ICGNet: Integration Context-Based Reverse-Contour Guidance Network for Polyp Segmentation

Xiuquan Du[1,2,*], **Xuebin Xu**[2] and **Kunpeng Ma**[2]
[1]Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei, China
[2]School of Computer Science and Technology, Anhui University, Hefei, China
dxqllp@ahu.edu.cn

Abstract

Precise segmentation of polyps from colonoscopic images is extremely significant for the early diagnosis and treatment of colorectal cancer. However, it is still a challenging task due to: (1) the boundary between the polyp and the background is blurred makes delineation difficult; (2) the various size and shapes causes feature representation of polyps difficult. In this paper, we propose an integration context-based reverse-contour guidance network (ICGNet) to solve these challenges. The ICGNet firstly utilizes a reverse-contour guidance module to aggregate low-level edge information and meanwhile constraint reverse region. Then, the newly designed adaptive context module is used to adaptively extract local-global information of the current layer and complementary information of the previous layer to get larger and denser features. Lastly, an innovative hybrid pyramid pooling fusion module fuses the multi-level features generated from the decoder in the case of considering salient features and less background. Our proposed approach is evaluated on the EndoScene, Kvasir-SEG and CVC-ColonDB datasets with ten evaluation metrics, and gives competitive results compared with other state-of-the-art methods in both learning ability and generalization capability[1].

1 Introduction

Colorectal cancer (CRC) accounts for 9.4% of cancer deaths worldwide, with nearly one million cases in 2020 [Sung *et al.*, 2021]. Polyps are abnormal tissue growth and it is the precursor to colon cancer. If polyp can be removed before colon cancer, it can effectively reduce the incidence and mortality of colon cancer. In clinical practice, colonoscopy is the gold standard, which provides information about the location and appearance of polyps. Whereas, in the process of colonoscopy, there are missing detection and involves expensive human labour. Therefore, an automatic and accurate polyp segmentation method is needed to help doctors locate the polyp area during surgery and provide a reliable prediction for further diagnosis.

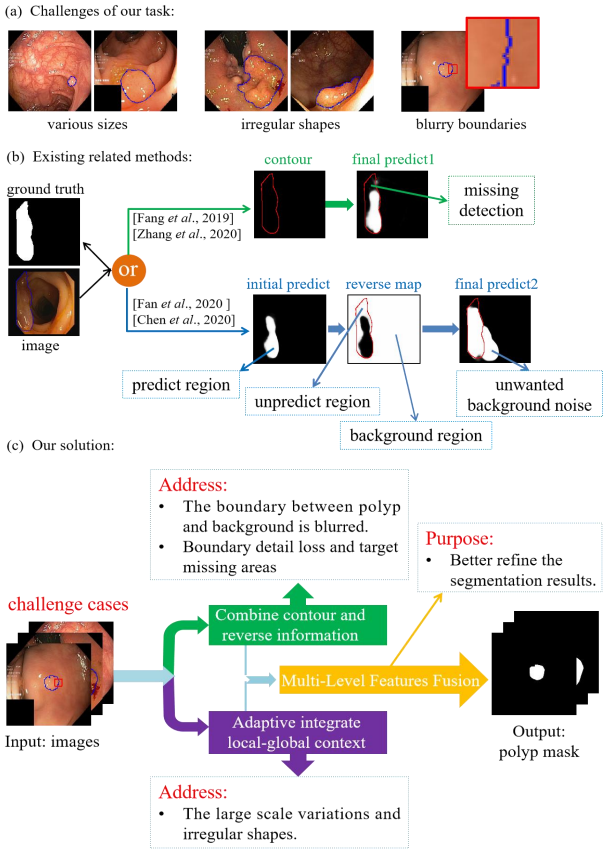

Figure 1: Challenges and method of our framework to handle the polyps segmentation via using the colonoscopic image. From (a) to (c), they are challenges of our tasks, existing related methods and our solution, respectively.

With the development of deep learning in medical image analysis, deep convolutional networks have made good progress in polyp segmentation. However, automatic polyp segmentation is still a challenging task, which is shown in Figure 1(a). (1) Various sizes and irregular shapes of the polyp. (2) Due to the lack of strong contrast, the boundary between polyp and background is blurred, making it difficult

*Corresponding Author
[1]Supplementary material: https:/ICGNet.github.io/Material.pdf

to be recognized. Recently, there have been some efforts to attempt to address these issues, which can be summarized as two solutions, one used reverse information (see the blue path in Figure 1(b)) and the other used edge information (see the green path in Figure 1(b)). For example, PraNet [Fan *et al.*, 2020] generated a global map as the initial guidance region and then used the reverse attention module to reveal more complete objects. ACSNet [Zhang *et al.*, 2020] extracted the global context from the bottom encoder and provided a local context with edge information to deal with the diversity of the shape and size of polyp areas. Nevertheless, they all have some limitations. First, reverse information contains the unpredicted and the background region (see reverse map in Figure 1(b), the red line represents the ground truth contour), the performance of the network will be limited when introducing unwanted background noise (see final predict2 in Figure 1(b)). Second, it is difficult to improve the segmentation accuracy and there is missing detection only by using edge information as shape constraint (see final predict1 in Figure 1(b)). Finally, they fail to make full use of the local-global information of the middle layer to handle polyps that change extremely in shapes and sizes.

In the face of these challenges, the above existing methods only utilized partial information, leading to unsatisfactory results. Therefore, our basic hypothesis is that reverse-contour information guide learning, local-global information adaptive learning and multi-level information fuse learning can improve polyps segmentation performance. Then, we propose an integration context-based reverse-contour guidance network (ICGNet). As shown in Figure 1(c), our approach mainly includes three parts. The green part is for solving the lack of strong contrast in the feature extraction process, through designing a novel reverse-contour guidance module (RCG) to avoid the influence caused by using reverse or contour information alone. Additionally, it compensates for boundary detail loss and considers missed detection in feature sampling. The purple part proposes an adaptive local-global context module (ALGM) to extract larger and denser features and aggerate complementary information so that adapt to polyps of various sizes and irregular shapes. Lastly, we further develop a hybrid pyramid pooling fusion module (HPPF) in the orange part to better refine the segmentation results, which capture global average and local maximum features to fuse the multi-level features generated by the decoder.

The contributions of this work are summarized as follows:

- We propose the reverse-contour guidance module to receive contour detail and reverse information to highlight the boundary needed for polyp segmentation, meanwhile focusing on unpredicted targets within the contour.
- We design the adaptive local-global context module to adaptively extract local and global information and also complementary information, which can help to segment polyps with large scale and various shapes.
- We present a hybrid pyramid pooling fusion module, which extracts global average and local maximum information to guide the fusion of multi-level features to achieve refined segmentation results.

2 Related Work

2.1 Boundary Method and Reverse Attention

To enhance boundary segmentation, [Hatamizadeh *et al.*, 2019] used a network edge branch to consider organ boundary information. SFANet [Fang *et al.*, 2019] applied region-boundary constraints to supervise the learning of polyp. To make up for the missing object part, [Chen *et al.*, 2020; Fan *et al.*, 2020] utilized reverse attention block to learn missing parts and details. However, using only edge information as shape constraint or reverse attention may cause incorrect detection and introduce unwanted noise, respectively. Therefore, [Xu *et al.*, 2021] employed both boundary and reverse attention mechanisms. Unfortunately, it is simply concatenated, without a good focus on the complementarity of each other information. Inspired by [Zhao *et al.*, 2019] focused on the complementarity between edge and object information, we believe that effectively implementing contour features guide reverse features can better optimize the segmentation results as well. Among them, contour information can compensate for detail loss and promote the network more attention to the unpredicted object region within the boundary.

2.2 Multi-Scale Context and Multi-Level Features Fusion

Multi-scale context is beneficial to improve the accuracy of semantic segmentation. DeepLab [Chen *et al.*, 2017] took advantage of the ASPP module to achieve multi-scale features. Although it enhances the network's ability to make use of context information, it lacks sufficient density to handle the complex cases where polyps vary greatly in size and shape. With the help of a dense connection manner [Yang *et al.*, 2018], information can be better mined. However, these efforts have limited the flexibility of the dilation rate and cannot learn long-range dependence. Non-local operation [Wang *et al.*, 2018] and dilation rate of hierarchy change can help solve these issues.

The fusing multi-level features have been proven to be an effective factor in improving segmentation performance [Xie *et al.*, 2020]. RAGCM [Wu *et al.*, 2021] utilized the contextual relationship among channels to fuse the multi-level features. Whereas using global average pooling alone overestimates the target scales and involves too much background because it averages all pixels in feature maps. [Ru *et al.*, 2021] adopted global max pooling to help solve this problem together. It follows that combining global information and maximum response subregion information can eliminate semantic gaps and further guide the fusion of multi-level features.

3 Proposed Method

3.1 Overview

The overall architecture of the proposed ICGNet is shown in Figure 2. It takes polyp images as inputs and outputs the segmentation results. We adopt ResNet34 [He *et al.*, 2016] as our encoder and it contains five blocks in total. To relieve the computation burden, they are compressed to 64 channels using a convolutional layer with 64 1x1 kernels. The

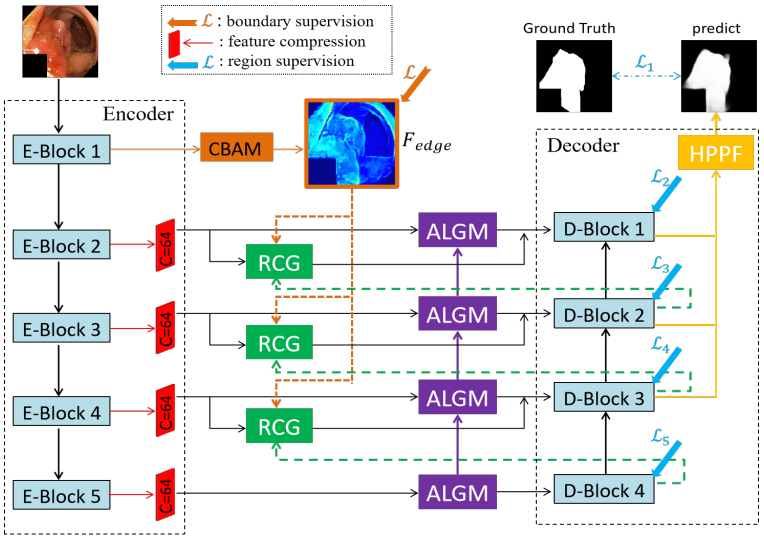

Figure 2: Overview of the proposed ICGNet.The ICGNet consists of encoder and decoder, including three parts with reverse-contour guidance module (RCG), adaptive local-glocal context module (ALGM) and hybrid pyramid pooling fusion module (HPPF).

model integrates three types of modules, Reverse-Contour Guidance Module (RCG), Adaptive Local-Global Context Module (ALGM) and Hybrid Pyramid Pooling Fusion Module (HPPF). The RCG receives boundary information of the first encoder layer and the output from the previous D-Block which is used as the reverse attention mechanism, making the network pay more attention to target edge and false-negative regions. The ALGM uses the output feature of the encoder as the input, and further adaptively extracts local-global context to adopt the variable scales of polyps. Subsequently, the new ALGM combines with complementary information from the previous ALGM to enhance feature representation. Finally, the HPPF fuses the last three decoder blocks to refine the segmentation result.

3.2 Reverse-Contour Guidance Module

As shown in Figure 3, the original input images have blurred boundaries that disguise areas of polyps, making the network difficult to segment them. Also, there may be the problem of missing detection in the process of segmentation (see predict map in Figure 3). In order to solve these issues, the reverse-contour guidance module (RCG) is proposed. RCG includes three inputs with edge feature, reverse feature and E-Block feature. The target edge feature F_{edge} is extracted from the low-layer of the network because the low-level features have rich detailed information (such as boundary, texture and appearance). However, these features are submerged by background noise and are difficult to be used directly. The convolutional block attention module (CBAM) [Woo *et al.*, 2018] is used to help remove background noise, which is shown in Figure 2. Specifically, we embed feature F_{edge} has two advantages. It can compensate for the loss of high-level detailed information as well as let the network focus on the region of unpredicted objects within the boundary. For feature F_{edge}, we apply a 3×3 Conv and interpolate it to the same size as E to generate feature F'_e. These operations are represented as $H(\cdot)$. For reverse feature input is to obtain a more complete target, the predictive map of the previous D-Block to calculate a reverse attention map and then channel-wise multiplied with the E-Block feature $E \in \mathbb{R}^{C\times H\times W}$ (C, H and W are the channel, height and width) get reverse feature map $R \in \mathbb{R}^{C\times H\times W}$. Next, we concatenate the features R and F'_e through a convolution operation. This operation effectively fuses reverse and boundary information according to their contributions. The process can be formulated as follows:

$$f = Conv_{3\times3}(\mathbb{C}[R, H(F_{edge})]) \tag{1}$$

where $H(\cdot)$ denotes convolution operation and the bilinear interpolation. $\mathbb{C}[\cdot]$ refers to concatenation.

After obtaining feature f, we through MLP operations generate a boundary location information weight vector multiplied by feature f to guide it to learn the region within the boundary. The results are added with feature E. The aim is to effectively combine the enhanced edge location information with the reverse semantic context to generate more accurate and complete object features. Although we achieve good results, the algorithm is still flawed. It can be calculated as:

$$Y = E \oplus (f \otimes (MLP(f))) \tag{2}$$

where MLP operations include single-channel 1×1Conv and sigmoid operation.

3.3 Adaptive Local-Global Context Module

Since the size and shape of polyps are continually changing, suffer from a challenge for feature expression in segmentation. Inspired by [Zhang *et al.*, 2020], local and global contextual information can help the network adapt to polyps of different scales and enhance feature representation. Therefore, we design an adaptive local-global context module (ALGM) to extract local and global information from encoder features to model multi-scale receptive fields. The ALGM also helps to integrate complementary information between

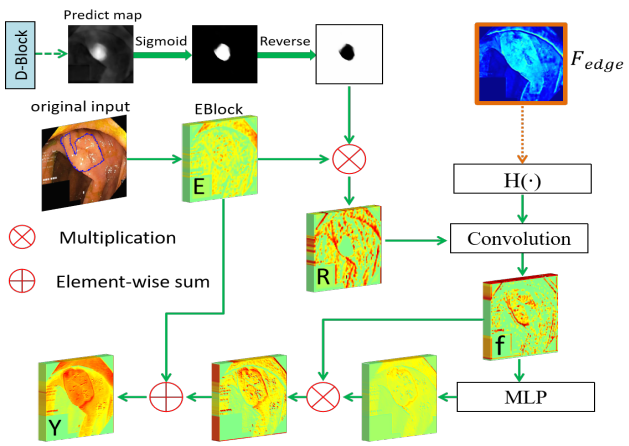

Figure 3: The architecture of the reverse-contour guidance module (RCG).

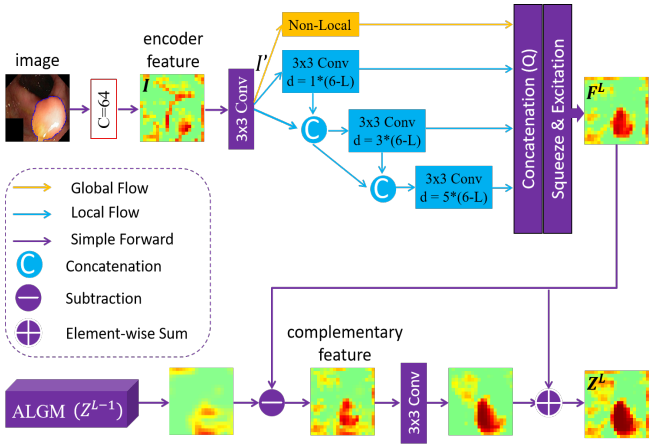

Figure 4: The architecture of adaptive local-global context module (ALGM).

the middle layers and the top layers, while reducing the redundant information generated by direct addition.

As shown in Figure 4. First, the encoder feature ($I \in \mathbb{R}^{C\times H\times W}$) serves as the input of ALGM and adjusts the channel dimension to a quarter ($I' \in \mathbb{R}^{(C/4)\times H\times W}$) through 3×3 convolution. Then, the module is divided into two parts, the first part is the global module G_{global}, which implements a non-local operation [Wang *et al.*, 2018] to capture the long-range spatial dependencies of the feature, and it can accurately extract the global dependencies of each locational feature. It can be calculated as:

$$G_{global} = F_{non}(I') \tag{3}$$

where F_{non} represents non-local operation, the feature I' generates global information G_{global}.

Specifically, the second part is the local module L_{local}, and it aims to obtain larger and denser receptive fields to achieve predominant multiscale representation. The input feature I' through a series of cascaded dilated convolutions with increasing dilation rates from $1 * (6 - L)$, $3 * (6 - L)$ to $5 * (6 - L)$ (for the L^{th} layer). The input to each dilated convolution layer is formed by concatenating the input feature with the outputs from previous convolutions. Last, the outputs Z^L from the global and local context are concatenated together and combine complementary information with the previous ALGM to obtain rich multiscale difference information and a more precise feature representation. Here, we use Squeeze-and-Excitation Block [Hu *et al.*, 2018] adaptively capture channel-wise dependencies and important feature correlations for feature Q to achieve better fusion. The calculation of L_{local} and Z^L as follows:

$$L_{local} = \mathbb{C}_{dl=1}^{dl=3}\Big(\sum_{D_{dl}^L=1*(6-L)}^{5*(6-L)} D_{dl}^L[H_{dl-1}^L, H_{dl-2}^L, \cdots, H_0^L]\Big) \tag{4}$$

$$F^L = SE(\mathbb{Q}(G_{global}, L_{local})) \tag{5}$$

$$Z^L = F^L + \Phi(F^L, Z^{L-1}) \tag{6}$$

where D_{dl}^L represents the dilation rate of the $(dl)^{th}$ dilated convolution at the L^{th} layer of the network. H_{dl}^L is dilated convolution. $\sum$ denotes the change of the dilation rates. $\mathbb{C}$ refers to concatenation. The $SE(\mathbb{Q}(\cdot))$ operation through Squeeze-and-Excitation block after concatenating both global and local information and $\Phi(\cdot)$ extracts complementary information between L^{th} and $(L-1)^{th}$ of the ALGM.

3.4 Hybrid Pyramid Pooling Fusion Module

A multi-level feature fusion strategy was applied and verified to availably improve segmentation performance [Wu *et al.*, 2021]. The feature aggregation layer has a great influence on the quality of the segmentation result, so we design a hybrid pyramid pooling fusion module (HPPF) to achieve more effective output feature fusion in our decoder.

The HPPF is illustrated in Figure 5. Given input feature maps $D_i \in \mathbb{R}^{C_i\times H_i\times W_i}$ from the decoder, we upsample the features to the same dimension, where C, H and W are the channel, height and width, $I \in \{1, 2, 3\}$. Then, the three features map is concatenated altogether to extract a large feature map $F \in \mathbb{R}^{C\times H\times W}$. Next, we use a global average pooling (GAP) to obtain global information and get the completeness of targets. Meanwhile, to solve the background problem brought by GAP, we also introduce global max pooling (GMP). GMP adopts 4×4 and 8×8 pooling to involve local maximum pixels and preserves more differentiated objects as well as less background area. And then through 1×1 Conv, ReLU operations to obtain three pooling feature maps $F_1 \in \mathbb{R}^{C\times 1\times 1}$, $F_2 \in \mathbb{R}^{C/16\times 4\times 4}$, $F_3 \in \mathbb{R}^{C/64\times 8\times 8}$. Finally, we flatten F_2 and F_3 into the output feature vector $Z \in \mathbb{R}^{C\times 1\times 1}$ is calculated by weighting three pooling feature maps.

$$Z = \frac{1}{\alpha+2}\Big(\sum_{k\in(4,8)} F_{GMP}^k + \alpha\frac{1}{hw}\sum_{i=1}^{h}\sum_{j=1}^{w} F_{GAP}^{i,j}\Big) \tag{7}$$

$$P = Conv_{3\times 3}(F \otimes \sigma(f(Z))) \tag{8}$$

where α is a weight factor that controls the ratio of the feature map of GAP, we set 1. F_{GMP}^k is a global max pooling feature

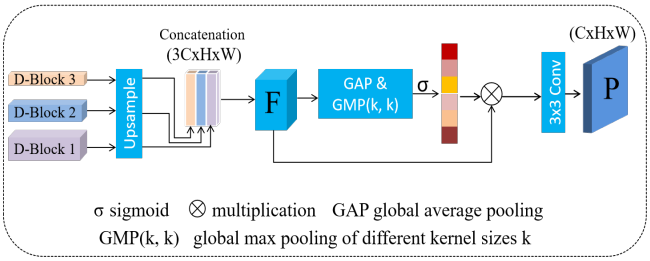

Figure 5: The architecture of hybrid pyramid pooling fusion module (HPPF).

with different kernel sizes k and $F_{GAP}^{i,j}$ is a global average pooling feature. $f(\cdot)$ denotes the transformation with two 1D convolutions. $\sigma(\cdot)$ is a Sigmoid function to produce channel attention weighting map. After applying an element-wise multiplication to reweight the F, we can get a more refined dense prediction P based on a 3×3 convolution.

4 Experiments

4.1 Experiment Setup

We evaluate our proposed methods on three benchmark polyp datasets: EndoScene [Vázquez *et al.*, 2017], Kvasir-SEG [Jha *et al.*, 2020], CVC-ColonDB [Bernal *et al.*, 2012]. EndoScene contains 912 images from 44 colonoscopy sequences which are acquired from 36 patients total. Kvasir-SEG has 1000 images with polyp regions, manually annotated by an experienced doctor released in 2020. CVC-ColonDB consists of 380 annotated frames extracted from 15 different colonoscopy sequences.

Following [Zhang *et al.*, 2020; Fan *et al.*, 2020], we adopt ten widely-used metrics to evaluate the model performance, including "Recall (Rec)", "Specificity (Spec)", "Precision (Prec)", "Dice Coefficient (Dice)", "Intersection-over-Union for Polyp (IoUp)", "IoU for Background (IoUb)", "Mean IoU (mIoU)" , "Accuracy (Acc)", "S-measure" and "E-measure". More details of hyperparameters and loss function are reported in the implementation details and loss function respectively of supplementary material.

4.2 Ablation Study

For the sake of evaluating the effectiveness of the three key components in our proposed method, we perform the ablation study on the Kvasir-SEG dataset. The results are presented in Table 1. The ICGNet without three modules is used as the baseline, and we continually add ALGM, RCG and HPPF to it, expressed as Baseline+ALGM, Baseline+RCG, Baseline+ALGM+RCG and Ours, respectively. When the introduction of aforementioned modules by degree, boosting Dice by 1.3%, 1.41%, 2.34% and 2.92%, also adding parameters by 0.56M, 0.24M, 1.02M, 1.15M, respectively. These experiments demonstrate that the proposed three modules are lightweight and can notably improve the segmentation performance with very few parameters.

To give a more intuitive result, we visualize the heat map of outputs in Figure S1 and the network structure for each experiment in Figure S2. Concrete analysis can be seen in supplementary material.

Method	Dice	mIoU	S_α	E_ϕ^{max}	Param.(M)
Baseline	89.43	89.29	90.40	93.68	21.60
+ALGM	90.73	90.59	91.50	95.09	22.16
+RCG	90.84	90.89	91.58	95.59	21.84
+ALGM+RCG	91.77	91.24	92.32	95.50	22.62
Ours	92.35	91.99	93.15	96.24	22.75

Table 1: Ablation experiment on the Kvasir-SEG dataset.

4.3 Comparison with State-of-the-art

To validate the effectiveness of our method, we implement learning ability and generalization capability between our method and several advanced methods, including U-Net++ [Zhou *et al.*, 2018], U-Net [Ronneberger *et al.*, 2015], ResUNet [Zhang *et al.*, 2018], SFANet [Fang *et al.*, 2019], PraNet [Fan *et al.*, 2020], ACSNet [Zhang *et al.*, 2020] and CCBANet [Nguyen *et al.*, 2021]. In our experiments, we conducted all competitors on the same train, validation and test sets and computing environments, meanwhile using their released code with default settings to guarantee a fair comparison.

Learning Ability

In this section, we use the Kvasir-SEG and EndoScene datasets to validate our model's learning ability. We refer to the setting in [Zhang *et al.*, 2020] to divide the training set, validation set and test set. We resize the images to 384×288 on the EndoScene dataset and fixed size of 320×320 on the Kvasir-SEG dataset. As shown in Table 2, our proposed approach achieves competitive results on all two datasets. On the EndoScene dataset, our model achieves 87.93% Dice and 89.56% mean IoU, meanwhile achieving 92.35% Dice and 91.99% mean IoU on the Kvasir-SEG dataset. It demonstrates that our network is lightweight and has a strong learning ability to effectively segment polyps. Additionally, we can see clearly in Figure S3 that the parameters of our model are smaller than advanced models while achieving compa-

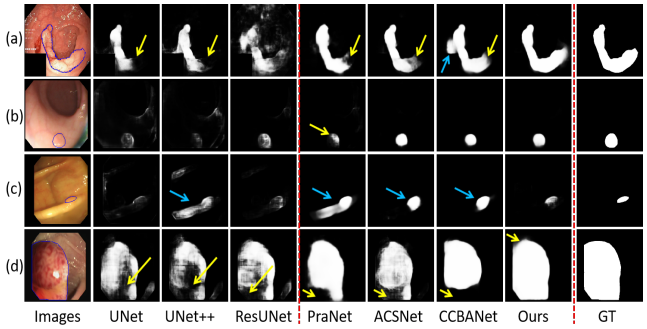

Figure 6: Qualitative results of different methods, where the white represents the polyps segmentation mask and the black represents the background. The yellow arrow indicates the missing segmentation location, and the light blue arrow indicates the location where the segmentation result does not fit well with the ground truth.

Model	EndoScene										Kvasir-SEG									
	Rec	Spec	Prec	Dice	IoUp	IoUb	mIoU	ACC	S_α	E_ϕ^{max}	Rec	Spec	Prec	Dice	IoUp	IoUb	mIoU	ACC	S_α	E_ϕ^{max}
UNet [Ronneberger *et al.*, 2015]	85.54	98.75	83.56	80.31	70.68	95.90	83.29	96.25	–	–	87.89	97.69	83.89	82.85	73.95	94.73	84.34	95.65	–	–
U-Net++ [Zhou *et al.*, 2018]	78.90	99.15	86.17	77.38	68.00	95.48	81.74	95.78	–	–	88.67	97.49	83.17	82.80	73.74	94.49	84.11	95.42	–	–
ResUNet [Zhang *et al.*, 2018]	82.18	98.64	85.38	79.52	70.98	95.62	83.30	96.37	–	–	81.25	98.31	87.88	81.14	72.23	94.00	83.11	94.90	–	–
SFANet [Fang *et al.*, 2019]	85.51	98.94	86.81	82.93	75.00	96.33	85.66	96.61	–	–	91.99	97.05	82.95	84.68	77.06	94.83	85.94	95.71	–	–
PraNet [Fan *et al.*, 2020]	82.94	99.03	90.52	83.34	75.85	96.56	86.20	96.80	90.39	92.91	91.41	97.94	89.56	90.75	84.50	96.59	90.54	97.25	91.79	95.51
ACSNet [Zhang *et al.*, 2020]	87.96	99.16	90.99	86.59	79.73	96.86	88.29	97.11	90.45	94.07	93.14	**98.55**	91.59	91.30	85.80	97.00	91.40	97.64	92.47	95.57
CCBANet [Nguyen *et al.*, 2021]	85.56	99.08	91.14	85.53	79.89	97.78	88.83	97.01	91.21	93.83	93.56	97.79	92.14	91.99	86.44	96.86	91.65	97.36	92.36	95.55
ICGNet(Ours)	**88.45**	**99.52**	**91.24**	**87.93**	**80.96**	**98.17**	**89.56**	**98.34**	**92.42**	**95.04**	**93.70**	98.31	**92.63**	**92.35**	**86.89**	**97.09**	**91.99**	**97.68**	**93.15**	**96.24**

Table 2: Quantitative results of the test datasets EndoScene and Kvasir-SEG. The best results are highlighted in **bold**.

rable performance on the Dice metric in the supplementary material.

Figure 6 illustrates some visual segmentation results of our model and the compared models in some complex situations. It is observed that **1)** the UNet, UNet++ and ResUNet will fail to segment them well when polyps have highly blurry boundaries (see Figure 6 (b)-(c)) and various scales (see Figure 6 (a) and (d)). These methods neither aggregate large receptive-field features nor consider the boundary constraints, so it is difficult to fit complex situations. **2)** The PraNet, ACSNet and CCBANet aggregated multi-scale features significantly improve the large scales segmentation performance, and the latter two models used boundary constraints to get better results in lack of strong contrast cases (see Figure 6 (b)-(c)). But, they still have the problems of missing segmentation (see yellow arrow in Figure 6) and the segmentation result does not fit well with the ground truth (see light blue arrow in Figure 6). **3)** Our model can accurately locate and segment the polyp areas by using associative learning with three modules.

Generalization Capability

In order to adapt unseen different types of polyps in clinical scenarios, it also need to improve the predictive power of the model for unknown data. In this section, we conduct one experiment to test the model's generalizability. EndoScene is a combination of CVC612 and CVC300. We use Kvasir-SEG and EndoScene-CVC612 to serve as the seen dataset with 90% as training set and the remaining 10% as a validation set, while CVC-ColonDB serves as the test set. All the images are set to 352x352.

To prove the generalization capability of ICGNet, we present six main metrics results in Table 3. As can be seen,

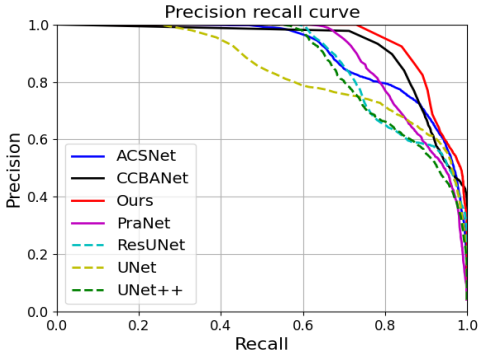

Figure 7: The PR curves of different methods.

Method	Spec	Dice	mIoU	ACC	S_α	E_ϕ^{max}
UNet [Ronneberger *et al.*, 2015]	98.52	54.77	70.00	94.44	72.14	80.23
U-Net++ [Zhou *et al.*, 2018]	98.61	58.99	71.56	94.93	75.26	82.13
ResUNet [Zhang *et al.*, 2018]	98.44	45.39	64.09	93.59	65.88	76.17
PraNet [Fan *et al.*, 2020]	98.97	74.28	82.34	**96.69**	84.56	88.48
ACSNet [Zhang *et al.*, 2020]	98.85	73.63	82.05	96.31	85.00	89.43
CCBANet [Nguyen *et al.*, 2021]	99.02	70.68	79.77	95.82	83.68	88.31
ICGNet(Ours)	**99.15**	**74.97**	**82.64**	96.56	**85.92**	**91.40**

Table 3: Quantitative results of the test datasets CVC-ColonDB. The best results are highlighted in **bold**.

our methods have the highest values on Dice (74.97%) and mIoU (82.64%) metrics and indicate that our prediction mask and ground truth have a high degree of coincidence. In the meantime, we also present the precision-recall curves in Figure 7, where the solid red line of the proposed method contains the largest area obviously, and its performance is better than other methods. These results show that ICGNet achieves a good generalization capability. It is mainly due to the following advantages. **1)** Larger and denser features are extracted through local-global information that can adaptive learning various sizes and shapes of unseen polyps. **2)** Performing reverse-contour layer on the middle feature map, further provide boundary localization information and encourage the network to learn more missing features. **3)** The multi-level feature fusion branch supplement the information lost during the delivery process, making the final predictions more reliable. Specific qualitative result in Figure S4 and analysis is reported in the supplementary material. To further improve the reliability of generalization capability, we also conduct two experiments on the polyp segmentation datasets. The quantitative results are shown in Table S1 of supplementary material.

4.4 Conclusion

In this paper, ICGNet is presented for automatic polyp segmentation from colonoscopic images. The reverse-contour guidance module integrates low-level edge information and reverses information to solve low contrast boundary and missing detection. Adaptive local-global context module combines the local and global information, which provides the capability to enable larger and denser receptive fields to help identify diverse scales polyp. And hybrid pyramid pooling fusion module focus on prominent areas and fewer background to fine segmentation results. Extensive experiments and ablation studies demonstrate the superiority of the proposed method and achieve new state-of-the-art performance.

Acknowledgments

This work was supported in part by the Provincial Natural Science Research Program of Higher Education Institutions of Anhui province under Grant KJ2020A0035. And then, the authors acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

References

[Bernal *et al.*, 2012] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[Chen *et al.*, 2020] Shuhan Chen, Xiuli Tan, Ben Wang, Huchuan Lu, Xuelong Hu, and Yun Fu. Reverse attention-based residual network for salient object detection. *IEEE Transactions on Image Processing*, 29:3763–3776, 2020.

[Fan *et al.*, 2020] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. *MICCAI*, pages 263–273, 2020.

[Fang *et al.*, 2019] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *MICCAI*, pages 302–310, 2019.

[Hatamizadeh *et al.*, 2019] Ali Hatamizadeh, Demetri Terzopoulos, and Andriy Myronenko. End-to-end boundary aware networks for medical image segmentation. In *MLMI*, pages 187–194, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[Jha *et al.*, 2020] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MMM*, pages 451–462, 2020.

[Nguyen *et al.*, 2021] Tan-Cong Nguyen, Tien-Phat Nguyen, Gia-Han Diep, Anh-Huy Tran-Dinh, Tam V Nguyen, and Minh-Triet Tran. Ccbanet: Cascading context and balancing attention for polyp segmentation. In *MICCAI*, pages 633–643, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[Ru *et al.*, 2021] Lixiang Ru, Bo Du, and Chen Wu. Learning visual words for weakly-supervised semantic segmentation. In *IJCAI*, pages 982–988, 2021.

[Sung *et al.*, 2021] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[Vázquez *et al.*, 2017] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017:1–9, 2017.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.

[Wu *et al.*, 2021] Huisi Wu, Jiasheng Liu, Wei Wang, Zhenkun Wen, and Jing Qin. Region-aware global context modeling for automatic nerve segmentation from ultrasound images. In *AAAI*, pages 2907–2915, 2021.

[Xie *et al.*, 2020] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. Mlcvnet: Multi-level context votenet for 3d object detection. In *CVPR*, pages 10447–10456, 2020.

[Xu *et al.*, 2021] Xiuqi Xu, Mingyu Zhu, Jinhao Yu, Shuhan Chen, Xuelong Hu, and Yuequan Yang. Boundary guidance network for camouflage object detection. *Image and Vision Computing*, 114:104283, 2021.

[Yang *et al.*, 2018] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, pages 3684–3692, 2018.

[Zhang *et al.*, 2018] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018.

[Zhang *et al.*, 2020] Ruifei Zhang, Guanbin Li, Zhen Li, Shuguang Cui, Dahong Qian, and Yizhou Yu. Adaptive context selection for polyp segmentation. In *MICCAI*, pages 253–262, 2020.

[Zhao *et al.*, 2019] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. *ICCV*, pages 8779–8788, 2019.

[Zhou *et al.*, 2018] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *DLMIA*, pages 3–11, 2018.