Revisiting Feature Propagation and Aggregation in Polyp Segmentation

Yanzhou Su[1], Yiqing Shen[3], Jin Ye[2], Junjun He[2], and Jian Cheng[1(✉)]

[1] School of Information and Communication Engineering,
University of Electronic Science and Technology of China, Chengdu 611731, China
chenjian@uestc.edu.cn
[2] Shanghai AI Laboratory, Shanghai 200232, China
[3] Johns Hopkins University, Baltimore, MD 21218, USA

Abstract. Accurate segmentation of polyps is a crucial step in the efficient diagnosis of colorectal cancer during screening procedures. The prevalent UNet-like encoder-decoder frameworks are commonly employed, due to their capability of capturing multi-scale contextual information efficiently. However, two major limitations hinder the network from achieving effective feature propagation and aggregation. Firstly, the skip connection only transmits a single scale feature to the decoder, which can result in limited feature representation. Secondly, the features are transmitted without any information filter, which is inefficient for performing feature fusion at the decoder. To address these limitations, we propose a novel feature enhancement network that leverages feature propagation enhancement and feature aggregation enhancement modules for more efficient feature fusion and multi-scale feature propagation. Specifically, the feature propagation enhancement module transmits all encoder-extracted feature maps from the encoder to the decoder, while the feature aggregation enhancement module performs feature fusion with gate mechanisms, allowing for more effective information filtering. The multi-scale feature aggregation module provides rich multi-scale semantic information to the decoder, further enhancing the network's performance. Extensive evaluations on five datasets demonstrate the effectiveness of our method, particularly on challenging datasets such as CVC-ColonDB and ETIS, where it can outperform the previous state-of-the-art models by a significant margin (3%) in terms of mIoU and mDice.

Keywords: Polyp Segmentation · Feature Propagation · Feature Aggregation

1 Introduction

Colorectal cancer is a life-threatening disease that results in the loss of millions of lives each year. In order to improve survival rates, it is essential to identify colorectal polyps early. Hence, regular bowel screenings are recommended, where

endoscopy is the gold standard. However, the accuracy of endoscopic screening can heavily rely on the individual skill and expertise of the domain experts involved, which are prone to incorrect diagnoses and missed cases. To reduce the workload on physicians and enhance diagnostic accuracy, computer vision technologies, such as deep neural networks, are involved to assist in the pre-segmentation of endoscopic images.
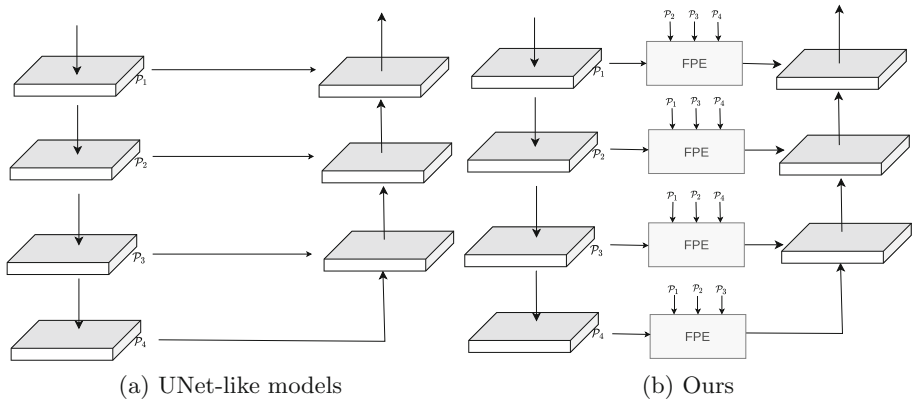
Fig. 1. Comparison of feature propagation methods. The UNet-like model uses skip connections that transmit only single-stage features. In contrast, our approach utilizes FPE to propagate features from all stages, incorporating a gate mechanism to regulate the flow of valuable information.

Deep learning-based image segmentation methods have gained popularity in recent years, dominated by UNet [11] in the field of medical image segmentation. UNet's success has led to the development of several other methods that use a similar encoder-decoder architecture to tackle polyp segmentation, including ResUNet++ [7], PraNet [3], CaraNet [10] and UACANet [8]. However, these methods are prone to inefficient feature fusion at the decoder due to the transmission of multi-stage features without filtering out irrelevant information.

To address these limitations, we propose a novel feature enhancement network for polyp segmentation that employs Feature Propagation Enhancement (FPE) modules to transmit multi-scale features from all stages to the decoder. Figure 1 illustrates a semantic comparison of our feature propagation scheme with the UNet-like model. While the existing UNet-like models use skip connections to propagate a single-scale feature, our method utilizes FPE to propagate multi-scale features from all stages in encoder. More importantly, this research highlights the usage of FPE can effectively replace skip connections by providing more comprehensive multi-scale characteristics from full stages in encoder. To further address the issue of high-level semantics being overwhelmed in the progressive feature fusion process, we also integrate a Feature Aggregation Enhancement (FAE) module that aggregates the outputs of FPE from previous stages at

decoder. Moreover, we introduce gate mechanisms in both FPE and FAE to filter out redundant information, prioritizing informative features for efficient feature fusion. Finally, we propose a Multi-Scale Aggregation (MSA) module appended to the output of the encoder to capture multi-scale features and provide the decoder with rich multi-scale semantic information. The MSA incorporates a cross-stage multi-scale feature aggregation scheme to facilitate the aggregation of multi-scale features. Overall, our proposed method improves upon existing UNet-like encoder-decoder architectures by addressing the limitations in feature propagation and feature aggregation, leading to improved polyp segmentation performance.

Our major contributions to accurate polyp segmentation are summarized as follows.

(1) The method addresses the limitations of the UNet-like encoder-decoder architecture by introducing three modules: Feature Propagation Enhancement (FPE), Feature Aggregation Enhancement (FAE), and Multi-Scale Aggregation (MSA).
(2) FPE transmits all encoder-extracted feature maps to the decoder, and FAE combines the output of the last stage at the decoder and multiple outputs from FPE. MSA aggregates multi-scale high-level features from FPEs to provide rich multi-scale information.
(3) The proposed method achieves state-of-the-art results in five polyp segmentation datasets and outperforms the previous cutting-edge approach by a large margin (3%) on CVC-ColonDB and ETIS datasets.

2   Method

Overview. Our proposed feature enhancement network illustrated in Fig. 2(a), is also a standard encoder-decoder architecture. Following Polyp-PVT [2], we adopt PVT [16] pretrained on ImageNet as the encoder. The decoder consists of three feature aggregation enhancement modules (FAE) and a multi-scale aggregation module (MSA). Given an input image $\mathcal{I}$, we first extract the pyramidal features using the encoder, which is defined as follows,

$$\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4 = \text{PVT}(\mathcal{I}) \tag{1}$$

where, $\{\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3, \mathcal{P}_4\}$ is the set of pyramidal features from four stages with the spatial size of 1/4, 1/8, 1/16, 1/32 of the input respectively. Features with lower spatial resolution usually contain richer high-level semantics. Then, these features are transmitted by the feature propagation enhancement module (FPE) to yield the feature set $\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}$, which provides multi-scale information from all the stages. This is different from the skip connection which only transmits the single-scale features at the present stage. Referring to [3,10], three highest-level features generated by FPEs are subsequently fed into a MSA module for aggregating rich multi-scale information. Afterwards, feature fusion is performed by
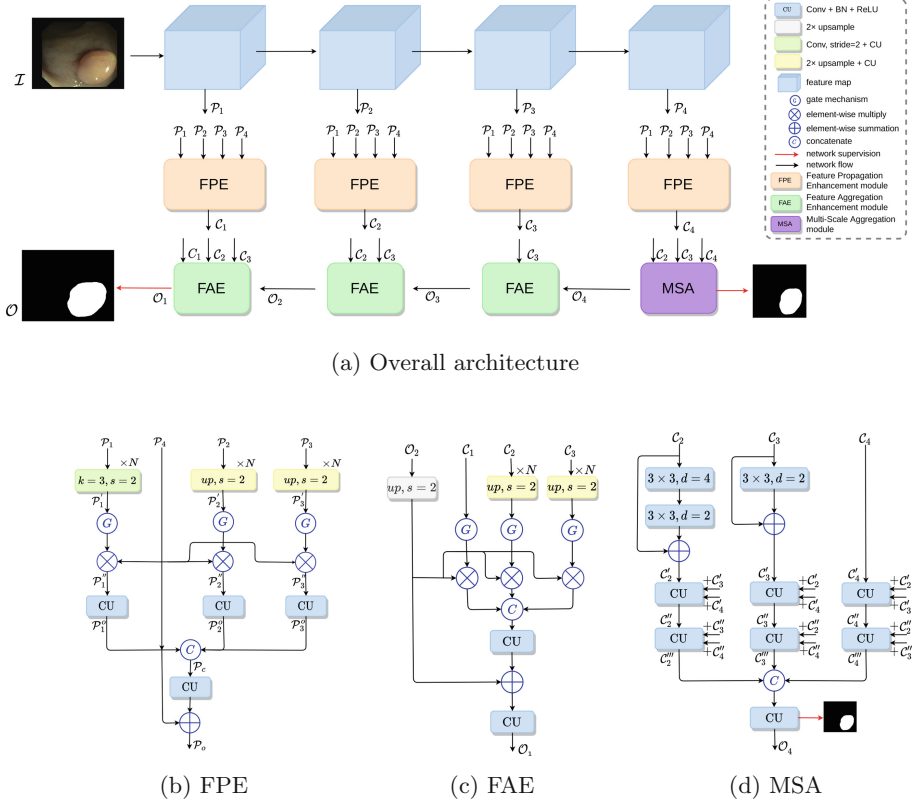
Fig. 2. An overview of the proposed framework. (a) Overall architecture; (b) FPE: Feature propagation enhancement module; (c) FAE: Feature aggregation enhancement module; (d) MSA: Multi-scale aggregation module

FAE in the decoder, whereby it progressively integrates the outputs from FPE and previous stages. The higher-level semantic features of the FPE output are capable of effectively compensating for the semantics that may have been overwhelmed during the upsampling process. This process is formulated as,

$$
\begin{aligned}
\mathcal{O}_4 &= \mathrm{MSA}(\{\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4\}) \\
\mathcal{O}_i &= \mathrm{FAE}(\{\mathcal{C}_i, \mathcal{C}_{i+1}, ..., \mathcal{C}_3\}, O_{i+1}) \quad i = 1, 2, 3.
\end{aligned}
\tag{2}
$$

A noteworthy observation is that the gating mechanism has been widely utilized in both FPE and FAE to modulate the transmission and integration of features. By selectively controlling the flow of relevant information, this technique has shown promise in enhancing the overall quality of feature representations [4,9]. The final features $\mathcal{O}_1$ are passed through the classifier (i.e., a 1×1 convolutional layer) to get the final prediction result in $\mathcal{O}$. Further details on FPE, FAE, and MSA will be provided in the following sections.

Feature Propagation Enhancement Module. In contrast to the traditional encoder-decoder architecture with skip connections, the FPE aims to transmit multi-scale information from full stage at the encoder to the decoder, rather than single-scale features at the current stage. The FPE architecture is illustrated in Fig. 2(b). The input of the FPE includes the features from the other three stages, in addition to the feature of the current stage, which delivers richer spatial and semantic information to the decoder.

However, these multi-stage inputs need to be downsampled or upsampled to match the spatial resolution of the features at the present stage. To achieve this, FPE employs a stepwise downsampling strategy, with each step only downsampling by a factor of 2, and each downsampling is followed by a Convolutional Unit (CU) to perform feature transformation. The number of downsamplings is denoted as $N = \log_2^T$, where T stands for the scale factor for downsampling. The CU consists of a 3×3 convolutional layer, a Batch Normalization layer, and an activation layer (i.e., ReLU). This strategy can be a highly effective means of mitigating the potential loss of intricate details during the process of large-scale interpolation. Similarly, this same strategy is employed in FAE.

The features from the other three stages, denoted as $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$, are downsampled or upsampled to generate $\mathcal{P}_1^{'}$, $\mathcal{P}_2^{'}$, and $\mathcal{P}_3^{'}$. Instead of directly combining the four inputs, FPE applies gate mechanisms to emphasize informative features. The gate mechanism takes the form of $\mathcal{Y}' = \mathrm{G}(\mathcal{X}) * \mathcal{Y}$, where G (in this work, Sigmoid is used) measures the importance of each feature vector in the reference feature $\mathcal{X} \in \mathbf{R}^{H \times W}$. By selectively enhancing useful information and filtering out irrelevant information, the reference features $\mathcal{X}$ assist in identifying optimal features $\mathcal{Y}$ at the current level. The output of G is in $[0, 1]^{H \times W}$, which controls the transmission of informative features from $\mathcal{Y}$ or helps filter useless information. Notably, $\mathcal{X}$ can be $\mathcal{Y}$ itself, serving as a reference feature. FPE leverages such gate mechanism to obtain informative features in $\mathcal{P}$ and passes them through a CU respectively. After that, FPE concatenates features from the four branches to accomplish feature aggregation. A CU is followed to boost the feature fusion.

Feature Aggregation Enhancement Module. The FAE is a novel approach that integrates the outputs of the last stages at the decoder with the FPE's outputs at both the current and deeper stages to compensate for the high-level semantics that may be lost in the process of progressive feature fusion. In contrast to the traditional encoder-decoder architecture with skip connections, the FAE assimilates the output of the present and higher-stage FPEs, delivering richer spatial and semantic information to the decoder.

The FAE, depicted in Fig. 2(c), integrates the outputs of the current and deeper FPE stages with high-level semantics. As an example, the last FAE takes as inputs $\mathcal{O}_2$ (output of the penultimate FAE), $\mathcal{C}_1$ (output of the current FPE stage), and $\mathcal{C}_2$ and $\mathcal{C}_3$ (outputs of FPE from deeper stages). Multiple outputs from deeper FPE stages are introduced to compensate for high-level semantics. Furthermore, gate mechanisms are utilized to filter out valueless features

for fusion, and the resulting enhanced feature is generated by a CU after concatenating the filtered features. Finally, $\mathcal{O}_2$ is merged with the output feature through element-wise summation, followed by a CU to produce the final output feature $\mathcal{O}_1$.

Multi-Scale Aggregation Module. The MSA module in our proposed framework, inspired by the Parallel Partial Decoder in PraNet [3], integrates three highest-level features $\mathcal{C}_2$, $\mathcal{C}_3$, and $\mathcal{C}_4$ from the FPE output. This provides rich multi-scale information for subsequent feature aggregation in the FAE and also helps to form a coarse localization of polyps under supervision. It benefits from an additional supervision signal, as observed in PraNet [3], CaraNet [8], and etc. by aiding in forming a coarse location of the polyp and contributing to improved accuracy and performance. As depicted in Fig. 2(d), the MSA module first processes these three features separately. $\mathcal{C}_2$, which has the highest feature resolution, is processed with multiple dilated convolutions to capture its multi-scale information while keeping its spatial resolution unchanged. $\mathcal{C}_3$ is processed with only one dilated convolution due to its higher spatial resolution, while $\mathcal{C}_4$ is not processed since it already contains the richest contextual information. The output features of the three branches are then upsampled to the size of $\mathcal{C}_2$. To better integrate these three multi-scale high-level features for subsequent fusion, additional cross-feature fusion operations (i.e., 2 CU layer) are performed in the MSA module.

3   Experiments

Datasets. We conduct extensive experiments on five polyp segmentation datasets, including Kvasir [6], CVC-ClinicDB [1], CVC-ColonDB [13], ETIS [12] and CVC-T [14]. Following the setting in [2,3,5,10,10,17,19], the model is trained using a fraction of the images from CVC-ClinicDB and Kvasir, and its performance is evaluated by the remaining images as well as those from CVC-T, CVC-ColonDB, and ETIS. In particular, there are 1450 images in the training set, of which 900 are from Kvasir and 550 from CVC-ClinicDB. The test set contains all of the images from CVC-T, CVC-ColonDB, and ETIS, which have 60, 380, and 196 images, respectively, along with the remaining 100 images from Kvasir and the remaining 62 images from CVC-ClinicDB.

Implementations. We utilize PyTorch 1.10 to run experiments on an NVIDIA RTX3090 GPU. We set an initial learning rate to 1e-4 and halve it after 80 epochs. We train the model for 120 epochs. The same multi-scale input and gradient clip strategies used in [3,8,10] are employed in the training phase. The batch size is 16 by default. AdamW is selected as the optimizer with a weight decay of 1e-4. We adopt the same data augmentation techniques as UACANet [8], including random flip, random rotation, and color jittering. In evaluation phase, we mainly focus on mDice, mIoU, the two most common metrics in medical image

segmentation, to evaluate the performance of the model. Referring to [3,8,10], we use a combination loss consisting of weighted Dice loss and weighted IoU loss to supervise network optimization.

Table 1. Quantitative results on CVC-ClinicDB, Kvasir, CVC-T, CVC-ColonDB and ETIS.

Methods	Published Venue	CVC-ClinicDB		Kvasir		CVC-ColonDB		ETIS		CVC-T	
		mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
UNet [11]	MICCAI'15	0.823	0.755	0.818	0.746	0.512	0.444	0.398	0.335	0.710	0.627
PraNet [3]	MICCAI'19	0.899	0.849	0.898	0.840	0.709	0.640	0.628	0.567	0.871	0.797
ResUNet++ [7]	JBHI'21	0.846	0.786	0.807	0.727	0.588	0.497	0.337	0.275	0.687	0.598
SANet [17]	MICCAI'21	0.916	0.859	0.904	0.847	0.752	0.669	0.750	0.654	0.888	0.815
MSNet [19]	MICCAI'21	0.915	0.866	0.902	0.847	0.747	0.668	0.720	0.650	0.862	0.796
UACANet-S [8]	MM'21	0.916	0.870	0.905	0.852	0.783	0.704	0.694	0.615	0.902	0.837
UACANet-L [8]	MM'21	0.926	0.880	0.912	0.859	0.751	0.678	0.766	0.689	**0.910**	**0.849**
Polyp-PVT [2]	arxiv'21	**0.937**	**0.889**	0.917	0.864	0.808	0.727	0.787	0.706	0.900	0.833
CaraNet [10]	SPIE MI'22	0.921	0.876	0.913	0.859	0.775	0.700	0.740	0.660	0.902	0.836
LDNet [18]	MICCAI'22	0.923	0.872	0.912	0.855	0.794	0.715	0.778	0.707	0.893	0.826
SSFormer-S [15]	MICCAI'22	0.916	0.873	0.925	0.878	0.772	0.697	0.767	0.698	0.887	0.821
SSFormer-L [15]	MICCAI'22	0.906	0.855	0.917	0.864	0.802	0.721	0.796	0.720	0.895	0.827
Ours	-	0.931	0.885	**0.928**	**0.880**	**0.837**	**0.759**	**0.822**	**0.746**	0.905	0.839

Comparison with State-of-the-Art Methods. We compared our proposed method with previous state-of-the-art methods. According to the experimental settings, the results on CVC-ClinicDB and Kvasir demonstrate the learning ability of the proposed model, while the results on CVC-T, CVC-ColonDB, and ETIS demonstrate the model's ability for cross-dataset generalization. The experimental results are listed in Table.1. It can be seen that our model is slightly inferior to Polyp-PVT on the CVC-ColonDB, but the gap is quite small, e.g., 0.6% in mDice and 0.4% in mIoU. On Kvasir, we are ahead of the previous best model by 1.1% in mDice and 1.6% in mIoU. This shows that our model is second to none in terms of learning ability, which demonstrates the effectiveness of our model.

Furthermore, our proposed method demonstrates strong cross-dataset generalization capability on CVC-T, CVC-ColonDB, and ETIS datasets, with particularly good performance on the latter two due to their larger and more representative datasets. Our model outperforms state-of-the-art models by 2.9% mDice and 3.2% mIoU on CVC-ColonDB and 3.5% mDice and 4.0% mIoU on ETIS. These results validate the effectiveness of feature-level enhancement and highlight the superior performance of our method. We also provide visual results in Fig. 3, where our predictions are shown to be closer to the ground truth.

Ablation Study. We carried out ablation experiments to verify the effectiveness of the proposed FPE, FAE, and MSA. For our baseline, we use the simple encoder-decoder structure with skip connections for feature fusion and perform element-wise summation at the decoder. Table 2 presents the results of our ablation experiments. Following the ablation study conducted on our proposed

approach, it is with confidence that we assert the significant contribution of each module to the overall performance enhancement compared to the baseline. Our results indicate that the impact of each module on the final performance is considerable, and their combination yields the optimal overall performance. Specifically, across the five datasets, our proposed model improves the mDice score by at least 1.4% and up to 3.4% on CVC-T, compared to the baseline. For mIoU, the improvements are 1.5% and 3.5% on the corresponding datasets. In summary, our ablation study underscores the crucial role played by each component of our approach, and establishes its potential as a promising framework for future research in this domain.

Table 2. Ablation study on FPE, MSA, and FAE. Each item presents mDice/mIoU.

FPE	MSA	FAE	CVC-T	CVC-ClinicDB	Kvasir	CVC-ColonDB	ETIS
			87.1/80.4	91.6/87.0	91.4/86.3	81.0/72.8	79.9/71.5
✓			90.2/83.7	92.7/87.8	91.5/86.4	81.0/73.1	80.2/71.9
	✓		89.9/83.2	91.6/86.8	91.9/87.0	82.1/73.9	81.3/73.1
		✓	90.0/83.3	91.8/87.2	91.7/86.6	81.4/73.3	81.2/73.3
✓	✓		90.1/83.4	92.6/87.7	92.2/87.2	81.6/73.4	80.9/72.9
✓		✓	90.2/83.5	93.0/88.4	92.1/87.2	81.5/73.4	80.9/72.9
	✓	✓	90.0/83.6	92.7/88.0	92.6/87.8	81.2/73.1	81.6/73.9
			90.5/83.9	**93.1/88.5**	**92.8/88.0**	**83.7/75.9**	**82.2/74.6**
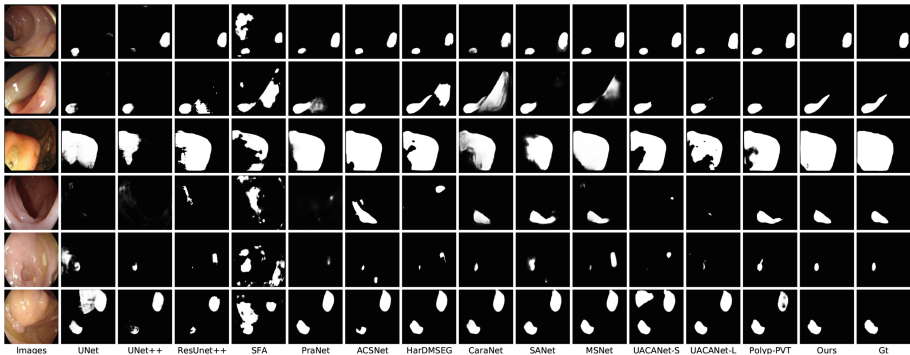
Fig. 3. Exemplary images and results that are segmented by different approaches.

4 Conclusion

We introduce a new approach to polyp segmentation that addresses inefficient feature propagation in existing UNet-like encoder-decoder networks. Specifically,

a feature propagation enhancement module is introduced to propagate multi-scale information over full stages in the encoder, while a feature aggregation enhancement module is attended at the decoder side to prevent the loss of high-level semantics during progressive feature fusion. Furthermore, a multi-scale aggregation module is used to aggregate multi-scale features to provide rich information for the decoder. Experimental results on five popular polyp datasets demonstrate the effectiveness and superiority of our proposed method. Specifically, it outperforms the previous cutting-edge approach by a large margin (3%) on CVC-ColonDB and ETIS datasets. To extend our work, our future direction focuses on exploring more effective approaches to feature utilization, such that efficient feature integration and propagation can be achieved even on lightweight networks.

References

1. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. CMIG **43**, 99–111 (2015)
2. Bo, D., Wenhai, W., Jinpeng, L., Deng-Ping, F.: Polyp-pvt: Polyp segmentation with pyramid vision transformers. arXiv preprint arXiv:2108.06932v3 (2021)
3. Fan, D.P., et al.: PraNet: parallel reverse attention network for polyp segmentation. In: MICCAI (2020)
4. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
5. Huang, C.H., Wu, H.Y., Lin, Y.L.: Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. arXiv preprint arXiv:2101.07172 (2021)
6. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37
7. Jha, D., et al.: Resunet++: An advanced architecture for medical image segmentation. In: ISM, pp. 225–2255. IEEE (2019)
8. Kim, T., Lee, H., Kim, D.: UACANet: Uncertainty augmented context attention for polyp segmentation. In: ACM MM, pp. 2167–2175 (2021)
9. Li, X., Zhao, H., Han, L., Tong, Y., Yang, K.: GFF: Gated fully fusion for semantic segmentation. arxiv 2019. arXiv preprint arXiv:1904.01803
10. Lou, A., Guan, S., Ko, H., Loew, M.H.: CaraNet: context axial reverse attention network for segmentation of small medical objects, pp. 81–92. SPIE (2022)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)

12. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. IJCARS **9**(2), 283–293 (2014)
13. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. TMI **35**(2), 630–644 (2015)
14. Vázquez, D., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. JHE (2017)
15. Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., Song, S.: Stepwise feature fusion: local guides global. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III, pp. 110–120. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_11
16. Wang, W., et al.: Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: ICCV, pp. 568–578 (2021)
17. Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S.K., Cui, S.: Shallow Attention Network for Polyp Segmentation. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I, pp. 699–708. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_66
18. Zhang, R., et al.: Lesion-aware dynamic kernel for polyp segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III, pp. 99–109. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_10
19. Zhao, X., Zhang, L., Lu, H.: Automatic polyp segmentation via multi-scale subtraction network. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I, pp. 120–130. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_12