Polyp-Mamba: Polyp Segmentation with Visual Mamba

Zhongxing Xu[1*], Feilong Tang[2*✉], Zhe Chen[3], Zheng Zhou[3], Weishan Wu[3], Yuyao Yang[3], Yu Liang[3], Jiyu Jiang[3], Xuyue Cai[3], and Jionglong Su[3✉]

[1] Weill Cornell Medicine, Cornell University
[2] Monash University
[3] Xi'an Jiaotong-Liverpool University

Abstract. Accurate segmentation of polyps is crucial for efficient colorectal cancer detection during the colonoscopy screenings. State Space Models, exemplified by Mamba, have recently emerged as a promising approach, excelling in long-range interaction modeling with linear computational complexity. However, previous methods do not consider the cross-scale dependencies of different pixels and the consistency in feature representations and semantic embedding, which are crucial for polyp segmentation. Therefore, we introduce Polyp-Mamba, a novel unified framework aimed at overcoming the above limitations by integrating multi-scale feature learning with semantic structure analysis. Specifically, our framework includes a **Scale-Aware Semantic** module that enables the embedding of multi-scale features from the encoder to achieve semantic information modeling across both intra- and inter-scales, rather than the single-scale approach employed in prior studies. Furthermore, the **Global Semantic Injection** module is deployed to inject scale-aware semantics into the corresponding decoder features, aiming to fuse global and local information and enhance pyramid feature representation. Experimental results across five challenging datasets and six metrics demonstrate that our proposed method not only surpasses state-of-the-art methods but also sets a new benchmark in the field, underscoring the Polyp-Mamba framework's exceptional proficiency in the polyp segmentation tasks.

Keywords: Polyp Segmentation · Mamba · Scale-Aware Semantic.

1 Introduction

Colonic polyps, identified as protrusions within the colonic mucosa, display considerable variability in shape, texture, and color [1]. Colonic polyps are considered precancerous lesions, closely associated with the development of colon cancer [2]. Multiple studies have shown that early colonoscopic examinations can reduce the incidence of colorectal cancer by 30% [3]. Therefore, accurate polyp segmentation is crucial in clinical practice. While traditional CNN models, such as FCN [4], struggle with long-term dependencies, advanced methods

[*] The first two authors contribute equally to this work.

(*e.g.,* UNet [5]) with encoder-decoder structures improve segmentation by combining features of different resolutions. Despite the great success of this method in dense prediction, it remains limited by inefficient non-local context modeling between arbitrary locations, dimming the prospects for further enhancing the accuracy of complex visual interpretations.

Although CNN-based [6–10] and Transformer-based [11, 12] models represent two dominant approaches in image segmentation and classification, each has its inherent limitations. The local receptive fields of CNN-based models restrict their ability to process long-range information, leading to inadequate feature extraction and, consequently, suboptimal segmentation outcomes. Transformer-based models, despite their superior global modeling capabilities, suffer from the drawback of requiring computational resources that scale quadratically with image size [13, 14]. This becomes particularly burdensome for dense prediction tasks, such as medical image segmentation. To address these current shortcomings, we develop a novel medical image segmentation architecture that is capable of capturing powerful long-range information while maintaining linear computational complexity.

Recently, State Space Models (SSMs) have garnered significant interest among researchers. Built on the foundation of classical SSM research [15], modern SSMs, such as Mamba [16], not only establish long-range dependencies but also exhibit linear complexity relative to input size. This architecture has been applied in various computer vision tasks, including Vision Mamba [17], UMamba [18], Segmamba [19], MambaUNet [20], and VM-UNet [21]. However, the current designs remain sub-optimal for medical image segmentation for two reasons. First, they primarily depend on self-attention mechanisms or Visual State Space (VSS) for context modeling at a singular scale, overlooking the cross-scale dependencies. Second, these approaches exhibit a lack of consistency in feature representations across scales. These omissions are particularly substantial in polyp segmentation, where the significant size variations of polyps and the blurred boundaries between polyps and surrounding mucosa make accurately locating polyp areas even more challenging.

To overcome these limitations, this paper introduces the Polyp-Mamba framework, which combines rich context modeling and semantic relationship mining to achieve accurate polyp segmentation. Polyp-Mamba advances beyond the limitations of existing models by incorporating a *Scale-Aware Semantic (SAS)* module and a *Global Semantic Injection (GSI)* module, as shown in Fig. 1. The SAS module utilizes VSS blocks to analyze and interpret semantic information at multiple scales, facilitating the modeling of semantic data from detailed to broad granularity. Meanwhile, the GSI module allows for the incorporation of scale-aware semantics into relevant features, promoting the synthesis of global and local information into potent hierarchical features. Specifically, we employ a cross-attention mechanism to facilitate interaction among feature representations. The significance of this mechanism lies in the capacity of global semantics to comprehend information across different scales, allowing for the integration of global semantics into the corresponding local features. Consequently, we can

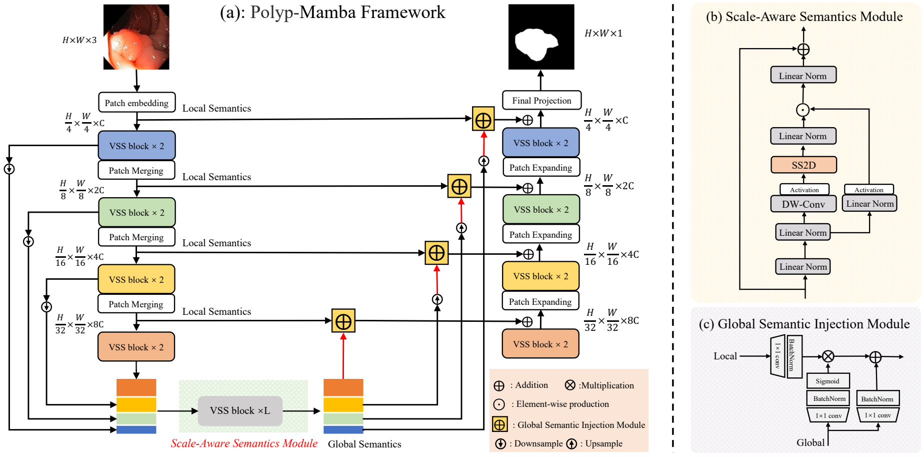

Fig. 1. Overview of the proposed Polyp-Mamba, which consists of (a) Polyp-Mamba Framework (b) Scale-Aware Semantics Module, and (c) Semantic Injection Module.

achieve a more nuanced and comprehensive feature enhancement, thereby improving the model's recognition of both details and global structures. We find that the guidance of global semantics helps to effectively handle polyps of varying sizes, shapes, and textures, while reducing instances of over-segmentation and under-segmentation.

In summary, the main contributions of this paper are threefold: (1) We pioneer the use of the Mamba model for polyp segmentation, effectively addressing long-term dependencies while maintaining linear computational complexity. Additionally, our proposed Polyp-Mamba framework uniquely enhances cross-scale contextual dependencies and fine-tunes semantic relationships for precise polyp segmentation. (2) We introduce an SAS Module and a GSI Module. SAS is responsible for implementing pixel-level context modeling across scales, while GSI infuses global information into local features. These two modules collaboratively improve the model's ability to recognize polyps of varying sizes, shapes, and textures. (3) Extensive experimental results demonstrate that the proposed Polyp-Mamba outperforms most contemporary models on five challenging datasets, showcasing our model's superior capability in accurate polyp detection.

2 Method

2.1 Architecture Overview

The proposed Polyp-Mamba architecture is outlined in Fig. 1, drawing inspiration from UNet [5] and Swin-UNet [22]. The Patch Embedding layer first divides the input image $x \in \mathbb{R}^{H\times W\times 3}$ into non-overlapping 4×4 patches, and then maps the image dimensions to C, resulting in an embedded image $x' \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$.

Layer Normalization [23] is applied to x' before it enters the encoder for feature extraction. The encoder consists of four stages, employing patch merging operations at the end of the first three stages to reduce the height and width of the input features while doubling the number of channels. Across the four stages, we use $[2, 2, 2, 2]$ VSS blocks, with the channel counts for each stage being $[C, 2C, 4C, 8C]$. Mirroring the encoder, the decoder comprises VSS blocks and patch-expanding layers, the latter restoring details lost in downsampling through skip connections to achieve the same feature size output as the encoder. Following the decoder, a Final Projection layer performs 4-times upsampling through patch expansion to recover the original height and width of the features, followed by a projection layer to adjust the number of channels to match the segmentation target. VSS blocks serve as SAS module, taking features of various scales from the encoder as input and generating semantics with scale-aware capabilities. These semantics are infused into corresponding scale features through the GSI module to enhance the model's representational power. Further details of these modules are discussed in subsequent sections.

2.2 Encoder and Decoder

In the Polyp-Mamba framework, the encoder first processes the reduced-resolution C-dimensional tokenized input through two consecutive VSS blocks for feature learning, while maintaining the original dimensions and resolution. The encoder then applies a triple patch merging operation as a downsampling process, dividing the input into quadrants one-fourth the size, concatenating them, and normalizing dimensions via layer normalization, thus halving the token count and doubling feature dimensions. Correspondingly, the decoder mirrors the structure of the encoder, utilizing two consecutive VSS blocks and patch-expanding layers for feature reconstruction. Unlike the merging layers in the encoder, the decoder's expanding layers enhance resolution through $2\times$ upsampling and halve the feature dimensions. This design allows the decoder to both enhance resolution and recover deep features, providing rich spatial details for the final predictions.

2.3 Scale-Aware Semantics Module

The VSS block, derived from VMamba [24], is the core module of Poly-Mamba, with the SAS module consisting of several stacked VSS blocks. The number of VSS blocks is denoted as L. Each VSS block includes a 2D-Selective-Scan (SS2D) module, linear layer, and residual connection, as shown in Fig. 1 (b). After layer normalization, the input is split into two branches. In the first branch, the input passes through a linear layer and activation function. In the second branch, the input is processed through a linear layer, depth-wise separable convolution, and activation function before being fed into the SS2D module for further feature extraction. Subsequently, the features are normalized using layer normalization and combined with the output of the first branch through element-wise multiplication to merge the two pathways. Finally, the features are mixed using a linear layer, and this outcome is combined with a residual connection to form the output of the VSS block. SiLU [25] is used as the default activation function.

As shown in Fig. 1 (a), the scale-aware semantic module takes features of various scales as inputs. To further reduce computational load, an average pooling operator is used to reduce the number of tokens from different scales to $\frac{1}{64\times64}$ of the input size. The pooled tokens from different scales now have the same resolution and are concatenated together as the input for the VSS block. This module is capable of obtaining a full-image receptive field and rich semantics. Specifically, the SS2D enables spatial information exchange, while the convolution layer enables cross-scale feature interaction. In each VSS block, after exchanging information of features from all scales, a residual mapping is learned and then added to the features to enhance representation and semantics. Finally, scale-aware semantics are obtained after processing through several VSS blocks.

2.4 Global Semantics Injection Module

After obtaining scale-aware semantics, we directly integrate them with other features $\mathbf{T}^N$. However, there exists a significant semantic gap between features $\{\mathbf{T}^1, \ldots, \mathbf{T}^N\}$ and scale-aware semantics, which may cause difficulties in accurately identifying the boundaries between polyps and normal tissues, affecting the quality of segmentation results. The GSI module is therefore introduced to bridge the semantic gap before merging features by cross-attention mechanism. As illustrated in Fig. 1 (c), GSI takes different local features from the encoder and global semantics from VSS as input. The local features pass through a 1×1 convolutional layer, followed by batch normalization, to generate the feature to be injected. The global semantics are fed into a 1×1 convolutional layer, followed by batch normalization and a sigmoid layer, to produce semantic weights. Simultaneously, the global semantics also pass through a 1×1 convolutional layer and batch normalization. The three outputs from these processes are uniform in size. These global semantics are then injected into the local tokens by Hadamard product, and the injected features are added to the global semantics. The outputs of several GSIs share the same number of channels, denoted as M.

2.5 Loss function

We aim to optimize the performance of Polyp-Mamba, an SSM-based model, in polyp segmentation tasks. [26] reports that combining multiple loss functions with adaptive weights at different levels can improve the performance of the framework with better convergence speed. Therefore, we use binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ and the weighted IoU loss $\mathcal{L}_{\text{Iou}}$ for supervision. λ_1 and λ_2 are the weighting coefficients. The loss $\mathcal{L}_{total}$ for the proposed Polyp-Mamba can be formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{Iou}} \tag{1}$$

Table 1. Quantitative results on Kvasir and ClinicDB datasets to validate our model's learning ability. 'n/a' denotes that the results are not available.

	Methods	mean Dice	mean IoU	F_β^w	S_α	$E_\phi^{\max}$	MAE
Kvasir	U-Net [MICCAI'15] [5]	0.818	0.746	0.794	0.858	0.893	0.055
	PraNet [MICCAI'20] [33]	0.898	0.840	0.885	0.915	0.948	0.030
	Transfuse [MICCAI'21] [34]	0.913	0.857	n/a	n/a	n/a	n/a
	SSformer [MICCAI'2022] [11]	0.926	0.874	n/a	n/a	n/a	n/a
	PolypPVT [AIR'2023] [32]	0.917	0.864	0.911	0.925	0.962	0.023
	CoInNet [TMI'2023] [35]	0.926	0.872	0.939	0.926	0.979	0.020
	VM-Unet [arxiv'24] [21]	0.913	0.856	0.902	0.918	0.958	0.027
	Polyp-Mamba (Ours)	**0.940**	**0.881**	**0.942**	**0.935**	**0.983**	**0.016**
ClinicDB	U-Net [MICCAI'15] [5]	0.823	0.755	0.811	0.889	0.954	0.019
	PraNet [MICCAI'20] [33]	0.899	0.849	0.896	0.936	0.979	0.009
	Transfuse [MICCAI'21] [34]	0.935	0.887	n/a	n/a	n/a	n/a
	SSformer [MICCAI'2022] [11]	0.927	0.876	n/a	n/a	n/a	n/a
	PolypPVT [AIR'2023] [32]	0.937	0.889	0.936	0.949	0.989	0.006
	CoInNet [TMI'2023] [35]	0.930	0.887	0.940	0.952	0.987	0.006
	VM-Unet [arxiv'24] [21]	0.926	0.871	0.927	0.933	0.971	0.009
	Polyp-Mamba (Ours)	**0.949**	**0.907**	**0.952**	**0.965**	**0.993**	**0.005**

3 Experiments and Results

Datasets and Metrics: Experiments are conducted on five polyp segmentation datasets (ETIS [27], CVC-ClinicDB (CVC-612) [28], CVC-ColonDB (ColonDB) [29], EndoScene-CVC300 (EndoScene) [30], Kvasir-SEG (Kvasir) [31]). We follow the same training/testing protocols in [32, 33], where 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. Specifically, the training set comprises 1,450 images, with 900 from Kvasir and 550 from CVC-ClinicDB. The test set contains all images from CVC-T, CVC-ColonDB, and ETIS, which are 60, 380, and 196 images, respectively, along with the remaining 100 images from Kvasir and the remaining 62 images from CVC-ClinicDB.

Implementation details. We use rotation and horizontal flips for data augmentation. The input resolution during training is set to 224 × 224, and the batch size is set to 8. The number of iterations during training is 50k. We use the AdamW optimizer for training with an initial learning rate of 0.0002, a momentum of 0.9, and a weight decay of 1e-4. All the experiments are conducted on one NVIDIA RTX 4090 GPU. The loss coefficients λ_1 and λ_2 are both set as 1 in Eq. 1. We employ two metrics (i.e., mean Dice and mean IoU) for quantitative evaluation, similar to [33, 36, 37, 36]. To provide deeper insight into the model performance, we introduce four key metrics to evaluate model performance followed by [33, 32]. The weighted Dice metric F_β^w addresses the "equal-importance flaw" inherent in the traditional Dice metric, while MAE measures pixel accuracy. For broader assessment, we use the enhanced alignment metric $E_\phi^{\max}$ [38] and S_α [39] to capture both pixel-level and structural similarities between predictions and ground truths.

Learning Ability. In this section, we conduct two experiments on two seen datasets to validate the learning capability of our model, namely Kvasir and

Table 2. Quantitative results on ColonDB, ETIS, and EndoScene datasets. 'n/a' denotes that the results are not available.

	Methods	mean Dice	mean IoU	F_β^w	S_α	$E_\phi^{\max}$	MAE
ColonDB	U-Net [MICCAI'15] [5]	0.512	0.444	0.498	0.712	0.776	0.061
	PraNet [MICCAI'20] [33]	0.709	0.640	0.696	0.819	0.869	0.045
	Transfuse [MICCAI'21] [34]	0.781	0.699	n/a	n/a	n/a	n/a
	SSformer [MICCAI'2022] [11]	0.802	0.721	n/a	n/a	n/a	n/a
	PolypPVT [AIR'2023] [32]	0.808	0.727	0.795	0.865	0.919	0.031
	CoInNet [TMI'2023] [35]	0.797	0.729	0.789	0.875	0.897	0.022
	VM-Unet [arxiv'24] [21]	0.798	0.712	0.782	0.861	0.904	0.036
	Polyp-Mamba (Ours)	**0.829**	**0.743**	**0.815**	**0.881**	**0.931**	**0.027**
ETIS	U-Net [MICCAI'15] [5]	0.398	0.335	0.366	0.684	0.740	0.036
	PraNet [MICCAI'20] [33]	0.628	0.567	0.600	0.794	0.841	0.031
	Transfuse [MICCAI'21] [34]	0.731	0.660	n/a	n/a	n/a	n/a
	SSformer [MICCAI'2022] [11]	0.796	0.720	n/a	n/a	n/a	n/a
	PolypPVT [AIR'2023] [32]	0.787	0.706	0.750	0.871	0.910	0.013
	CoInNet [TMI'2023] [35]	0.759	0.690	0.820	0.859	0.898	0.024
	VM-Unet [arxiv'24] [21]	0.761	0.692	0.743	0.869	0.900	0.015
	Polyp-Mamba (Ours)	**0.825**	**0.747**	**0.766**	**0.889**	**0.923**	**0.012**
EndoScene	U-Net [MICCAI'15] [5]	0.710	0.627	0.684	0.843	0.875	0.022
	PraNet [MICCAI'20] [33]	0.871	0.797	0.843	0.925	0.972	0.010
	Transfuse [MICCAI'21] [34]	0.893	0.824	n/a	n/a	n/a	n/a
	SSformer [MICCAI'2022] [11]	0.895	0.827	n/a	n/a	n/a	n/a
	PolypPVT [AIR'2023] [32]	0.900	0.833	0.884	0.935	0.981	0.007
	CoInNet [TMI'2023] [35]	0.909	0.863	0.881	0.942	0.989	**0.005**
	VM-Unet [arxiv'24] [21]	0.886	0.818	0.849	0.921	0.968	0.009
	Polyp-Mamba (Ours)	**0.921**	**0.875**	**0.895**	**0.948**	**0.993**	**0.005**

CVC-ClinicDB. Kvasir contains 1,000 images selected from a subclass (polyp class) of the Kvasir dataset. CVC-ClinicDB includes 612 open-access images from 31 colonoscopy clips. As shown in Table 1, our Polyp-Mamba significantly outperforms all SOTA models across all metrics on both datasets. This demonstrates that our model has a strong learning ability to acquire sufficient features from complex data for accurately identifying and segmenting polyps.

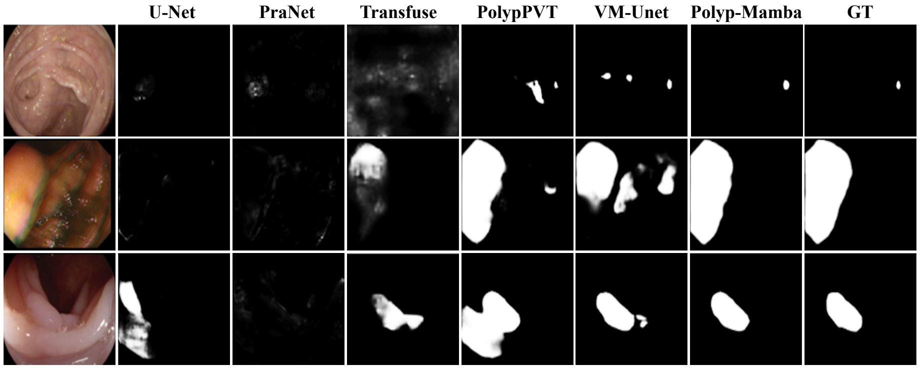

Fig. 2. Qualitative results of different methods.

Generalization Capability. We conducted three experiments to test the model's generalization capability on three unseen datasets, namely CVC-ColonDB, ETIS, and EndoScene (a combination of CVC-612 and CVC-300), each with its own

Table 3. Ablation study on main components of the proposed framework on the CVC-612 and CVC300 datasets. Backbone: U-shape architecture model for medical image segmentation, like VM-UNet. “w/ GSI": Only add the GSI module and employ the simple addition injection method replaces SAS. “w/ SAS": Only add the SAS module and employ the last layer features of the encoder as global semantics. “w/ GSI, SAS": Add GSI and SAS modules to Backbone.

Settings	ClinicDB (***seen***)		EndoScene(***unseen***)	
	mean Dice	mean IoU	mean Dice	mean IoU
Backbone (No.1)	0.916	0.861	0.871	0.811
w/ GSI (No.2)	0.926	0.889	0.895	0.831
w/ SAS (No.3)	0.941	0.877	0.915	0.859
w/ GSI, SAS (No.4)	**0.949**	**0.907**	**0.921**	**0.875**

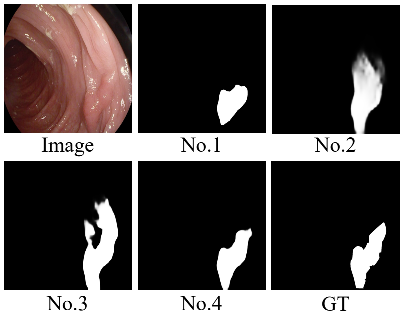

Fig 3. Qualitative results on ClinicDB and EndoScene datasets using different components, corresponding to settings (No.1-4) in Table 3.

challenging scenarios and attributes. Following the methodology of Fang et al. [40], we divided them into training, validation, and testing subsets. As shown in Table 2, our Polyp-Mamba once again significantly outperforms the existing classic medical segmentation baselines (*i.e.,* U-Net, Transformer) on all three unseen datasets. Our model also demonstrates its strong generalization ability in handling complex and diverse data.

Qualitative Results. As shown in Fig. 2, we present the polyp segmentation results of our Polyp-Mamba on the test set, accurately identifying and segmenting polyp in various challenging situations. These challenges include the varying sizes of polyps, homogeneous regions, and the diverse textures on the polyp surface. Whether amidst highly similar backgrounds or complex textures, Polyp-Mamba effectively differentiates and precisely locates polyps.

Ablation Study. We analyze each component of our Polyp-Mamba on the *seen* and *unseen* datasets to provide deeper insight into our model. As shown in Table 3, we investigate the importance of the SAS Module, and quantitative results indicate that configuration No. 2 (backbone + SAS) outperforms No. 1 (backbone only), clearly demonstrating the necessity of the SAS mechanism for performance enhancement. Note that we simply add global semantics to each feature layer. Furthermore, we investigate the contribution of the GSI Module, with results listed in the first and third columns of Table 3. We observe that configuration No. 3 improves the performance of the backbone (No. 1) on the ClinicDB dataset, increasing the average Dice score from 0.916 to 0.941. These improvements suggest that introducing the GSI component enables our model to accurately distinguish true polyp tissues. To assess the combined effect of the SAS and GSI modules, we test the performance of No. 4 (SAS + GSI + Backbone). Our Polyp-Mamba (No. 4) performs better than other settings (No. 1, No. 2, No. 3). As shown in Fig. 3, visual comparisons indicate that the baseline performs poorly. Our method improves polyp segmentation accuracy through SAS and GSI modules, eliminating false-positive areas, especially near adherent edges or in low-contrast areas.

4 Conclusion

In this study, we introduce a novel framework Polyp-Mamba for colonic polyp segmentation. Our key idea is to integrate advanced context modeling and semantic relationship mining, tailored to address the unique challenges presented by the variability in the appearance of polyps. To this end, we have developed two novel modules: the SAS Module for discerning semantic information across scales, and the GSI Module for merging this information with local features to construct a robust hierarchical representation. Extensive experimental results on five polyp segmentation datasets demonstrate that we outperform previous state-of-the-art results by a large margin.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. B. D. Pooler, D. H. Kim, K. A. Matkowskyj, M. A. Newton, R. B. Halberg, W. M. Grady, C. Hassan, and P. J. Pickhardt, "Growth rates and histopathological outcomes of small (6–9 mm) colorectal polyps based on ct colonography surveillance and endoscopic removal," 2023.
2. R. Djinbachian, R. Iratni, M. Durand, P. Marques, and D. von Renteln, "Rates of incomplete resection of 1-to 20-mm colorectal polyps: a systematic review and meta-analysis," *Gastroenterology*, 2020.
3. F. A. Haggar and R. P. Boushey, "Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors," *Clinics in colon and rectal surgery*, 2009.
4. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.
5. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.
6. W. Li, X. Xiong, S. Li, and F. Fan, "Hybridvps: Hybrid-supervised video polyp segmentation under low-cost labels," *IEEE Signal Processing Letters*, 2023.
7. W. Li, W. Lu, J. Chu, and F. Fan, "Lacinet: A lesion-aware contextual interaction network for polyp segmentation," *IEEE Transactions on Instrumentation and Measurement*, 2023.
8. F. Tang, Z. Xu, Z. Qu, W. Feng, X. Jiang, and Z. Ge, "Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation," in *CVPR*, 2024.
9. P. Xia, M. Hu, F. Tang, W. Li, W. Zheng, L. Ju, P. Duan, H. Yao, and Z. Ge, "Generalizing to unseen domains in diabetic retinopathy with disentangled representations," in *MICCAI*, 2024.
10. X. Zhao, F. Tang, X. Wang, and J. Xiao, "Sfc: Shared feature calibration in weakly supervised semantic segmentation," in *AAAI*, 2024.
11. J. Wang, Q. Huang, F. Tang, J. Meng, J. Su, and S. Song, "Stepwise feature fusion: Local guides global," in *MICCAI*, Springer, 2022.
12. F. Tang, Z. Xu, Q. Huang, J. Wang, X. Hou, J. Su, and J. Liu, "Duat: Dual-aggregation transformer network for medical image segmentation," in *PRCV*, 2023.

13. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
14. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017.
15. R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.
16. A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.
17. L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.
18. J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.
19. Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation," *arXiv preprint arXiv:2401.13560*, 2024.
20. Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li, "Mamba-unet: Unet-like pure visual mamba for medical image segmentation," *arXiv preprint arXiv:2402.05079*, 2024.
21. J. Ruan and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *arXiv preprint arXiv:2402.02491*, 2024.
22. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *ECCV*, 2022.
23. J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
24. Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.
25. S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural networks*, 2018.
26. J. Wei, S. Wang, and Q. Huang, "F^{3}net: fusion, feedback and focus for salient object detection," in *AAAI*, 2020.
27. D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of healthcare engineering*, 2017.
28. J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International journal of computer assisted radiology and surgery*, 2014.
29. J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.
30. N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE transactions on medical imaging*, 2015.
31. D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. d. Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MMM*, 2020.
32. B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," in *AIR*, 2023.
33. D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," in *MICCAI*, 2020.

34. Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” in *MICCAI*, 2021.
35. S. Jain, R. Atale, A. Gupta, U. Mishra, A. Seal, A. Ojha, J. Kuncewicz, and O. Krejcar, “Coinnet: A convolution-involution network with a novel statistical attention for automatic polyp segmentation,” *IEEE Transactions on Medical Imaging*, 2023.
36. Y. Su, Y. Shen, J. Ye, J. He, and J. Cheng, “Revisiting feature propagation and aggregation in polyp segmentation,” in *MICCAI*, Springer, 2023.
37. H. Shao, Y. Zhang, and Q. Hou, “Polyper: Boundary sensitive polyp segmentation,” in *AAAI*, 2024.
38. D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” *arXiv preprint arXiv:1805.10421*, 2018.
39. D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in *ICCV*, 2017.
40. Y. Fang, C. Chen, Y. Yuan, and K.-y. Tong, “Selective feature aggregation network with area-boundary constraints for polyp segmentation,” in *MICCAI*, 2019.