Technical Appendix of Supplementary Material

1. Implementation Details

We use Pytorch to implement our model. The whole network is trained in an end-to-endway, and use Adam optimizer with weight decay of 1e-5 to optimize the model. We set the batch size of 4 and the learning rate of 1e-4 for 150 epochs. During training, we utilize data augmentation strategies to enlarge the training set, including random horizontal andvertical flips, rotation, zoom and shift. All experiments are run on a hardware environmentusing 16G NVIDIA GPU.

2. Loss Function

We employ a hybrid loss including segmentation loss $\mathcal{L}_{seg}$ and edge loss. Segmentation loss is the combination of a Dice loss $\mathcal{L}_{dice}$ and a binary cross-entropy(BCE) loss $\mathcal{L}_{bce}$. The $\mathcal{L}_{dice}$ can better evaluate the similarities between ground truth and prediction while $\mathcal{L}_{bce}$ promotes the model to predict the right result for each pixel category. The edge loss is proposed to refine the segmentation boundary, and we adopt focal loss $\mathcal{L}_{focal}$ as its loss function. Also, we refer to[1] to obtain the boundary mask as the ground truth of the boundary prediction. The decoding part of the network uses a deep supervision mechanism and total loss is shown in Eq S1 and Eq S2.

$$\mathcal{L}_{seg} = \lambda_1 \mathcal{L}_{dice}(P, G) + \lambda_2 \mathcal{L}_{bce}(P, G) \qquad \text{(S1)}$$

$$Loss = \sum_{i=1}^{5} \mathcal{L}_{seg}^{i} + \lambda_3 \mathcal{L}_{focal}(P, G) \qquad \text{(S2)}$$

where P, G are the prediction and ground truth, respectively. $\lambda_1, \lambda_2, \lambda_3$ are the weight used to balance cross-entropy, Dice loss and Focal loss, we set to 1.

3. Ablation Study

To give a more intuitive result, we visualize the heat map of outputs as shown in Figure S1. Where several challenging cases with various scales and blurry boundaries can be seen. Intuitively, the results of the Baseline are less than satisfactory in these cases. With the continuous addition of three modules, the results are gradually optimized and very close to the ground truth. It further shows that our proposed modules can adapt to challenging tasks and enhance segmentation completeness. Finally, we present the framework for each experiment, which is shown in Figure S2.

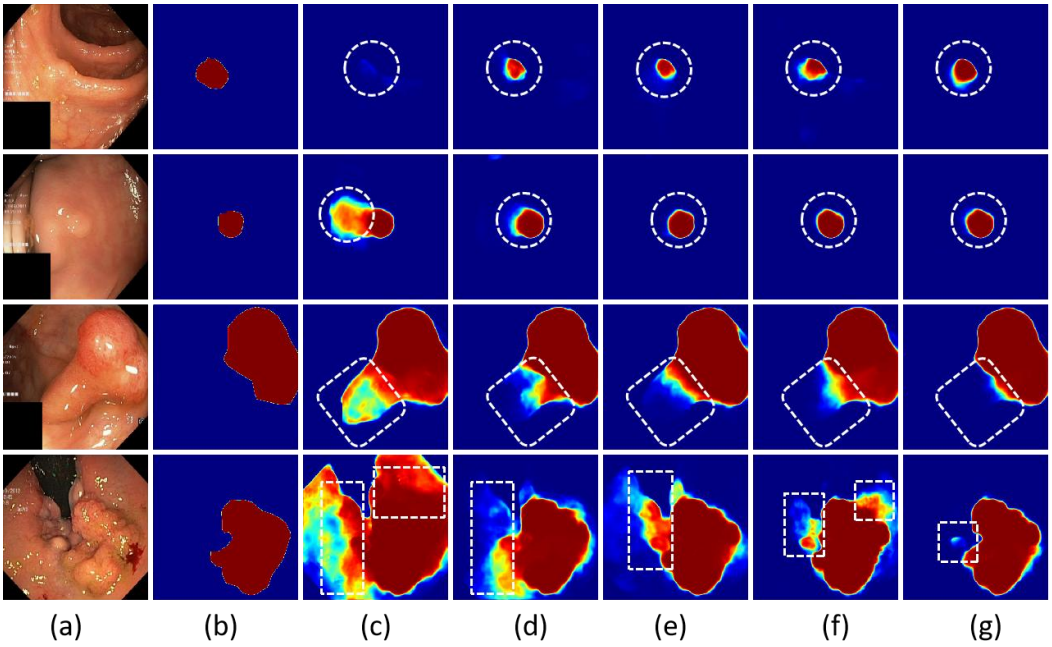
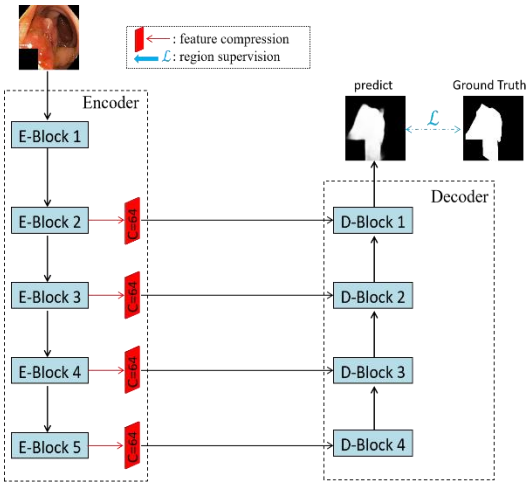
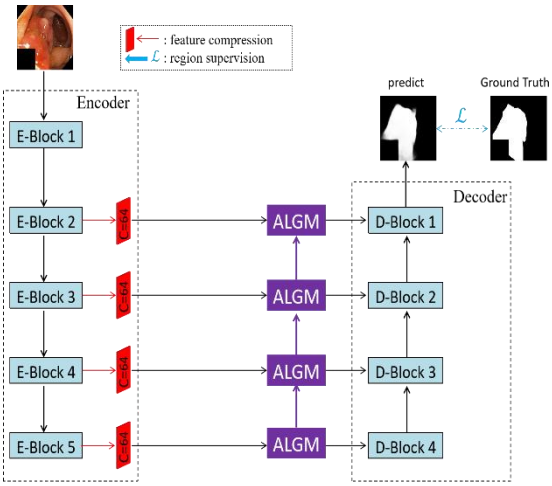


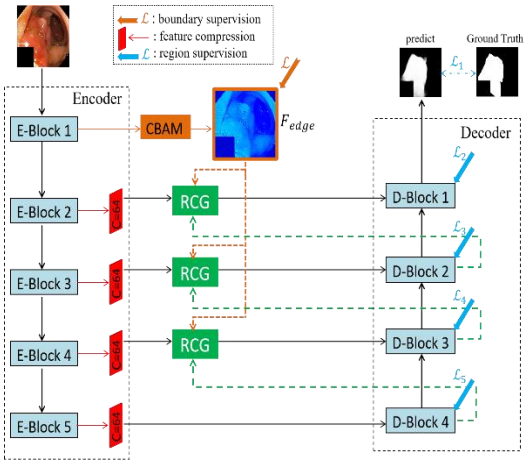
Figure S1: Visualization of the ablation study result, which are converted from the output into a heat map. (a) Images. (b) Ground truth. (c) Baseline. (d) Baseline+ALGM. (e) Baseline+RCG. (f) Baseline+ALGM+RCG. (g)Ours (Baseline+ALGM+RCG+HPPF). The white circles and rectangles inside the range are gradually optimized objects.



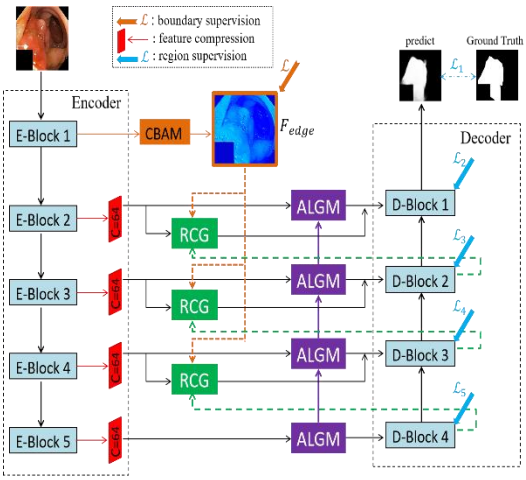
(a) Illustration of the Baseline.



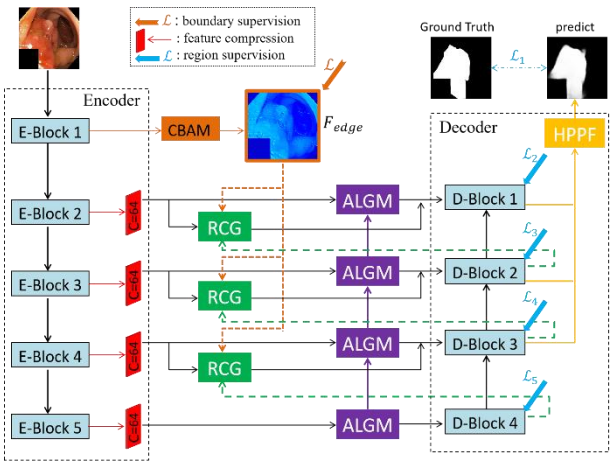
(b) Illustration of the Baseline+ALGM.



(c) Illustration of the Baseline+RCG.



(d) Illustration of the Baseline+ALGM+RCG.



(e) Illustration of the Ours (Baseline+ALGM+RCG+HPPF).

Figure S2: Overview of our proposed network of ablation study.

4. Comparison on Dice metric and model size

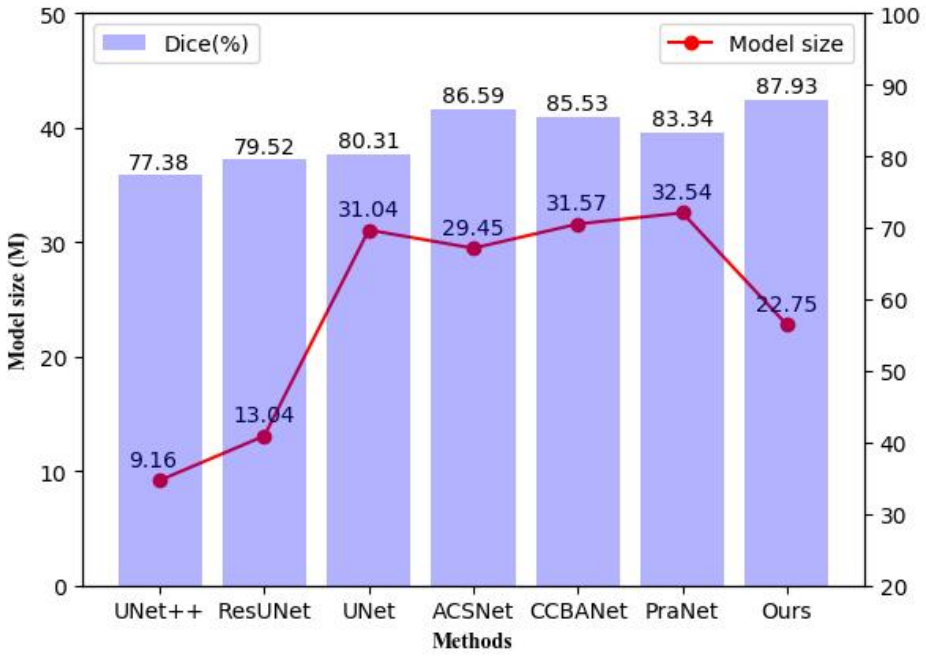


Figure S3: Statistical comparison on Dice metric and model size with different methods on the EndoScene dataset. As we can see that our model has few parameters(22.75M) and achieved the best Dice score(87.93%).

5. Qualitative analysis of generalization capability

In Figure S4, we provide the polyp segmentation qualitative results of different methods (the red curve represents the outline of ground truth). In these cases, the target area is dim and the background is bright, which was not learned before. As shown in Figure S4 (a)-(b), the existing methods tend to favour bright background areas and miss target areas. We can see that the inner part of the red curve has less predicted region, while the outer part of the curve with more prediction happens to be the bright background region. Specifically, the target brightness is improved(see Figure S4 (c)) and the predicted area is relatively complete, but a large amount of unexpected background noise (mainly from bright background) is also introduced. In contrast, our model eliminates the interference of the background and can effectively focus on the target to achieve accurate location, and obtain competitive results.

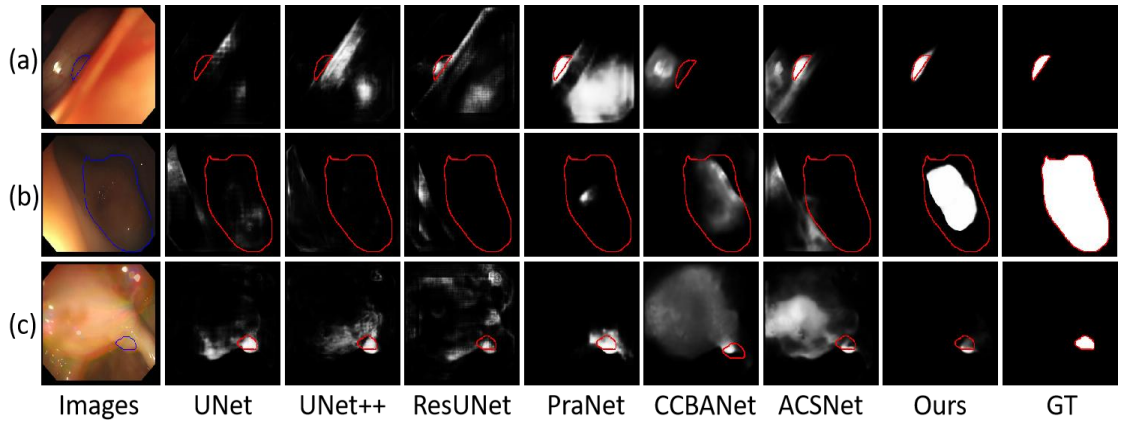


Figure S4: Qualitative results of different methods on the CVC-ColonDB dataset. The red curve represents the outline of ground truth, with the interior as the target area and the exterior as the background area.

6. Quantitative results of generalization capability

	Method	Rec	Spec	Prec	Acc	Dice	mIoU	S_α	E_ϕ^{max}
ETIS [2]	UNet	53.10	93.05	47.67	91.66	34.52	59.85	61.81	65.05
	UNet++	54.76	93.24	42.98	92.09	35.14	60.08	63.90	61.91
	ResUNet	46.49	92.85	30.64	90.80	27.80	55.30	56.25	66.88
	PraNet	**75.61**	97.00	69.78	96.62	62.98	77.12	81.43	85.75
	ACSNet	74.57	**97.11**	69.10	**96.72**	62.50	76.02	79.14	83.58
	CCBANet	69.49	96.71	73.80	96.08	62.58	76.37	78.20	84.02
	Ours	74.69	96.46	**75.24**	96.40	**63.71**	**77.95**	**83.18**	**87.70**
CVC300 [3]	UNet	70.60	98.10	73.52	96.58	65.15	75.52	77.28	84.85
	UNet++	73.81	98.34	75.70	96.98	69.01	77.72	79.50	86.65
	ResUNet	55.62	98.26	63.68	95.98	53.68	68.60	70.52	81.96
	PraNet	86.80	98.98	85.34	98.56	86.31	88.56	91.23	95.05
	ACSNet	87.50	98.85	86.91	98.29	84.19	87.57	90.05	93.24
	CCBANet	**87.86**	99.08	85.74	98.50	86.50	88.35	91.08	94.54
	Ours	87.29	**99.38**	**88.62**	**99.16**	**88.42**	**90.56**	**93.64**	**97.64**

Table S1: Quantitative results of the test datasets ETIS and CVC300.

References

[1]Guo L, Lei B, Chen W, et al. Dual attention enhancement feature fusion network for segmentation and quantitative analysis of paediatric echocardiography[J]. Medical Image Analysis, 2021, 71: 102042.

[2]Silva J, Histace A, Romain O, et al. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer[J]. International journal of computer assisted radiology and surgery, 2014, 9(2): 283-293

[3] Bernal J, Sánchez J, Vilarino F. Towards automatic polyp detection with a polyp appearance model[J]. Pattern Recognition, 2012, 45(9): 3166-3182.