	**0.893**	**0.796**	**0.845**	**0.677**	**0.890**	**0.020**
C2FNet [60]	0.699	0.624	0.902	0.745	**0.827**	0.668	0.875	0.022
CFA-Net (Ours)	**0.732**	**0.655**	**0.910**	**0.804**	**0.845**	**0.693**	**0.892**	**0.014**

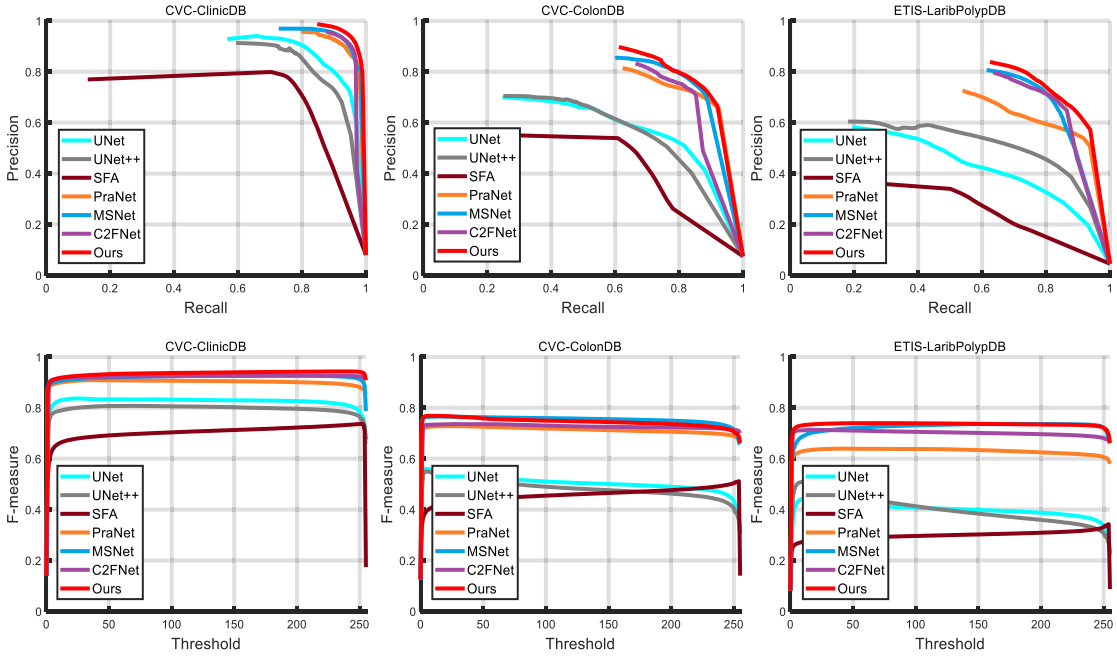

Fig. 5. Precision-Recall and F-measure curves of our model and other six state-of-the-art methods across three datasets (*i.e.*, CVC-ClinicDB, CVC-ColonDB, and ETIS).

achieves 2.6%, 6.6%, 3.3%, and 3.5% over C2FNet in terms of mDice, mIou, F_β^w, and E_ϕ. This is because our model can provide boundary-aware features to help locate the boundaries of polyps, resulting in accurate segmentation of polyps. Table 3 provides the quantitative comparison between our model and six state-of-the-art methods on the ETIS [48] dataset. According to the results, the effectiveness of our method can be further validated.

In addition to the overall quantitative comparisons using the above evaluation metrics, precision-Recall and F-measure curves are further presented in Fig. 5 and Fig. 6. From the results, the proposed model achieves much better results compared to the other state-of-the-art polyp segmentation methods.

4.3.3. Qualitative Comparison

Fig. 7 depicts the segmentation results by comparing our model with six state-of-the-art polyp segmentation methods. Based on the visual results, the results of our model are closest to the ground truth maps, and our method outperforms the other compared methods in dealing with different challenging factors. Specifically, in the 1st and 2nd rows, the polyps have extremely small sizes, and our method still can accurately segment small polyps. However, UNet and UNet++ completely fail to segment them. In this case, SFA, MSNet, and C2FNet produce several errors with over-segmented regions. In the 3rd and 4th rows, the polyps have different shapes and large sizes (*e.g.*, in the 4th row), making it challenging to accurately segment polyps. Accordingly, SFA and UNet++ perform worse than the other methods. In the 5th and 6th rows, the boundaries between the polyps and background are not sharp since the polyps are visually embedded in their background, thus it is highly challenging for segmentation methods to identify them. In this case, our method segments polyps more accurately than the other compared methods. Overall, the visual results further demonstrate that our model can achieve good performance in handling different challenging factors for polyp segmentation.

Moreover, we visualize the predicted edges and segmentation results using the proposed model in Fig. 8. From the results, we can be observed that the boundary extraction network can effectively predict the main edge parts of polyps. Although some fine details are missing, the boundary-aware features can still capture edge information for boosting the segmentation performance. Overall, the proposed boundary extraction network can learn boundary-aware features and its effectiveness has been well validated.

4.3.4. Model Complexity and Inference Time Comparison

To investigate the model complexity and inference time, we report the model sizes and inference times of our model and other compared methods in Table 4. In Table 4, #Param is measured in million (M), floating point operations (FLOPs) are measured in Giga (G), and the inference time is measured by frames per second (FPS). As can be observed, our model is with minimal parameters in comparison with the other methods. Because our model

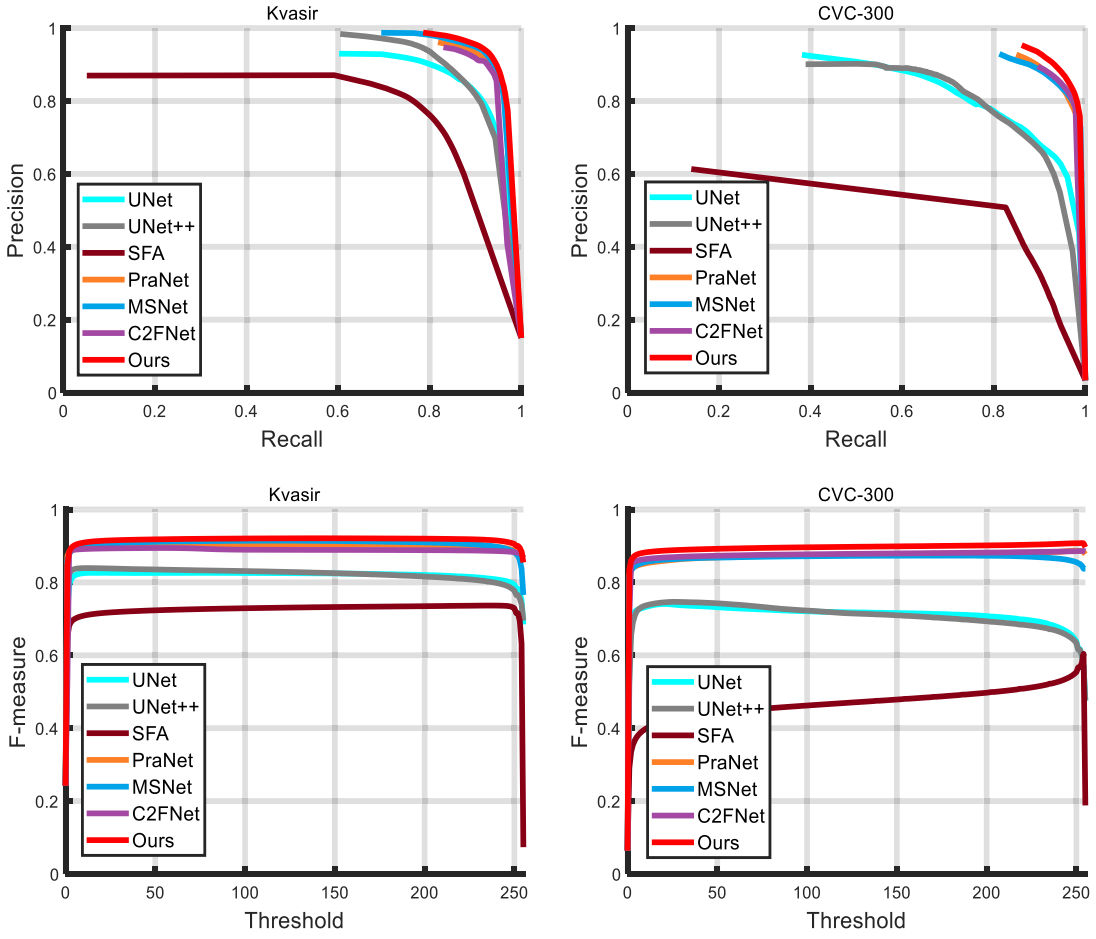

Fig. 6. Precision-Recall and F-measure curves of our model and other six state-of-the-art methods across two datasets (*i.e.*, Kvasir and CVC-300).

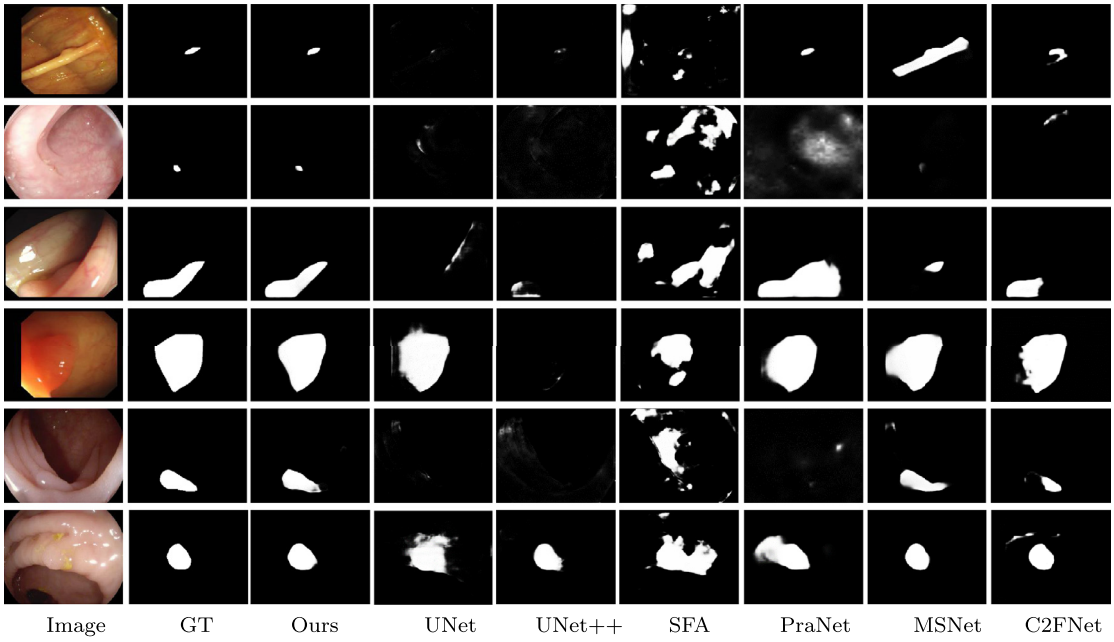

Fig. 7. Qualitative visualization of polyp segmentation results comparing our model with six state-of-the-art methods, including UNet [17], UNet++ [10], SFA [11], PraNet [5], MSNet [59], and C2FNet [60].

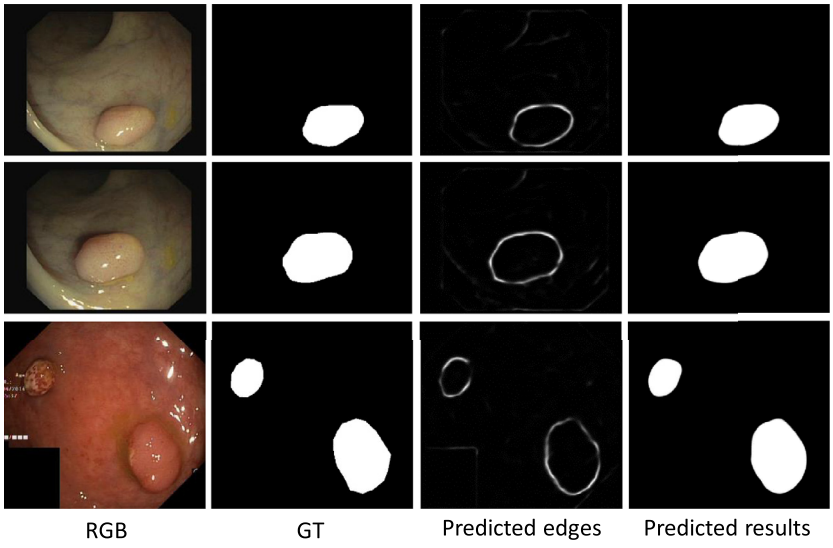

Fig. 8. Visualization of the predicted edges and segmentation results using the proposed model.

Table 4
Comparison of model size and inference time.

Models	UNet [17]	UNet+ [10]	PraNet [5]	MSNet [59]	C2FNet [60]	CFA-Net (ours)
Speed (FPS)	123.11	82.51	25.05	31.08	20.93	23.50
FLOps	123.87	262.16	13.15	17.00	13.16	55.36
Param (M)	34.52	36.63	30.50	27.69	26.36	25.24

adopts a boundary prediction network and a polyp segmentation network to generate the boundary map and segmentation maps, respectively, it takes much more inference time for the polyp segmentation than other compared methods. Therefore, we can design lightweight networks to improve the efficiency of the proposed model for real-time polyp segmentation in future work.

4.4. Ablation Study

Effectiveness of Boundary-aware Features. In our model, the boundary prediction network is designed to generate boundary-aware features, which are incorporated into the segmentation network for providing the boundary context information. To investigate the effectiveness of the boundary prediction network, we perform ablation studies by removing it from our model, denoted as "w/o Boundary". Additionally, we adopt a simple and effective gate fusion strategy to integrate low-level features (*i.e.*, $\mathbf{F}_1$ and $\mathbf{F}_2$). To validate its effectiveness, we utilize the concatenation and addition operations to replace the gate fusion strategy, which are denoted as "Concat" and "Addition". The experimental results of ablation studies are provided in Table 5. As shown in Table 5, compared "w/o boundary" with our full model, it can be observed that our method using the boundary-aware features can improve the segmentation performance. Moreover, our model, using the gate fusion strategy, could perform better compared to using concatenation or addition operations. This is probably because low-level features contain some noises, and the gate fusion strategy helps to filter out these noises and then enhance the features. The visual comparison results in Fig. 9 further indicate that boundary-aware features can improve the segmentation performance.

Effectiveness of CFF Module. To validate the effectiveness of the CFF module in our model, we directly utilize a concatenation operation to fuse the two adjacent features, followed by a 3×3 convolutional layer, which is further fed to the segmentation network, denoted as "w/o CFF". Based on the results (Table 5), the CFF module could enable our method to accurately segment polyp regions. This is because the proposed CFF module could effectively fuse cross-level features and capture multi-scale information for dealing with scale variations. The visual comparison results in Fig. 9 further indicate that the proposed CFF module can boost the segmentation performance.

Effectiveness of BAM. To validate the effectiveness of BAM in our model, we directly utilize a concatenation operation to incorporate boundary-aware features into the segmentation network, denoted as "w/o BAM". As shown in Table 5, the proposed BAM boosts the segmentation performance, highlighting the effectiveness of the proposed BAM in incorporating the boundary cues into the segmentation network. As shown in Fig. 9, without using the proposed BAM, some boundary details can not be accurately detected.

Effectiveness of Two-stream Structure in Segmentation Network. In the proposed model, we adopt a two-stream decoder structure in the polyp segmentation network, which can effectively exploit the hierarchical semantic information. To validate the effectiveness of the two-stream structure, we compare it with the proposed method using a one-stream structure (as shown in Fig. 10). As shown in Table 5, the results confirm that our model outperforms when using the two-stream structure rather than only using a one-stream structure, indicating that much hierarchical semantic information can be exploited to boost the segmentation performance. Moreover, the visual comparison results in Fig. 9 also demonstrate that our model can accurately segment polyps.

4.5. Failure Cases and Limitations

The qualitative and quantitative evaluations show the effectiveness and superiority of the proposed CFA-Net, however, our CFA-Net fails to segment polyps when dealing with some challenging scenes such as big polyps and complex backgrounds. Some failure cases of our model are shown in Fig. 11. In the 1st and 2nd rows, we can see that the polyps have a big size, which makes our model can segment the coarse regions without fine boundary details. In

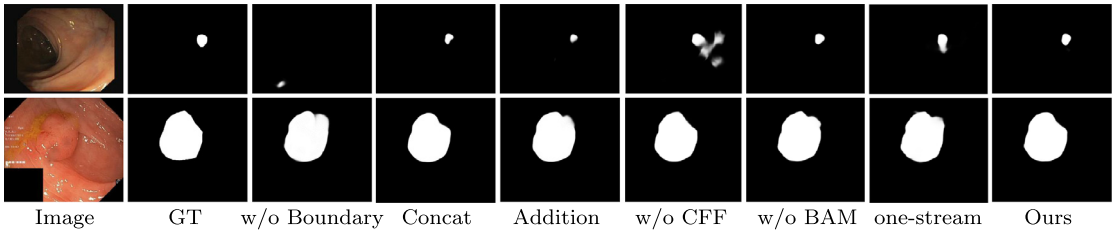

Fig. 9. Visual comparisons for validating the benefits of different modules.

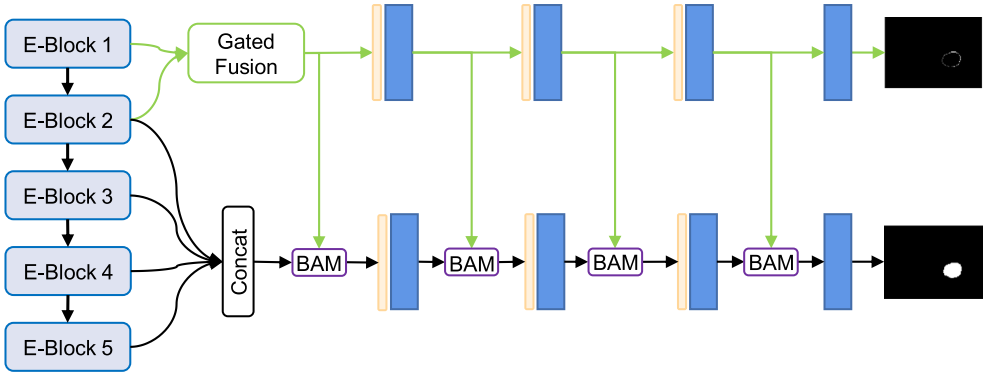

Fig. 10. The architecture of our method uses a one-stream structure in the segmentation network.

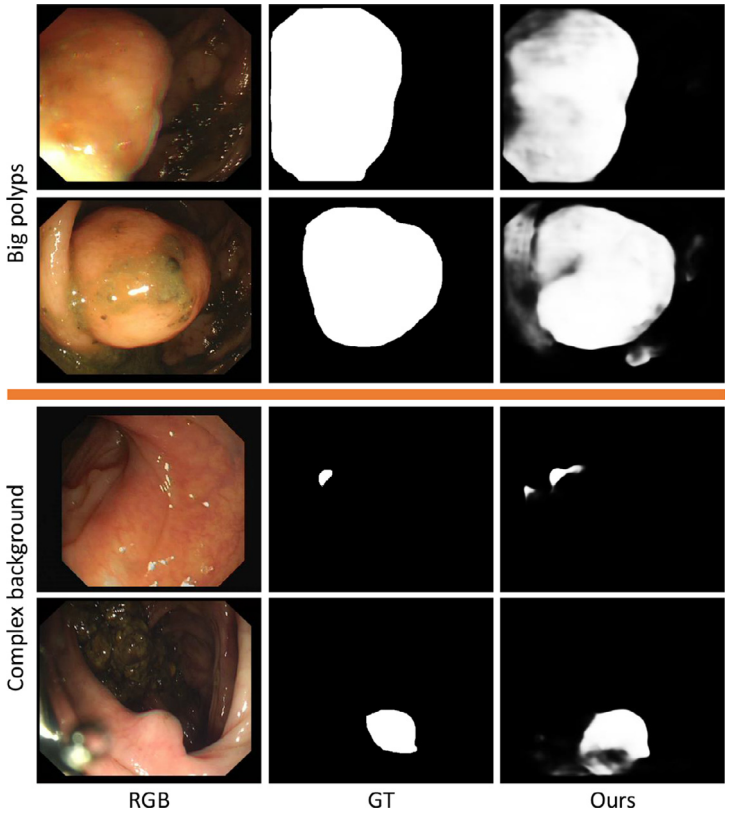

Fig. 11. Some failure cases of the proposed CFA-Net.

the 3rd and 4th rows, it can be seen that polyps have a similar appearance to background regions in the scene, which makes it challenging to accurately segment polyps without sharp boundaries. As a result, our CFA-Net fails to accurately locate and segment the polyps under this condition. Therefore, dealing with large-scale variation and segmenting polyps under complex backgrounds will be investigated in future work. Moreover, it is worth noting that a real-time detection system with high accuracy is needed in clinical practice, which can help doctors take necessary action during colonoscopy procedures. Although our CFA-Net has achieved satisfactory segmentation performance, it still requires a huge computational cost. In the future, we can compress the proposed CFA-Net by network pruning and knowledge distillation [61] to develop a lightweight network for real-time polyp segmentation in clinics.

Table 5
Ablation study on the effectiveness of different key components using the CVC-ClinicDB and Kvasir datasets.

Methods	CVC-ClinicDB [47]								Kvasir [49]							
	mDice ↑	mIou ↑	SPE ↑	SEN ↑	S_α ↑	F_β^w ↑	E_ϕ ↑	$\mathcal{M}$ ↓	mDice ↑	mIou ↑	SPE ↑	SEN ↑	S_α ↑	F_β^w ↑	E_ϕ ↑	$\mathcal{M}$ ↓
w/o Boundary	0.894	0.839	0.958	0.921	0.931	0.885	0.977	0.011	0.876	0.814	0.969	0.889	0.901	0.861	0.936	0.037
Concat	0.926	0.872	0.990	0.954	0.945	0.918	0.984	0.009	0.894	0.837	0.984	0.910	0.911	0.881	0.945	0.027
Addition	0.922	0.871	0.991	0.941	0.943	0.916	0.985	0.007	0.895	0.836	0.982	0.907	0.909	0.881	0.949	0.030
w/o CFF	0.916	0.863	0.990	0.960	0.940	0.902	0.985	0.009	0.898	0.842	**0.986**	0.900	0.910	0.891	0.955	0.029
w/o BAM	0.910	0.856	0.976	0.925	0.935	0.907	0.980	0.013	0.903	0.847	0.981	0.913	0.915	0.890	0.954	0.028
One-stream	0.910	0.856	0.989	0.944	0.936	0.901	0.976	0.012	0.892	0.837	0.983	0.903	0.910	0.880	0.951	0.030
CFA-Net (Ours)	**0.933**	**0.883**	**0.991**	**0.960**	**0.950**	**0.924**	**0.989**	**0.007**	**0.915**	**0.861**	0.985	**0.926**	**0.924**	**0.903**	**0.962**	**0.023**

5. Conclusion

In this paper, we present a Cross-level Feature Aggregation Network (CFA-Net) for Polyp Segmentation. Specifically, we first propose a boundary prediction network to learn boundary-aware features, which capture boundary information to boost the segmentation performance. To effectively exploit hierarchical semantic information, we propose a two-stream segmentation network. In the segmentation network, we propose a Cross-level Feature Fusion (CFF) module to fuse cross-level features and exploit multi-scale context information for handling scale variations. Furthermore, we propose a Boundary Aggregated Module (BAM) to fully incorporate the boundary cues into the segmentation network. Experiments on five public datasets demonstrate that our CFA-Net outperforms other state-of-the-art methods, and a comprehensive ablation study has validated the effectiveness of all key components in the proposed CFA-Net.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Nos. 62172228, 62106043, 61973162), NSF of Jiangsu Province (No: BZ2021013), CAAI-Huawei MindSpore Open Fund, NSF for Distinguished Young Scholar of Jiangsu Province (No: BK20220080), the Fundamental Research Funds for the Central Universities (Nos: 30920032202, 30921013114), and an Open Project of the Key Laboratory of System Control and Information Processing, Ministry of Education (Shanghai Jiao Tong University, ID: Scip202102) and A*STAR Central Research Fund.

References

[1] J. Bernal, J. Sánchez, F. Vilarino, Towards automatic polyp detection with a polyp appearance model, Pattern Recognition 45 (9) (2012) 3166–3182.
[2] M. Navarro, A. Nicolas, A. Ferrandez, A. Lanas, Colorectal cancer population screening programs worldwide in 2016: An update, World journal of gastroenterology 23 (20) (2017) 3632.
[3] R. Zhang, Y. Zheng, C.C. Poon, D. Shen, J.Y. Lau, Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker, Pattern Recognition 83 (2018) 209–219.
[4] X. Guo, C. Yang, Y. Liu, Y. Yuan, Learn to threshold: Thresholdnet with confidence-guided manifold mixup for polyp segmentation, IEEE Transactions on Medical Imaging 40 (4) (2020) 1134–1146.
[5] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020, pp. 263–273.
[6] M. Akbari, M. Mohrekesh, E. Nasr-Esfahani, S.R. Soroushmehr, N. Karimi, S. Samavi, K. Najarian, Polyp segmentation in colonoscopy images using fully convolutional network, in: International Conference of the IEEE Engineering in Medicine and Biology Society, 2018, pp. 69–72.
[7] X. Sun, P. Zhang, D. Wang, Y. Cao, B. Liu, Colorectal polyp segmentation by u-net with dilation convolution, in: IEEE International Conference On Machine Learning And Applications, 2019, pp. 851–858.
[8] D.V. Sang, T.Q. Chung, P.N. Lan, D.V. Hang, D. Van Long, N.T. Thuy, AG-CUResNeSt: A novel method for colon polyp segmentation, arXiv preprint arXiv:2105.00402 (2021).
[9] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H.D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: IEEE International Symposium on Multimedia, 2019, pp. 225–2255.
[10] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Transactions on Medical Imaging 39 (6) (2019) 1856–1867.
[11] Y. Fang, C. Chen, Y. Yuan, K.-y. Tong, Selective feature aggregation network with area-boundary constraints for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 302–310.

[12] B. Murugesan, K. Sarveswaran, S.M. Shankaranarayana, K. Ram, J. Joseph, M. Sivaprakasam, Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation, in: International Conference of the IEEE Engineering in Medicine and Biology Society, 2019, pp. 7223–7226.
[13] H. Jia, P.-T. Yap, D. Shen, Iterative multi-atlas-based multi-image segmentation with tree-based registration, NeuroImage 59 (1) (2012) 422–430.
[14] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, et al, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88.
[15] D. Zhang, G. Huang, Q. Zhang, J. Han, J. Han, Y. Yu, Cross-modality deep feature learning for brain tumor segmentation, Pattern Recognition 110 (2021) 107562.
[16] A. Oulefki, S. Agaian, T. Trongtirakul, A.K. Laouar, Automatic COVID-19 lung infected region segmentation and measurement using CT-scans images, Pattern Recognition 114 (2021) 107747.
[17] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.
[18] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, P.-A. Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, IEEE Transactions on Medical Imaging 37 (12) (2018) 2663–2674.
[19] C. Van Wijk, V.F. Van Ravesteijn, F.M. Vos, L.J. Van Vliet, Detection and segmentation of colonic polyps on implicit isosurfaces by second principal curvature flow, IEEE Transactions on Medical Imaging 29 (3) (2010) 688–698.
[20] S.Y. Park, D. Sargent, I. Spofford, K.G. Vosburgh, A. Yousif, et al., A colon video analysis framework for polyp detection, IEEE Transactions on Biomedical Engineering 59 (5) (2012) 1408–1418.
[21] A.V. Mamonov, I.N. Figueiredo, P.N. Figueiredo, Y.-H.R. Tsai, Automated polyp detection in colon capsule endoscopy, IEEE Transactions on Medical Imaging 33 (7) (2014) 1488–1502.
[22] N. Tajbakhsh, S.R. Gurudu, J. Liang, Automated polyp detection in colonoscopy videos using shape and context information, IEEE Transactions on Medical Imaging 35 (2) (2015) 630–644.
[23] S. Ameling, S. Wirth, D. Paulus, G. Lacey, F. Vilarino, Texture-based polyp detection in colonoscopy, in: Bildverarbeitung für die Medizin 2009, 2009, pp. 346–350.
[24] S.A. Karkanis, D.K. Iakovidis, D.E. Maroulis, D.A. Karras, M. Tzivras, Computer-aided tumor detection in endoscopic video using color wavelet features, IEEE Transactions on Information Technology in Biomedicine 7 (3) (2003) 141–152.
[25] J. Li, F. Fang, K. Mei, G. Zhang, Multi-scale residual network for image super-resolution, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 517–532.
[26] K. Jiang, Z. Wang, P. Yi, C. Chen, B. Huang, Y. Luo, J. Ma, J. Jiang, Multi-scale progressive fusion network for single image deraining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 8346–8355.
[27] J. He, Z. Deng, Y. Qiao, Dynamic multi-scale filters for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3562–3572.
[28] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 447–456.
[29] Y. Pang, Y. Li, J. Shen, L. Shao, Towards bridging semantic gap to improve semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4230–4239.
[30] Z. Wang, H. Liu, J. Tang, S. Yang, G.Y. Huang, Z. Liu, Learning multi-level dependencies for robust word recognition, in: Proceedings of the Conference on Artificial Intelligence, volume 34, 2020, pp. 9250–9257.
[31] H. Chen, Y. Li, Progressively complementarity-aware fusion network for rgb-d salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3051–3060.
[32] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, L. Shao, RGB-D salient object detection: A survey, Computational Visual Media 7 (1) (2021) 37–69.
[33] T. Zhou, Y. Zhou, C. Gong, J. Yang, Y. Zhang, Feature aggregation and propagation network for camouflaged object detection, IEEE Transactions on Image Processing (2022) (2022).
[34] G. Chen, S.-J. Liu, Y.-J. Sun, G.-P. Ji, Y.-F. Wu, T. Zhou, Camouflaged object detection via context-aware cross-level fusion, IEEE Transactions on Circuits and Systems for Video Technology (2022) (2022).
[35] J. Fan, X. Cao, P.-T. Yap, D. Shen, Birnet: Brain image registration using dual-supervised fully convolutional networks, Medical Image Analysis (2019).
[36] X. Fang, P. Yan, Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction, IEEE Transactions on Medical Imaging 39 (11) (2020) 3619–3629.
[37] Q. Shao, L. Gong, K. Ma, H. Liu, Y. Zheng, Attentive ct lesion detection using deep pyramid inference with multi-scale booster, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 301–309.
[38] A. Sinha, J. Dolz, Multi-scale self-guided attention for medical image segmentation, IEEE Journal of Biomedical and Health Informatics 25 (1) (2020) 121–130.
[39] D. Banik, D. Bhattacharjee, M. Nasipuri, A multi-scale patch-based deep learning system for polyp segmentation, in: Advanced Computing and Systems for Security, 2020, pp. 109–119.
[40] S. Wang, Y. Cong, H. Zhu, X. Chen, L. Qu, H. Fan, Q. Zhang, M. Liu, Multi-scale context-guided deep network for automated lesion segmentation with endoscopy images of gastrointestinal tract, IEEE Journal of Biomedical and Health Informatics 25 (2) (2020) 514–525.
[41] Y. Lin, J. Wu, G. Xiao, J. Guo, G. Chen, J. Ma, Bsca-net: Bit slicing context attention network for polyp segmentation, Pattern Recognition 132 (2022) 108917.
[42] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (2) (2021) 652–662.
[43] Z. Zhang, H. Fu, H. Dai, J. Shen, Y. Pang, L. Shao, ET-Net: A generic edge-attention guidance network for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019, pp. 442–450.
[44] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Inf-net: Automatic COVID-19 lung infection segmentation from CT images, IEEE Transactions on Medical Imaging 39 (8) (2020) 2626–2637.
[45] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, G. Zeng, Bi-directional cross–modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 561–577.
[46] J. Wei, S. Wang, Q. Huang, F^3Net: Fusion, feedback and focus for salient object detection, in: Proceedings of the Conference on Artificial Intelligence, volume 34, 2020, pp. 12321–12328.
[47] J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Computerized Medical Imaging and Graphics 43 (2015) 99–111.
[48] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, International Journal of Computer Assisted Radiology and Surgery 9 (2) (2014) 283–293.
[49] D. Jha, P.H. Smedsrud, M.A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H.D. Johansen, Kvasir-seg: A segmented polyp dataset, in: International Conference on Multimedia Modeling, 2020, pp. 451–462.
[50] D. Vázquez, J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, A.M. López, A. Romero, M. Drozdzal, A. Courville, A benchmark for endoluminal scene segmentation of colonoscopy images, Journal of Healthcare Engineering (2017) 1–9.
[51] C. Yang, X. Guo, M. Zhu, B. Ibragimov, Y. Yuan, Mutual-prototype adaptation for cross-domain polyp segmentation, IEEE Journal of Biomedical and Health Informatics 25 (10) (2021) 3886–3897.
[52] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, L. Shao, Specificity-preserving rgb-d saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 4681–4691.
[53] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, L. Shao, Camouflaged object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 2777–2787.
[54] M.-M. Chen, D.-P. Fan, Structure-measure: A new way to evaluate foreground maps, International Journal of Computer Vision 129 (2021) 2622–2638.
[55] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1597–1604.
[56] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, in: International Joint Conferences on Artificial Intelligence, 2018, pp. 698–704.
[57] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 733–740.
[58] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115 (3) (2015) 211–252.
[59] X. Zhao, L. Zhang, H. Lu, Automatic polyp segmentation via multi-scale subtraction network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2021, pp. 120–130.
[60] Y. Sun, G. Chen, T. Zhou, Y. Zhang, N. Liu, Context-aware cross-level fusion network for camouflaged object detection, in: Proceedings of the International Joint Conference on Artificial Intelligence, 2021, pp. 1025–1031.
[61] T.-Y. Chiu, D. Gurari, Pca-based knowledge distillation towards lightweight and content-style balanced photorealistic style transfer models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7844–7853.

Tao Zhou received the Ph.D. degree in Pattern Recognition and Intelligent Systems from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, in 2016. From 2016 to 2018, he was a Postdoctoral Fellow in the BRIC and IDEA lab, University of North Carolina at Chapel Hill. From 2018 to 2020, he was a Research Scientist at the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi. He is currently a Professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include machine learning, computer vision, and medical image analysis.

Yi Zhou is currently an Associate Professor at Southeast University. He received a M.Sc. degree from the Department of Electronic and Electrical Engineering, University of Sheffield, U.K. in 2014, and a Ph.D. degree from the School of Computing Sciences, University of East Anglia, U.K. in 2018. He was a Research Scientist with IIAI. His research interests include computer vision, pattern recognition, machine learning, and medical imaging.

Kelei He received the Ph.D. degree in computer science and technology from Nanjing University, China. He is currently the assistant dean of National Institute of Healthcare Data Science at Nanjing University. He is also an assistant researcher of Medical School at Nanjing University, China. His research interests include medical image analysis, computer vision and deep learning.

Chen Gong received his B.E. degree from East China University of Science and Technology (ECUST) in 2010, and dual doctoral degree from Shanghai Jiao Tong University (SJTU) and University of Technology Sydney (UTS) in 2016 and 2017, respectively. Currently, he is a full professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests mainly include machine learning, data mining, and learning-based vision problems.

Jian Yang received the PhD degree from Nanjing University of Science and Technology (NUST), on the subject of pattern recognition and intelligence systems in 2002. In 2003, he was a Postdoctoral researcher at the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow at Biometrics Centre of Hong Kong Polytechnic University. From 2006 to 2007, he was a Postdoctoral Fellow at Department of Computer Science of New Jersey Institute of Technology. Now, he is a Chang-Jiang Professor in the School of Computer Science and Technology of NUST. He is the author of more than 200 scientific papers in pattern recognition and computer vision. His papers have been cited more than 6000 times in the Web of Science, and 15000 times in the Scholar Google. His research interests include pattern recognition, computer vision and machine learning.

Huazhu Fu is currently a Senior Scientist at Inception Institute of Artificial Intelligence, UAE. He received a Ph.D. degree from Tianjin University in 2013 and was a Research Fellow at NTU for two years. From 2015 to 2018, he was a Research Scientist with I2R, A*STAR, Singapore. His research interests include computer vision, machine learning, and AI in healthcare. He serves as an Associate Editor for IEEE TMI and IEEE JBHI, and also served as the Area Chair for MICCAI 2021 and Co-Chair for the OMIA Workshop.

Dinggang Shen Professor, IEEE Fellow, AIMBE Fellow, IAPR Fellow. His research interests include medical image analysis, computer vision, and pattern recognition. He has published more than 1000 papers in the international journals and conference proceedings, with h-index of 105. He serves as an editorial board member for eight international journals, and was General Chair for MICCAI 2019.