2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995022

ICBNet: Iterative Context-Boundary Feedback Network for Polyp Segmentation

1st Yefan Xiao
College of Intelligence and Computing.
Tianjin University
Tianjin, China
fnxyf@tju.edu.cn

2nd Zhihao Chen
College of Intelligence and Computing.
Tianjin University
Tianjin, China
zh_chen@tju.edu.cn

3rd Liang Wan*
College of Intelligence and Computing.
Tianjin University
Tianjin, China
lwan@tju.edu.cn

4th Lequan Yu
Department of Statistics and Actuarial Science.
The Hong Kong University
Hong Kong SAR, China
lqyu@hku.hk

5th Lei Zhu
The Hong Kong University of Science and Technology(Guangzhou).
Guangdong, China
The Hong Kong University of Science and Technology.
Hong Kong SAR, China
leizhu@ust.hk

***Abstract*—Accurate polyp segmentation from colonoscopy images, which is critical to automatic colorectal cancer diagnosis, attracts increasing attentions in recent years. Most existing deep learning-based methods adopt the one-stage processing pipeline, by usually fusing features from different levels or employing boundary-related attention. In this paper, we propose an novel Iterative Context-Boundary feedback Network, namely ICBNet, for robust and accurate polyp segmentation. By mimicking the "from-Preliminary-to-Refined" working paradigm of doctors, ICBNet adopts an iterative feedback learning strategy. Differently from other feedback methods which only use the prediction mask as a guide for foreground features, ICBNet refines encoder features with contextual and boundary-aware details from the preliminary segmentation and boundary predictions, and conducts such strategy in an iterative manner to achieve progressive improvement. Moreover, a dual-branch iterative feedback unit (IFU) is developed to enhance features under the guidance of segmentation and boundary predictions to enable the iterative learning. Extensive experiments on five widely-used polyp segmentation datasets demonstrate that the proposed ICBNet can utilize progressive refinement to effectively address the challenges of large appearance variations and obscure boundaries, and hence achieves more accurate and robust results against the state-of-the-arts methods.**

***Index Terms*—Polyp segmentation, Colonoscopy, Iterative feedback learning, Contextual and boundary-aware.**

I. Introduction

Colorectal cancer (CRC) ranks the third most commonly diagnosed cancer, yet has become the second leading cause (9.4%) of cancer-related deaths around the world in 2020 [18]. Most CRC cases develop from tiny protrusions raised on the surface of the colon or rectum, known as polyps [7]. For CRC screening and prevention, endoscopists usually used a standard Colonoscopy [11] technique to visually examine and identify polyps. The situation that polyp diagnose highly depends on the experiences and skills of endoscopists, puts high demands on automatic accurate segmentation of polyps from colonoscopy images.

Automatic polyp segmentation is a challenging task since the detail image information appearances such as color and texture are easily confused, and polyps may have low boundary contrast against surroundings. The early learning-based methods rely on the extraction of hand-crafted features [12] [19], such as color, texture, shape, appearance, or a combination of these features. In these methods, trained classifiers are usually used to distinguish a polyp from its surroundings. However, these models often obtain high miss-detection rate result.for the reason that the representation capability of hand-crafted features is quite limited when it comes to dealing with the high intra-class variations of polyps and low inter-class variations between polyps and hard mimics [26].

In recent years, many deep learning-based methods have been developed for polyp segmentation, reporting promising progress. For instance, fully convolutional neural networks are employed with a pre-trained model to identify and segment polyps [2]. Inspired by the success of U-Net [14] applied in biomedical image segmentation, U-Net++ [29] and ResUNet++ [9] are developed for polyp segmentation and obtain good performance. In order to further improve the segmentation accuracy, more recent methods use the characteristics of polyp images to improve the segmentation results. As polyps arise from the gut in the form of hyperplastic tissue so that the uncertain information of the boundary area is very rich, more methods exploit the boundary features of polyp images to improve segmentation results. PraNet [6] aggregates encoder features and reverse boundary attention at multiple levels to mine the boundary cues. HRENet [15], LOD-Net [4], extract the boundary area or border candidate from the segmentation prediction as explicit supervision. These experiment results show that the boundary information can improve the final segmentation results both visually and in accuracy. Later on, some methods explored the complementary information among dif-

Liang Wan is the corresponding author of this work.

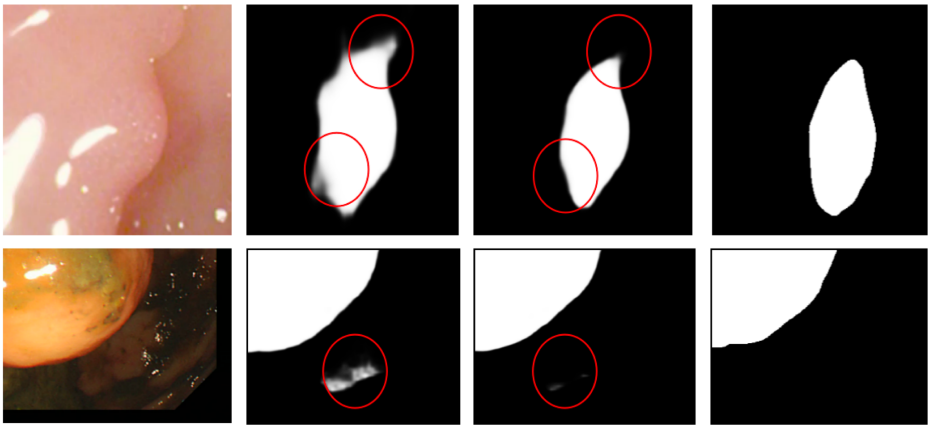


Fig. 1: Iterative refinement of our proposed ICBNet: (a) polyp images; (b) one-stage segmentation prediction (without feedback refinement); (c) multi-stage segmentation prediction (with context-boundary feedback refinement at the second iteration); (d) ground truth. The red circle highlights the refined areas.

ferent resolutions, including MSNet [28] which proposes a multi-scale subtraction network, and Polyp-PVT [5] which is built on a transformer backbone and develops cross-level fusion. Also, SANet [25] pays more attention to compensate color differences among colonoscopy images. Although these methods improve polyp segmentation accuracy from different aspects, they may still fail in case of existing a high degree of similarity in the polyp(foreground) and background. The polyp feature discrimination between the polyp and background regions is insufficient, so the learned weights of these networks are not enough to capture inter-class differences exist in polyp image. In practice, failure recognition or blurred boundary still exist in their segmentation results.

To remedy these problem and achieve accurate polyp segmentation, we use the last prediction masks to optimize semantic segmentation of polyps in iterative manner. The prediction masks, i.e. the preliminary polyp segmentation results, naturally include regions with high foreground confidence and regions with high background confidence, namely quasi-foreground and quasi-background areas. Thus, we can use the divided regions to act on complex semantic features, and divide them into quasi-foreground features and quasi-background features, respectively. In this way, We can extract information from quasi-foreground features and quasi-background features separately to obtain intra-class differences of polyps. Besides, prediction masks also provide prior information and helps learn sample variability. In recent years, there are some methods use prediction masks as an iteratively feedback strategy, such as [21] [13] [16] [23]. Although they are not fully applicable in polyp segmentation since they only use prediction masks for foreground guidance, these methods proved that prediction masks help us pruning of the predicted masks during inference and allowing the network to learn local and global features, which can be output from the learned weight rectified mask.

Specifically, we propose an noval **I**terative **C**ontext-**B**oundary feedback **Net**work, termed ICBNet, for robust and accurate polyp segmentation. By mimicking the "from-Preliminary-to-Refined" working paradigm of doctors, ICBNet adopts an iterative feedback learning strategy which generates a preliminary estimation at the first glance and further updates the predication iteratively. In ICBNet, semantic features from the encoder will be used for boundary generation and polyp region segmentation, respectively. Besides, we proposed an iterative feedback unit (IFU) which uses proposed iterative feedback prediction mask to considers the contextual and boundary-aware information. The dual-branch structure of the IFU can analyze the information of the quasi-foreground region and the quasi-background region respectively, which are combined with generated boundary information. Semantic features optimized by IFU will generate optimized segmentation results through the decoder, which will be used in the next iteration of the segmentation process. Through this iterative strategy, the network effectively handles appearance variations and obscure boundaries. Experimental results on five benchmark polyp segmentation datasets demonstrate that our network outperforms the state-of-the-art methods.

II. Method

Figure 1 shows two cases for the proposed ICBNet. As we can see from Figure 1(b), the first-stage segmentation predication generates blurred boundaries near the polyp area, and may also result in false estimations. The second-stage predication in Figure 1(c) sharpens the boundary and suppresses false estimations effectively, yielding more accurate and robust segmentation results.

Our main contributions can be summarized as follows:

- We propose a noval feedback model, namely ICBNet, to obtain preliminary-to-refined segmentations in an iterative way via feedbacking contextual and boundary-aware

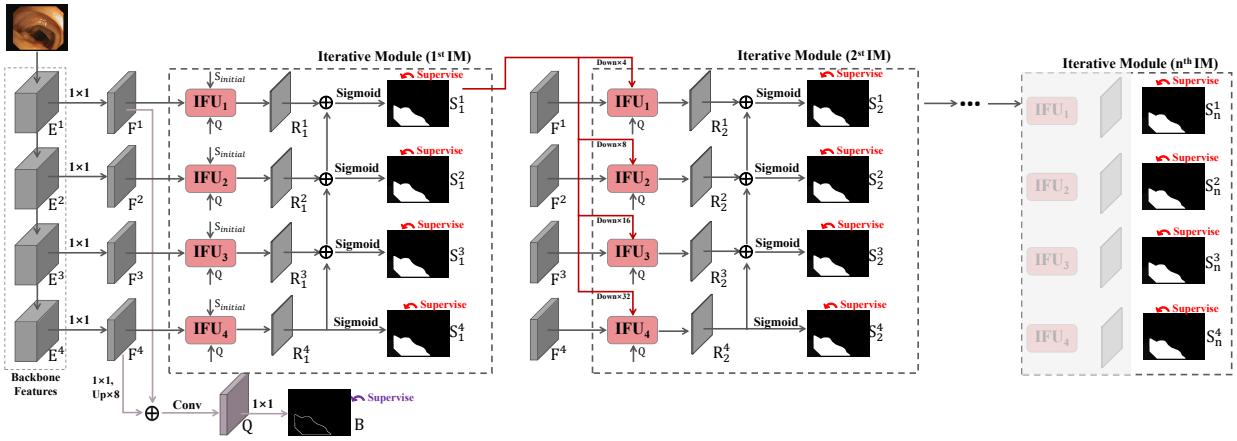


Fig. 2: An overview of the proposed ICBNet.

detail from the preliminary segmentation and boundary predictions.

- We design a dual-branch Iterative Feedback Unit (IFU) to better strengthen important features in foreground/background and mix boundary features.
- Extensive experiments show that the proposed ICBNet outperforms the state-of-the-arts methods on five widely used public benchmarks with more accurate and robust results.

Fig. 2 illustrates the architecture of the proposed ICBNet, which estimates the boundary and segmentation maps at multiple levels, and utilizes an iterative module (IM) to feedback contextual/boundary-aware information.

A. Iterative Feedback Learning of ICBNet

In the proposed network ICBNet, we obtain semantic features at different resolutions through the encoder backbone. These semantic features will be used for iteratively optimizing semantic segmentation results. The concatenation of high-level features and low-level features will also be used to generate boundary features. N iterative modules (IMs) is set up in ICBNet to predict and optimize the segmentation results. The semantic features, boundary features and the previous segmentation results will be feeded to each IM which be connected each other through the adjacent previous prediction results. Finally, through n iterations of optimization, the final segmentation result is obtained.

For more details, ICBNet use a transformer-based method, the PVT [24], as the encoder backbone to extract more powerful and robust features following [5] for the reason that PVT uses global receptive field to extract features. Given an input polyp image with size of $[H, W]$, the encoder backbone extracts multi-level features $\{E^i, i = 1, \ldots, 4\}$ with resolution $[\frac{H}{2^{i+1}}, \frac{W}{2^{i+1}}]$. To decrease the parameter number for subsequent processing, we compress the channel dimension of E^i to 32, yielding F^i, by applying 1×1 convolutions. Since the boundary represents the global shape or salient topology of structures, it is highly related to the areas with large differences in the global feature and the fine-detail feature. Hence, we leverage the lowest-level and highest-level features together for offering boundary feature $Q \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 32}$ under the supervision of the ground-truth boundary follow [3], which is computed as

$$Q = \mathcal{F}(Concat(F^1, UP_{\times 8}(F^4))), \tag{1}$$

where $\mathcal{F}(\cdot)$ is a convolutional unit containing one 1×1 convolution and two 3×3 convolutions, $Concat(\cdot)$ denotes the channel concatenation, and $UP_{\times 8}(\cdot)$ means a $8\times$ upsampling.

In the iterative module (IM), an iterative feedback unit (IFU_i) is applied at each layer, which takes the last segmentation predication and boundary feature as feature-refinement guidance in the skip connection. Then four decoding layers are used for combining multi-scale semantic features to obtain segmentation prediction results. Specifically, the Fig. 2 shows ICBNet architecture details.

At the first iteration, we begin with an initial segmentation predication $\text{S}_{initial}$, which is set as a whole-white/whole-1 image with size $[H, W]$. This allows us to retain the complete semantic information of the encoder during the first iteration, so as to generate better preliminary segmentation results. By feeding F^i, Q and $S_{initial}$ to the IFUs, we obtain refined compressed features $R_1^i \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times 1}$ which are used for estimating prediction results S_1^i at different levels.

At the t-th iteration, we take the last full-resolution prediction result S_{t-1}^1 and reuse F^i/Q as the inputs to the IFUs, to iteratively refine the features R_t^i and get those refined binary predictions S_t^i. During the decoding, the refined features R_t^i are accumulated across different levels via summation. The computation of S_t^i can be formulated as follows,

$$\begin{aligned} R_t^i &= IFU(F^i, Q, S_{t-1}^i), \\ S_t^4 &= \sigma(R_t^4), \\ S_t^i &= \sigma(R_t^i + R_t^{i+1}), i = 1, 2, 3, \\ S_t^0 &= S_{initial}, \end{aligned} \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid function. It is worth mentioning that for a given layer, the IFUs at different iterations use the same parameters. During the inference stage, we follow

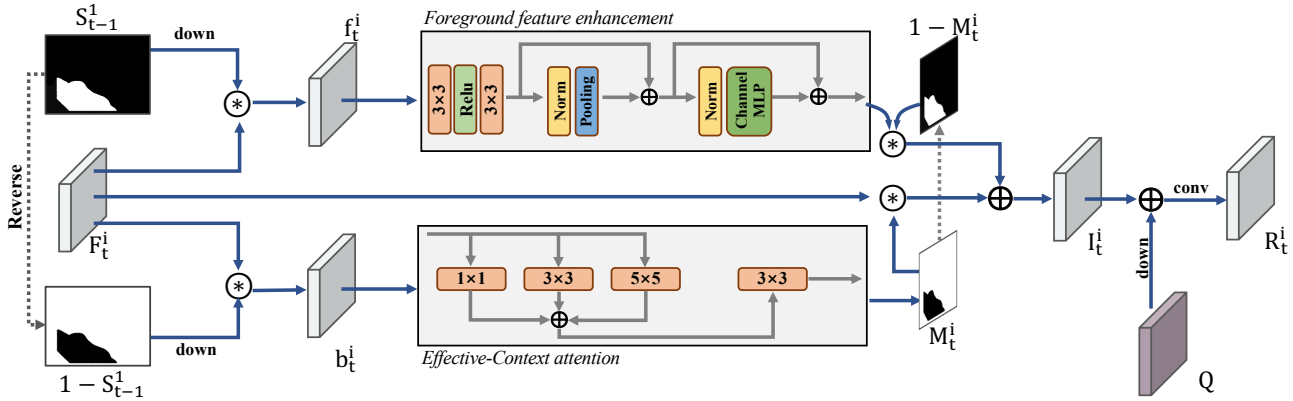


Fig. 3: Illustration of the proposed IFU module.

the above iterative learning procedure, and finally outputs S_T^1 as the polyp segmentation prediction.

B. Iterative Feedback Unit (IFU)

Taking into account the difference between the front and background contextual text, we develop a dual-branch Iterative Feedback Unit (IFU) which processes the foreground and background features separately to better strengthen important features under the guidance of segmentation prediction S_{t-1}^1 and boundary features Q.

As shown in Fig. 3, the IFU consists of three main parts: foreground feature enhancement (FFE) branch, effective-context attention (ECA) branch and feature mixture. With IFU, semantic feature will be refined under the guidance of prediction mask, and then be forwarded to decoders for generating better segmentation results. To be specific, we first make use of the segmentation mask S_{t-1}^1 to initially divide the encoder feature F_t^i into the foreground feature f_t^i and the background feature b_t^i, which are given by

$$\begin{aligned} f_t^i &= F_t^i \odot Down_{\times 2^{i-1}}(S_{t-1}^1), \\ b_t^i &= F_t^i - f_t^i, \end{aligned} \tag{3}$$

where operator $\odot$ denotes element-wise multiplication, and $Down_{\times 2^{(i-1)}}(\cdot)$ means a $2^{i-1}\times$ downsampling.

Foreground Feature Enhancement (FFE). The extracted foreground feature f_t^i reflects the appearance distribution of preliminary estimated polyp regions. Therefore, we proposed Foreground Feature Enhancement (FFE) block which aims to strengthens existing attentive foreground features that exist in both predicted masks and semantic feature. In order to enhance more global semantic information while maintaining local details in f_t^i, we apply two 3×3 convolution layers and one poolformer layer in the FFE block. The poolformer layer [27] is an effective MLP-based layer which replaces complex and computationally expensive self-attention with a simple pooling layer in Transformer structure without affecting performance.

Effective-Context Attention (ECA). The ECA block attempts to estimate a refined attention mask by enhancing background feature b_t^i, i.e. $M_t^i = ECA(b_t^i)$, which is used to make up for residual attentive features that exist in the learned predicted masks but not or weakly exist in et‘. In ECA block, we apply three convolutions with varying kernel sizes in parallel, followed by summation and one 3×3 convolution, which captures contextual information from background in different proportions. The resulted feature undergoes one 1×1 convolution to yield the refined mask M_t^i.

Feature Mixture. The reverse of the refined mask, which is (1-M_t^i), includes both enhanced foreground areas and uncertain areas, which can provide additional supplementary information for encoder features. Hence, the improved feature I_t^i can be computed as the following combination,

$$I_t^i = (1 - M_t^i) \odot FFE(f_t^i) + M_t^i \odot F_t^i. \tag{4}$$

In the end, we further mix the improved feature I_t^i with boundary feature Q, obtaining the final refined feature R_t^i, which is computed as

$$R_t^i = \mathcal{F}(I_t^i + Down_{2^{i-1}}(Q)). \tag{5}$$

C. Loss Function & Implementation Details

The network uses the combination of binary cross entropy (BCE) loss and IOU loss for both the poly region and boundary supervision. Let $L(\cdot) = L_{BCE} + L_{IOU}$ denote the loss function. The total loss L_t at the t-th iteration consists of $L(S_t^i)$ and $L(B)$, where B is the boundary prediction. As for the boundary ground truth, we infer it by eroding the mask and subtract the eroded result from the original mask. Specially, We choose a convolution kernel of size 5*5, 5 iteration times for image eroding. And we calculate the difference between the original ground truth image and the eroded image, which as the ground truth of the final boundary area. Then the whole loss L_{final} of the network is the sum of losses from T iterations, given by

$$L_{final} = \sum_{t=1}^{T} L_t = \sum_{t=1}^{T} (\sum_{i=1}^{4} L(S_t^i)) + L(B). \tag{6}$$

TABLE I: Performance comparison with different polyp segmentation models on Kvasir-seg and CVC-ClinicDB. The highest scores are bolded.

	Methods	References	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
Kvasir-seg	U-Net	(MICCAI'15) [14]	0.818	0.746	0.794	0.858	0.893	0.055
	U-Net++	(DLMIA'18) [29]	0.821	0.743	0.808	0.862	0.910	0.048
	PraNet	(MICCAI'20) [6]	0.898	0.840	0.885	0.915	0.948	0.030
	SANet	(MICCAI'21) [25]	0.904	0.847	0.892	0.915	0.949	0.028
	MSNet	(MICCAI'21) [28]	0.907	0.862	0.893	0.922	0.944	0.028
	ERRNet	(PR'22) [10]	0.901	0.840	0.891	0.911	0.952	0.027
	TGANet	(MICCAI'22) [20]	0.886	0.822	0.883	0.898	0.942	0.032
	FANet	(TMI'22) [21]	0.852	0.791	0.837	0.872	0.902	0.059
	Polyp-PVT	(TMI'22) [5]	0.917	0.864	0.911	0.925	0.956	0.023
	ICBNet(ours)	-	**0.929**	**0.883**	**0.924**	**0.937**	**0.962**	**0.021**
CVC-ClinicDB	U-Net	(MICCAI'15) [14]	0.823	0.755	0.811	0.889	0.954	0.019
	U-Net++	(DLMIA'18) [29]	0.794	0.729	0.785	0.873	0.931	0.022
	PraNet	(MICCAI'20) [6]	0.899	0.849	0.896	0.936	0.979	0.009
	SANet	(MICCAI'21) [25]	0.916	0.859	0.909	0.940	0.972	0.012
	MSNet	(MICCAI'21) [28]	0.921	0.879	0.914	0.941	0.972	0.008
	ERRNet	(PR'22) [10]	0.918	0.868	0.917	0.940	0.986	0.009
	TGANet	(MICCAI'22) [20]	0.863	0.805	0.863	0.903	0.952	0.015
	FANet	(TMI'22) [21]	0.823	0.756	0.807	0.870	0.902	0.045
	Polyp-PVT	(TMI'22) [5]	0.937	0.889	0.936	0.949	**0.985**	**0.006**
	ICBNet(ours)	-	**0.938**	**0.892**	**0.937**	**0.955**	0.983	**0.006**

We use Pytorch to implement our ICBNet. All input images are uniformly resized to 352×352. For data augmentation, we adopt multi-scale training which is the same as PraNet [6]. The whole network is trained in an end-to-end way with an Adam optimizer. Initial learning rate and batch size are empirically set to 5e-5 and 16, respectively. The number of training epochs are 150 on a single RTX 3090 GPU. Our network has a parameter size of 98.3M, and an average inference-speed of 17fps for five experiment datasets.

III. Experiment

A. Datasets and Evaluation Metrics

We evaluate the performance of our ICBNet on five benchmark datasets, including Kvasir-seg [8], CVC-ClinicDB [22], CVC-ColonDB [19], ETIS [17] and CVC-300 [1]. Following the recent work PraNet [6], we utilize the same training set containing 900 samples from the Kvasir and 550 samples from the CVC-ClinicDB for testing different segmentation methods on the five benchmark datasets.

We employ six metrics for conducting quantitative comparisons, i.e., mDice, mIoU, F_β^w, S_α, E_ϕ^{max}, and MAE; please refer to PraNet [6] for the definitions of six metrics. In general, a better polyp segmentation method shall have a smaller MAE score, and a larger score at the other five metrics.

B. Comparisons with State-of-the-art methods

We compare our network against seven state-of-the-art methods, which are U-Net [14], U-Net++ [29], PraNet [6], SANet [25], MSNet [28], ERRNet [10], TGANet [20], FANet [21], and Polyp-PVT [5]. To provide fair comparisons, we utilize their public polyp segmentation results, following the same experimental settings.

Quantitative Comparisons. Table I shows quantitative results of six metrics (mDice, mIoU, F_β^w, S_α, E_ϕ^{max}, and MAE) on Kvasir-seg and CVC-ClinicDB datasets. Apparently, Polyp-PVT achieves the best performance of six metrics among all compared methods on Kvasir-seg. More importantly, our ICBNet further outperforms Polyp-PVT on all six metrics for Kvasir-seg. In comparison, our ICBNet has an mDice improvement of 1.31%, an mIoU improvement of 2.20%, and an MAE improvement of 8.70% on for Kvasir-seg. For CVC-ClinicDB, Polyp-PVT takes the 2nd rank on mDice, mIou, F_β^w, and S_α, and they are 0.937, 0.889, 0.936, 0.949. On the contrary, our method takes the 1st rank on four metrics, and they are 0.938, 0.892, 0.937, 0.955. Regarding E_ϕ^{max}, Polyp-PVT and our method have the largest score (0.985) and the second largest score (0.983), while they have the same MAE score (0.006) on CVC-ClinicDB.

Table II summarizes metrics of different segmentation methods on the other three unseen benchmark datasets (i.e., CVC-CloneDB, ETIS and CVC-300). From these quantitative results, we have the observations below:

(1) Our method has an mDice improvement of 1.91% and an mIoU improvement of 3.12% on for ETIS. Although our MAE score (0.015) ranks the third, it is very close to the 1st rank (0.013).

TABLE II: Performance comparison with different polyp segmentation models on CVC-CloneDB, ETIS and CVC-300. The highest scores are bolded.

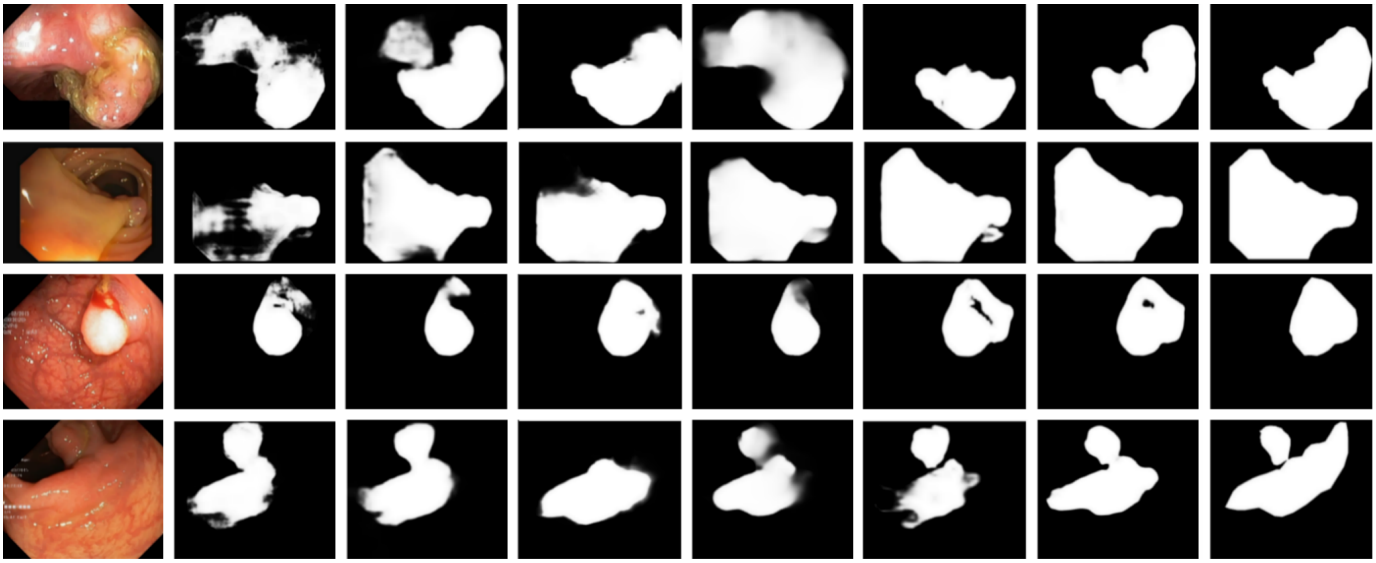
	Methods	References	mDice ↑	mIoU ↑	F_β^w ↑	S_α ↑	E_ϕ^{max} ↑	MAE ↓
CVC-CloneDB	U-Net	(MICCAI'15) [14]	0.512	0.444	0.498	0.712	0.776	0.061
	U-Net++	(DLMIA'18) [29]	0.483	0.410	0.467	0.691	0.760	0.064
	PraNet	(MICCAI'20) [6]	0.709	0.640	0.696	0.819	0.869	0.045
	SANet	(MICCAI'21) [25]	0.752	0.669	0.725	0.837	0.867	0.043
	MSNet	(MICCAI'21) [28]	0.755	0.678	0.737	0.836	0.883	0.041
	ERRNet	(PR'22) [10]	–	–	–	–	–	–
	TGANet	(MICCAI'22) [20]	0.695	0.609	0.674	0.789	0.836	0.053
	FANet	(TMI'22) [21]	0.558	0.486	0.527	0.683	0.697	0.153
	Polyp-PVT	(TMI'22) [5]	0.808	0.727	0.795	0.865	**0.913**	0.031
	ICBNet(ours)	-	**0.812**	**0.738**	**0.801**	**0.866**	0.912	**0.030**
ETIS	U-Net	(MICCAI'15) [14]	0.398	0.335	0.366	0.684	0.740	0.036
	U-Net++	(DLMIA'18) [29]	0.401	0.344	0.390	0.683	0.776	0.035
	PraNet	(MICCAI'20) [6]	0.628	0.567	0.600	0.794	0.841	0.031
	SANet	(MICCAI'21) [25]	0.750	0.654	0.685	0.849	0.881	0.015
	MSNet	(MICCAI'21) [28]	0.719	0.664	0.678	0.840	0.830	0.020
	ERRNet	(PR'22) [10]	0.691	0.611	0.658	0.822	0.886	0.014
	TGANet	(MICCAI'22) [20]	0.574	0.488	0.542	0.744	0.805	0.028
	FANet	(TMI'22) [21]	0.415	0.361	0.388	0.622	0.577	0.161
	Polyp-PVT	(TMI'22) [5]	0.787	0.706	0.750	0.871	0.906	**0.013**
	ICBNet(ours)	-	**0.802**	**0.728**	**0.774**	**0.884**	**0.914**	0.015
CVC-300	U-Net	(MICCAI'15) [14]	0.710	0.627	0.684	0.843	0.876	0.022
	U-Net++	(DLMIA'18) [29]	0.707	0.624	0.687	0.839	0.898	0.018
	PraNet	(MICCAI'20) [6]	0.871	0.797	0.843	0.925	0.972	0.010
	SANet	(MICCAI'21) [25]	0.888	0.815	0.859	0.928	0.962	0.008
	MSNet	(MICCAI'21) [28]	0.869	0.807	0.849	0.925	0.943	0.010
	ERRNet	(PR'22) [10]	0.889	0.813	0.864	0.931	**0.978**	**0.006**
	TGANet	(MICCAI'22) [20]	0.822	0.733	0.794	0.879	0.948	0.011
	FANet	(TMI'22) [21]	0.668	0.600	0.641	0.781	0.816	0.082
	Polyp-PVT	(TMI'22) [5]	**0.900**	**0.833**	**0.884**	**0.935**	0.973	0.007
	ICBNet(ours)	-	0.898	**0.833**	0.881	0.934	0.966	0.008

(2) Regarding CVC-CloneDB, our method consistently has the best performance on mDice, mIoU, F_β^w, and S_α, and MAE, and they are 0.812, 0.738, 0.801, 0.866 and 0.030. Except the E_ϕ^{max} metric (0.912) ranks second, which only 0.1% worse than the first (0.913).
(3) Our method has the second best performance on mDice (0.898), slightly worse than the best one (0.900), and the best mIoU (0.833) for CVC-300. On other evaluation metrics, our method also has a thin gap with the best number.

Visual Comparisons. From the visual results shown in Figure 4, we can find that compared segmentation networks tend to neglect parts of polyp regions or wrongly identify non-polyp regions as the target ones. As these methods insufficiently consider boundary information and lack of guidance of coarse segmentation, our results are more far away from the ground truths. On the contrary, our method can more accurately segment the polyp regions from different input colonoscopy images than all compared methods, and our results are more consistent with the ground truths; our segmentation results basically have no segmentation artifacts, which represents the uncertain area is very small; see Figures 4(g) and (g), The last row shows a challenging case, in which our result still misses right corner part.

C. Ablation Study

Effectiveness of IFU at iterative modules (IMs). We construct a network (denoted as "basic") by removing all IFU modules from our ICBNet. Then, we construct a network (denoted as "basic+boundary") by adding the IFU modules with only the boundary information into "basic", and another network (denoted as "basic+contextual") by adding the IFU modules with only the contextual information. Table III reports the mDice and mIoU scores of our ICBNet and three construc-



(a) Input Image (b) U-Net (c) PraNet (d) SANet (e) MSNet (f) Polyp-PVT (g) ICBNet (ours) (h) Ground Truth

Fig. 4: Visual results produced by our ICBNet and state-of-the-art methods.

TABLE III: Ablation study on IFUs with boundary or context on Kvasir-seg and CVC-ClinicDB.

Methods	Kvasir-seg		
	mDice ↑	mIoU ↑	MAE ↓
basic	0.902	0.851	0.030
basic+boundary	0.923	0.869	0.024
basic+context	0.925	0.873	**0.021**
ICBNet	**0.929**	**0.883**	**0.021**
Methods	CVC-ClinicDB		
	mDice ↑	mIoU ↑	MAE ↓
basic	0.910	0.861	0.008
basic+boundary	0.916	0.866	0.014
basic+context	0.936	0.889	0.006
ICBNet	**0.938**	**0.892**	**0.006**

tive baselines on Kvasir-seg and CVC-ClinicDB. Apparently, both "basic+boundary" and "basic+contextual" have larger mDice and mIoU scores than "basic", demonstrating that the boundary information or the contextual information in our IFU modules has a capability to enhance features at CNN layers for boosting polyp segmentation. More importantly, combining boundary information and the contextual information together in our IFU modules can further improve polyp segmentation accuracy, as indicated by the superior mDice and mIoU performance of our ICBNet over "basic+boundary" and "basic+contextual" in Table III.

The number of iterative modules (IMs) in our method. Table IV shows the segmentation results of our method with different number of iterative modules (IM). We can find that using two iterations gets the best mDice, mIoU and MAE values. Hence, in all our experiments, we apply two iterations.

TABLE IV: Ablation study on the number of IMs on Kvasir-seg.

IM number	Proformance Result		
	mDice ↑	mIoU ↑	MAE ↓
1	0.926	0.878	0.021
2 (ours)	0.929	0.883	0.021
3	0.928	0.882	0.021
4	0.927	0.881	0.021

IV. Conclusion

This paper presents an iterative context-boundary feedback network, ICBNet, for boosting the Polyp Segmentation. Our key idea is to iteratively feedback the preliminary predications of both segmentation and boundary into different encoder levels, so as to obtain progressively refined segmentation results. For feature enhancement, the contextual and boundary-aware information is strengthened by the iterative feedback unit (IFU). Through this iterative strategy, the network effectively handles appearance variations and obscure boundaries. Experimental results on five benchmark polyp segmentation datasets demonstrate that our network outperforms the state-of-the-art methods.

V. Acknowledgments

This work was supported by the grant from Tianjin Natural Science Foundation (Grant No. 20JCYBJC00960) and National Natural Science Foundation of China (Grant No. 61902275).

REFERENCES

[1] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[2] Patrick Brandao, Evangelos Mazomenos, Gastone Ciuti, Renato Caliò, Federico Bianchi, Arianna Menciassi, Paolo Dario, Anastasios Koulaouzidis, Alberto Arezzo, and Danail Stoyanov. Fully convolutional neural networks for polyp segmentation in colonoscopy. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, pages 101–107. SPIE, 2017.

[3] Zhihao Chen, Lei Zhu, Liang Wan, Song Wang, Wei Feng, and Pheng-Ann Heng. A multi-task mean teacher for semi-supervised shadow detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 5611–5620, 2020.

[4] Mengjun Cheng, Zishang Kong, Guoli Song, Yonghong Tian, Yongsheng Liang, and Jie Chen. Learnable oriented-derivative network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 720–730. Springer, 2021.

[5] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[6] 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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 263–273. Springer, 2020.

[7] Fatima A Haggar and Robin P Boushey. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in Colon and Rectal Surgery*, 22(04):191–197, 2009.

[8] 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 *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

[9] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE, 2019.

[10] Ge-Peng Ji, Lei Zhu, Mingchen Zhuge, and Keren Fu. Fast camouflaged object detection via edge-based reversible re-calibration network. *Pattern Recognition*, 123:108414, 2022.

[11] Xiao Jia, Xiaohan Xing, Yixuan Yuan, Lei Xing, and Max Q-H Meng. Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition. *Proceedings of the IEEE*, 108(1):178–197, 2019.

[12] Alexander V Mamonov, Isabel N Figueiredo, Pedro N Figueiredo, and Yen-Hsi Richard Tsai. Automated polyp detection in colon capsule endoscopy. *IEEE transactions on medical imaging*, 33(7):1488–1502, 2014.

[13] Agata Mosinska, Pablo Marquez-Neila, Mateusz Koziński, and Pascal Fua. Beyond the pixel-wise loss for topology-aware delineation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2018.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[15] Yutian Shen, Xiao Jia, and Max Q-H Meng. Hrenet: A hard region enhancement network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–568. Springer, 2021.

[16] Eisuke Shibuya and Kazuhiro Hotta. Feedback u-net for cell image segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*, pages 974–975, 2020.

[17] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand 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):283–293, 2014.

[18] 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.

[19] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35(2):630–644, 2015.

[20] Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. Tganet: Text-guided attention for improved polyp segmentation. *arXiv preprint arXiv:2205.04280*, 2022.

[21] Nikhil Kumar Tomar, Debesh Jha, Michael A Riegler, Håvard D Johansen, Dag Johansen, Jens Rittscher, Pål Halvorsen, and Sharib Ali. Fanet: A feedback attention network for improved biomedical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[22] 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, 2017.

[23] Feigege Wang, Yue Gu, Wenxi Liu, Yuanlong Yu, Shengfeng He, and Jia Pan. Context-aware spatio-recurrent curvilinear structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12648–12657, 2019.

[24] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[25] Jun Wei, Yiwen Hu, Ruimao Zhang, Zhen Li, S Kevin Zhou, and Shuguang Cui. Shallow attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 699–708. Springer, 2021.

[26] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng Ann Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE journal of biomedical and health informatics*, 21(1):65–75, 2016.

[27] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10819–10829, 2022.

[28] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multi-scale subtraction network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 120–130. Springer, 2021.

[29] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.