MSWMNet: Enhancing State-Space Models with Multi-Scale Sliding Windows for Polyp Segmentation

1st Yang Meng
School of Computer Science
Shenyang Aerospace University
Shenyang, China

2nd Guoxu Zhang
the People's Hospital of Liaoning Province
Department of Radiology
Shenyang, China

3rd Wei Guo
School of Computer Science
Shenyang Aerospace University
Shenyang, China

4th Zhaoxuan Gong
School of Computer Science
Shenyang Aerospace University
Shenyang, China

5th Guodong Zhang*
School of Computer Science
Shenyang Aerospace University
Shenyang, China

***Abstract*—Accurate segmentation of diminutive polyps in colonoscopy images is crucial for early colorectal cancer screening. However, due to their small scale, low contrast, and blurred boundaries, such polyps are easily obscured by complex backgrounds, and the detail loss introduced by downsampling further exacerbates foreground–background confusion, leading to frequent missed detections. To address these challenges, we propose a novel Mamba-based network, termed MSWMNet (Enhancing State-Space Models with Multi-Scale Sliding Windows for Polyp Segmentation), which adopts a collaborative design that integrates uncertainty-aware local refinement, multi-scale context aggregation, and global structural modeling. Specifically, a Structure-aware Uncertainty Context Refinement (SUCR) module is introduced to enhance the representation of ambiguous regions; a Multi-channel Shifted Window Transformer (MSWinTransformer) module is designed to capture multi-scale local contexts and mitigate small-object feature dilution; and an Enhanced Dual-path Cross Attention (EDCA) module is employed to strengthen cross-level feature interaction and alleviate boundary over-smoothing. In addition, the Mamba module further reinforces global dependency modeling in a computationally efficient manner. Extensive experiments on five public datasets demonstrate the superiority of MSWMNet. On the challenging ETIS-LaribPolypDB dataset, it achieves a Dice score of 75.8%, surpassing previous state-of-the-art methods by up to 5.4%, highlighting its potential for reliable clinical polyp detection.**

***Index Terms*—Colonoscopy Polyp Segmentation, Missed Detection of Small Polyps, Multi-Channel Sliding Window Transformer, Multi-Scale Fusion, Selective Scanning Mechanism.**

I. Introduction

Early methods for colonoscopy polyp segmentation primarily relied on traditional techniques such as edge detection and region growing, whose performance is highly sensitive to image quality and hand-crafted parameter settings. With the advent of deep learning, convolutional neural networks (CNNs), especially U-Net [1], have significantly advanced segmentation performance through a symmetric encoder–decoder architecture. However, such CNN-based methods still exhibit limitations in modeling long-range dependencies and complex lesion morphologies. To address this limitation, numerous variants have been proposed, including Res-UNet [2], Attention U-Net [3], and Dense-UNet [4], which enhance feature representation via residual connections, attention mechanisms, and feature reuse, respectively.

Following the success of Transformers [5] in natural language processing, they have been progressively introduced into computer vision tasks. Transformer-based segmentation models have shown great promise in medical image analysis. For example, TransUNet [6] combines the local representation capability of U-Net with the global context modeling of Transformers, achieving improved performance in segmenting polyps and tumors in complex colonoscopy images. Vision Transformer (ViT) [7] helps alleviate the inherent limitation of CNNs in receptive field for medical image segmentation. Swin Transformer [8] reduces computational complexity through a hierarchical architecture with window-based self-attention and has been adopted in models such as Swin-Unet [9]. Moreover, hybrid architectures such as V-Unet [10] introduce Transformer blocks between the encoder and decoder to enable cross-level or multi-scale global dependency modeling, thereby improving the segmentation of complex lesion regions.

CNN- and Transformer-based segmentation models typically incur high parameter counts and substantial computational and memory costs when modeling long-range dependencies. The Mamba network [11], built on state space models (SSMs) [12], enables linear-complexity long-range modeling by capturing bidirectional dependencies, thereby alleviating the trade-off between global modeling capability and computational efficiency. Vision Mamba [13] further combines Transformers and SSMs in a hybrid

*Corresponding author.

design, maintaining efficient long-range modeling while reducing overall complexity. Recently, several studies have integrated Mamba into U-Net–style architectures for enhanced global context modeling. For example, U-Mamba [3] improves multi-scale feature fusion via inter-layer interactions; SegMamba [14] balances high-resolution representation and memory-efficient multi-scale/volumetric modeling; LKM-UNet [15] strengthens local spatial representation with large-window modeling; T-Mamba [16] introduces frequency-domain features; Mamba-UNet [17] leverages visual state space (VSS) blocks to capture fine-grained structures and global context; and Swin-UMamba [18] replaces attention with VSS blocks while combining shifted windows with a U-shaped design.

Despite recent progress, existing Mamba- and Transformer-based methods for colon polyp segmentation still suffer from insufficient coordination between local detail preservation and global dependency modeling, as well as inadequate multi-scale feature interaction. Consequently, small and low-contrast polyps are prone to feature dilution during hierarchical encoding, while their boundaries are often over-smoothed during decoding, ultimately leading to frequent missed detections. Unlike existing works that typically focus on either global modeling or local enhancement in isolation, we argue that reducing small-polyp missed detection requires a collaborative design that jointly considers uncertainty-aware local refinement, multi-scale context aggregation, and global structural consistency. To this end, we propose MSWMNet, a heterogeneous encoder–decoder network that integrates multi-level feature fusion with complementary local–global modeling, thereby effectively reducing missed detections and improving the segmentation performance of small polyps.

(1) The SUCR module is designed to explicitly model and refine ambiguous regions by aggregating reliable foreground and background information, thereby effectively mitigating missed detections of small and low-contrast polyps.

(2) The MSWinTransformer and Mamba modules are jointly integrated to combine local-detail modeling and global-structure modeling, which preserves small and weak-texture polyp features during deep propagation and multi-channel feature fusion, preventing feature dilution.

(3) The EDCA module is introduced to alleviate insufficient multi-scale feature alignment and boundary over-smoothing, thus enhancing the localization accuracy of small polyps in boundary-ambiguous regions.

The framework demonstrates strong generalization across datasets, and the code will be publicly available at https://github.com/Meng527-maker/MSWMNet

II. Method

A. Overall Network Architecture

As shown in Fig. 1, the network adopts an encoder–decoder architecture. The encoder consists of four stages

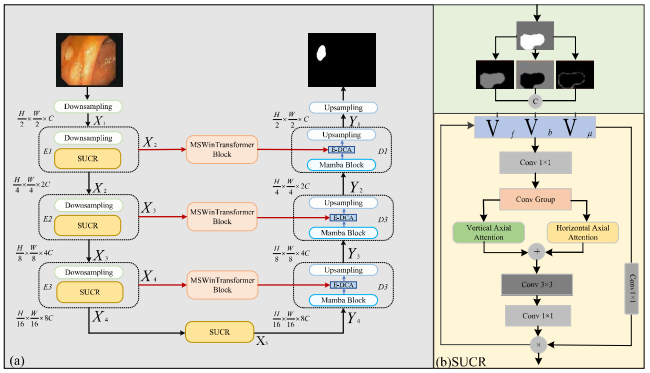

Fig. 1. Illustration of the overall network architecture and detailed SUCR module. (a) Overall structure of MSWMNet. (b) Expanded view of the SUCR module.

(E1–E4), progressively downsampling to extract multi-scale features with resolutions of feature maps of sizes $\frac{H}{2} \times \frac{W}{2}$, $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$ and $\frac{H}{16} \times \frac{W}{16}$, while the number of channels is gradually increased. At each encoding stage, the SUCR module and multi-channel MSWinTransformer module are integrated. In the decoder, feature resolution is progressively restored via upsampling and skip connections, while EDCA modules and Mamba modules are incorporated at each decoding layer for multi-scale fusion and global context modeling.

B. SUCR Module

Given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, the features are partitioned into foreground, background, and uncertain regions by a coarse segmentation, yielding masks m_f, m_b, and m_u. The coarse foreground, background, and uncertainty masks are generated by an auxiliary prediction head attached to the corresponding encoder stage, and are used only for feature refinement during training and inference. To mitigate errors in the uncertain region, the masked features $X_f = X \odot m_f$, $X_b = X \odot m_b$, and $X_u = X \odot m_u$ are first computed. High-confidence information from X_f and X_b is then injected into X_u to produce X'_u, which is subsequently fused with X_f and X_b via a 1×1 convolution to obtain an intermediate feature $\tilde{X} \in \mathbb{R}^{H\times W\times C}$. Next, 1×1 convolutions are applied to generate value features V_f, V_b, and V_u. The fused features are then reorganized through channel compression and grouped convolutions, followed by axial attention along the horizontal and vertical directions. Finally, a residual gating mechanism produces the structure-enhanced feature $Y \in \mathbb{R}^{H\times W\times C}$, which improves global structural consistency while maintaining low computational cost.

C. MSWinTransformer Module

The MSWinTransformer module adopts a dual-branch architecture consisting of an SE-ResNet branch and an MSWMSA branch. The MSWMSA branch serves as the primary feature extractor, while the SE-ResNet

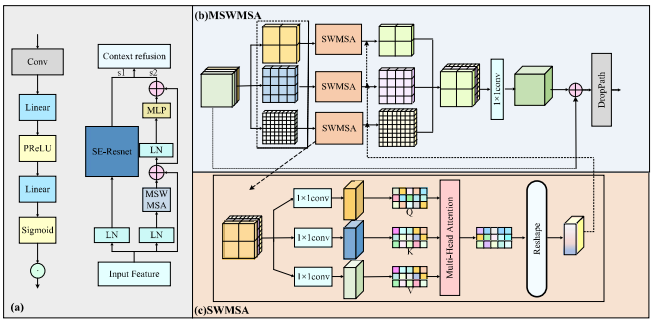

Fig. 2. MSWinTransformer Module Details. (a) Expanded view of the MSWinTransformer structure. (b) Expanded view of the MSWMSA structure. (c) Expanded view of the SWMSA structure.

branch complements it by providing adaptive channel-wise reweighting. The outputs of the two branches are then fused through a context-aware feature fusion strategy. As shown in Fig. 2, the MSWMSA module first splits the input feature map along the channel dimension. Given $X \in \mathbb{R}^{B\times C\times H\times W}$, it is divided into three sub-features $[\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3]$ via channel grouping, where $\mathbf{X}_i \in \mathbb{R}^{B\times \frac{C}{3}\times H\times W}$. Each sub-feature is processed with a different window size $(7\times7, 14\times14, 21\times21)$ to model multi-scale contexts. The small, medium, and large windows capture local details, local context, and global context, respectively.This design is particularly beneficial for preserving the responses of small and low-contrast polyps, whose features are easily overwhelmed during deep encoding. For each $\mathbf{X}_i$, three linear projections are applied to generate the query, key, and value:

$$\mathbf{Q}_i = \mathbf{X}_i\mathbf{W}_Q, \quad \mathbf{K}_i = \mathbf{X}_i\mathbf{W}_K, \quad \mathbf{V}_i = \mathbf{X}_i\mathbf{W}_V \tag{1}$$

Let $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d_k}$ be learnable projection matrices. Sliding-window multi-head self-attention is then applied within each window:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

Since attention is restricted to local windows, each layer only models features at a single scale. The window features are reshaped back and fused by a 1×1 convolution across channel branches to obtain multi-scale representations:

$$\mathbf{F} = \text{Conv}_{1\times1}(\text{Concat}(\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3)) \tag{3}$$

To enable cross-window interaction, a shifted-window mechanism is adopted by shifting the feature map by $M/2$ pixels along both spatial axes before window partition. This allows each window to aggregate features from neighboring windows in the previous layer. The residual update is formulated as:

$$\mathbf{z_l} = \text{MSWMSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1} \tag{4}$$

Each $M\times M$ window performs self-attention within dynamically grouped channels to fuse multi-scale features, achieving linear complexity $4hwC^2 + 2M^2hwC$. Meanwhile, DropPath randomly drops residual branches during training to improve generalization and robustness.

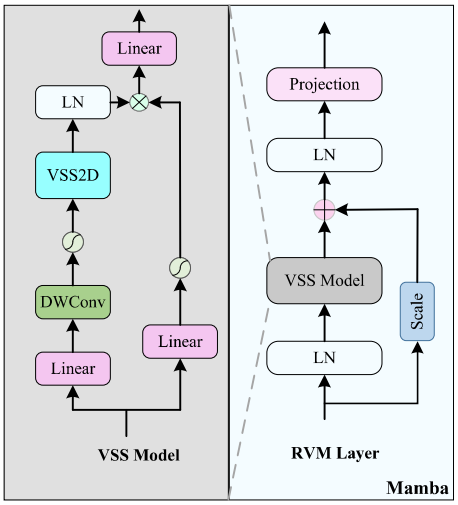

Fig. 3. Mamba Module Architecture

D. Mamba Module

As illustrated in Fig. 3, the Mamba Module consists of a Residual Visual Mixed (RVM) layer and a Visual Selective State-space (VSS) module. The input feature $X \in \mathbb{R}^{N\times C}$ is first normalized and processed by a local convolution before being fed into the VSS module for multi-directional sequence modeling. The features are then mapped back to the 2D space via weighted fusion:

$$Y = \alpha\cdot S_{\text{row}} + \beta\cdot S_{\text{col}} + \gamma\cdot S_{\text{diag}} + \delta\cdot S_{\text{anti}}, \tag{5}$$

where $\alpha, \beta, \gamma, \delta$ are learnable weights. The overall update process of VSS2D is formulated as:

$$Y_t = Y_{t-1} + \text{VSS2D}(\text{LN}(\text{DWConv}(Y_{t-1}))). \tag{6}$$

Its internal recurrent form is formulated as:

$$h_t = A_t h_{t-1} + B_t Z_{\text{local},t}, \quad Y_t = C_t h_t, \tag{7}$$

where A_t, B_t, C_t are input-dependent selective parameters. The output of the state-space branch is first normalized and adaptively fused with the input through a gated linear transformation. Subsequently, a learnable residual scaling is introduced to enhance feature preservation and stabilize training.The fused representation is then normalized and projected to obtain the final output.

$$Y = \text{Projection}\Big(\text{LN}\big(\text{LN}(Z_{\text{ssm}})\odot\text{Linear}(X)+\alpha\cdot X\big)\Big). \tag{8}$$

TABLE I
COMPARISON OF SMALL-SCALE COLONOSCOPY POLYP SEGMENTATION PERFORMANCE OF DIFFERENT METHODS ACROSS FIVE DATASETS.

Method	CVC-300		CVC-clinicDB		CVC-ColonDB		ETIS-LaribPolpyDB		Kvasir	
	mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU
U-Net++ [19]	0.707	0.624	0.794	0.729	0.583	0.490	0.541	0.444	0.821	0.743
ResUNet++ [20]	0.733	0.626	0.779	0.712	0.572	0.481	0.532	0.464	0.791	0.764
Mamba-UNet [17]	0.851	0.785	0.902	0.835	0.777	0.701	0.704	0.632	0.901	0.833
Swin-UMamba [18]	0.883	0.824	0.903	0.816	0.764	0.733	0.672	0.638	0.894	0.791
KMUnet [21]	0.887	0.798	0.901	0.845	0.795	0.734	0.687	0.604	0.902	0.815
H-vmunet [22]	0.894	0.796	0.899	0.788	0.792	0.695	0.696	0.600	0.904	0.813
Ours	**0.905**	**0.799**	**0.926**	**0.817**	**0.814**	**0.738**	**0.758**	**0.654**	**0.911**	**0.869**

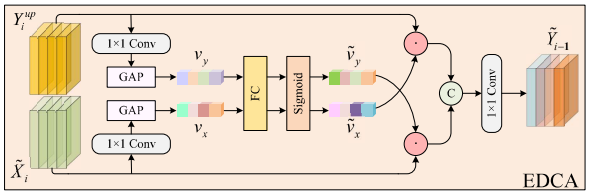

Fig. 4. EDCA Module Architecture

TABLE II
COMPARISON OF MODEL COMPLEXITY AND INFERENCE SPEED

Method	Params (M)	FLOPs (G)	FPS
U-net++	9.1	34.8	158
ResUNet++	11.3	46.1	142
Mamba-Unet	41.8	132.5	49
Swin-UMamba	36.9	124.3	53
KMUnet	27.6	95.4	74
H-vmunet	38.2	120.1	55
Ours	**29.8**	**88.6**	**82**

E. EDCA Module

To alleviate information loss during encoder–decoder fusion, we design a multi-scale fusion module to enhance feature interaction at the same level (Fig. 4). Given Y_i^{up} and X_i, the two features are first aligned and linearly projected. Channel descriptors are obtained by global average pooling:

$$v_y = \text{GAP}(Y_i^{\text{up}}), \quad v_x = \text{GAP}(X_i). \tag{9}$$

They are then passed through FC layers and a Sigmoid function to produce channel weights:

$$v_y = \sigma(\text{FC}(v_y)), \quad v_x = \sigma(\text{FC}(v_x)). \tag{10}$$

We perform bidirectional modulation and extract residual differences to suppress redundancy:

$$d_y = Y_i^{\text{up}} - (v_x \odot X_i), \quad d_x = X_i - (v_y \odot Y_i^{\text{up}}), \tag{11}$$

where $\odot$ denotes channel-wise weighting. The two residual features are concatenated and fused:

$$d = \text{Concat}(d_y, d_x), \tag{12}$$

By enhancing cross-path interaction and boundary discrimination, this design alleviates semantic gaps and reduces missed detections of small polyps.

III. EXPERIMENTS

A. Cross-Dataset Evaluation Protocol

To evaluate the cross-dataset generalization capability of the proposed model, the training set consists of Kvasir [23] and CVC-ClinicDB [24], totaling 1,450 images. The backbone network employs Res2Net50-v1b-26w-4s with pre-trained weights to enhance feature extraction. During the testing phase, the model is independently evaluated on CVC-ColonDB [25] (380 images), CVC-ClinicDB [24] (62 images), ETIS-LaribPolypDB [26] (392 images), Kvasir [23] (100 images), and CVC-300 [27] (120 images). All images are resized to 352×352 to ensure experimental consistency.All experiments in this study were conducted on a workstation equipped with an Intel Core i9-13900K CPU and an NVIDIA GeForce RTX 4090 GPU (24 GB memory) running Ubuntu 22.04. The models were implemented using the PyTorch deep learning framework with CUDA 11.7 (driver version 560.94).

B. Comparative Experiments

The proposed MSWMNet achieves superior segmentation performance on five public colonoscopy datasets (Fig. 5 shows visual comparisons of small polyps). Compared with baseline methods, which often suffer from missed detections, fragmented regions, or over-smoothed predictions, MSWMNet provides more stable and accurate localization and segmentation of small polyps. Even for extremely small, irregularly shaped, or background-similar polyps, it maintains complete coverage and clear boundaries, while producing outputs with higher structural consistency to the ground truth, significantly reducing edge discontinuities and missegmentation.

Table 1 summarizes the quantitative comparison of our method against several representative segmentation models across five public datasets. Traditional CNN-based methods (e.g., U-Net++ and ResUNet++) show limited performance on datasets with a high proportion of small polyps, achieving mDice scores below 60% on both CVC-ColonDB and ETIS-LaribPolypDB. In comparison, advanced models with global modeling capability, such as Mamba-UNet, Swin-UMamba, KMUnet, and H-vmunet, achieve notably improved performance; however, their results still exhibit performance degradation

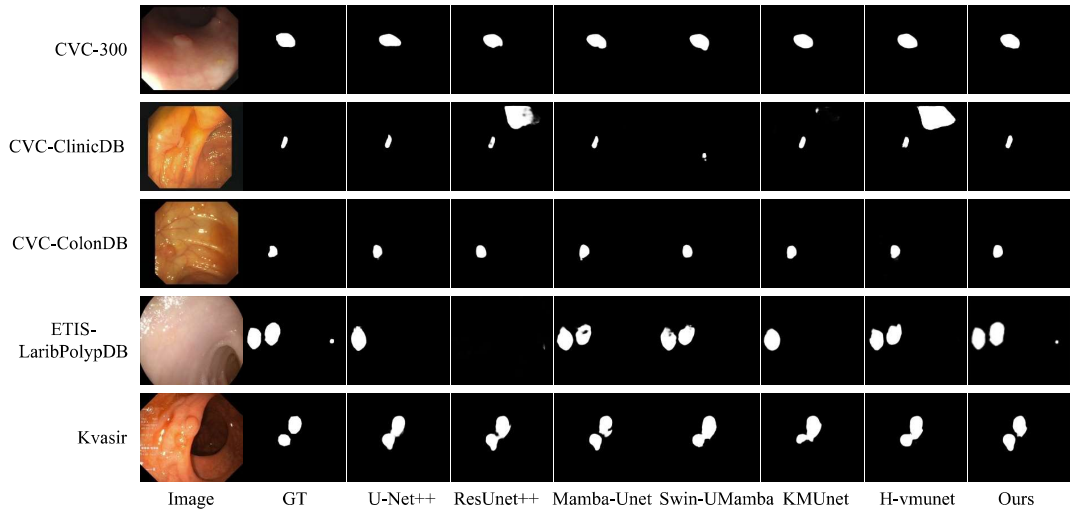

Fig. 5. Qualitative comparison of different methods on small-scale colonoscopy polyp segmentation.

TABLE III
ABLATION EXPERIMENTS SHOWING THE INCREMENTAL EFFECTS OF INTRODUCING SUCR, MSWINTRANSFORMER, MAMBA, AND EDCA MODULES.

Method	CVC-300		CVC-ClinicDB		CVC-ColonDB		ETIS-LaribPolypDB		Kvasir	
	mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU	mDice	mIOU
Base	0.664	0.589	0.792	0.682	0.593	0.476	0.594	0.501	0.737	0.612
Base+SUCR	0.712	0.665	0.823	0.726	0.669	0.543	0.639	0.512	0.752	0.679
Base+SUCR+MSWin	0.843	0.741	0.894	0.765	0.745	0.612	0.654	0.556	0.782	0.728
Base+SUCR+MSWin+Mamba	0.890	0.753	0.901	0.812	0.754	0.675	0.706	0.611	0.891	0.818
Ours (Full Model)	**0.905**	**0.799**	**0.926**	**0.817**	**0.814**	**0.738**	**0.758**	**0.654**	**0.911**	**0.869**

on challenging small-polyp scenarios, with the best competing methods reaching only 79.5% and 70.4% mDice on CVC-ColonDB and ETIS-LaribPolypDB, respectively. Our method achieves the best performance across all test datasets, with particularly notable advantages on datasets containing a high proportion of small polyps. Specifically, on CVC-ColonDB, our method attains an mDice of 81.4%, outperforming the second-best method by approximately 2 percentage points. On ETIS-LaribPolypDB, the mDice reaches 75.8%, exceeding Mamba-UNet by 5.4 percentage points. On larger-scale datasets such as Kvasir and CVC-ClinicDB, our method consistently maintains stable performance advantages, demonstrating strong cross-dataset generalization capability.

Table 2 compares different methods in terms of model complexity and inference efficiency. While U-net++ and ResUNet++ have lower parameter counts and computational costs, their representation capability is limited, whereas Mamba-Unet, Swin-UMamba, and H-vmunet achieve stronger modeling performance at the expense of increased complexity and reduced speed. In contrast, the proposed method attains a higher inference speed (82 FPS) with relatively low computational cost (88.6 G FLOPs) and moderate parameters (29.8 M), achieving a better trade-off between efficiency and performance and demonstrating strong practical potential.

C. Ablation Study

To assess the contribution of each module, starting from the baseline model, we conducted ablation studies on five public datasets by sequentially adding SUCR, MSWinTransformer, Mamba, and EDCA.

Table 3 reports the quantitative results of the ablation studies. The baseline achieves relatively low mDice scores (66.4% on CVC-300 and 59.4% on ETIS), indicating limited capability in handling small and challenging polyps. Adding the SUCR module increases the scores to 71.2% and 63.9%, respectively, validating its effectiveness in enhancing contextual representation for ambiguous and low-contrast regions. Incorporating the MSWinTransformer module further boosts performance (e.g., 84.3% on CVC-300 and 89.4% on CVC-ClinicDB), demonstrating that multi-scale window-based context modeling helps preserve small-object features and suppress background interference. With the Mamba module, the performance is further improved, especially on challenging datasets (e.g., ETIS increases from 65.4% to 70.6%), reflecting the benefit of enhanced global structural modeling. Finally, integrating the EDCA module achieves the best results across all datasets (90.5% on CVC-300, 75.8% on ETIS, and 91.1% on Kvasir), confirming that boundary- and structure-aware feature fusion effectively mitigates boundary over-smoothing. Overall, the modules complement each other

at local, global, and structural levels, jointly improving segmentation performance and reducing missed detections of small polyps.

IV. Conclusion

To address the missed detection of small polyps in colonoscopy images, we propose an end-to-end segmentation network, termed MSWMNet. Instead of relying on a single modeling paradigm, MSWMNet adopts a collaborative design that jointly considers uncertainty-aware local refinement, multi-scale context aggregation, and global structural modeling to strengthen the representation of small and weak-textured targets in complex intestinal scenes. Specifically, the SUCR module leverages uncertainty-guided context modeling to refine ambiguous regions; the MSWinTransformer module employs a multi-channel multi-scale window design to enhance inter-channel feature interactions and preserve small-object responses during deep encoding; the Mamba module further improves global dependency modeling and structural consistency; and the EDCA module facilitates cross-path feature interaction to refine multi-scale alignment and boundary localization. Extensive experiments on multiple public colonoscopy datasets demonstrate that MSWMNet consistently outperforms state-of-the-art methods, especially on challenging scenarios with a high proportion of small polyps. Overall, the proposed network effectively reduces small-polyp miss rates and improves the reliability of computer-aided clinical diagnosis.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *arXiv preprint arXiv:1505.0459*, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[3] O. Oktay, J. Schlemper, L. L. Folgoc, M. J. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[4] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-unet: a novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," *Quantitative imaging in medicine and surgery*, vol. 10, no. 6, p. 1275, 2020.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[6] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *Medical Image Analysis*, 2022.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations (ICLR)*, 2021.

[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002.

[9] 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 Workshops*, 2021.

[10] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *International Conference on 3D Vision (3DV)*, 2016.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[12] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024, pp. 578–588.

[13] 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.

[14] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024.

[15] J. Wang, J. Chen, D.-Z. Chen, and J. Wu, "Lkm-unet: Large kernel vision mamba unet for medical image segmentation," *arXiv preprint*, 2024.

[16] J. Hao, Y. Zhu, L. He, M. Liu, J. K. Tsoi, and K. F. Hung, "T-mamba: A unified framework with long-range dependency in dual-domain for 2d and 3d tooth segmentation," *arXiv preprint*, 2024.

[17] Z. Wang, J. 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.

[18] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, C. Li, Y. Liang, G. Shi, Y. Yu, and S. Zhang, "Swin-umamba: Mamba-based unet with imagenet-based pretraining," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024, pp. 615–625.

[19] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *DLMIA and ML-CDS Workshops*, 2018, pp. 3–11.

[20] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. de Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *IEEE International Symposium on Multimedia (ISM)*, 2019.

[21] K. Wang, M. Zhang, Y. Liu, and L. Chen, "Kmunet: Knowledge-guided multi-scale u-net for medical image segmentation," *arXiv preprint arXiv:2404.05678*, 2024.

[22] Q. Zhao, H. Wang, S. Liu, and W. Zhang, "H-vmunet: Hybrid vision mamba u-net for medical image segmentation," *arXiv preprint arXiv:2402.16789*, 2024.

[23] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and S. A. Pettersen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*, 2020, pp. 451–462.

[24] J. Bernal, N. Tajbakhsh, F. Sánchez, B. Matuszewski, H. Chen, L. Yu, and Q. Angermann, "Wm-dova maps for accurate polyp highlighting in colonoscopy," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[25] J. Bernal, F. Sánchez, and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," in *International Conference on Pattern Recognition (ICPR)*, 2012, pp. 1172–1175.

[26] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward automatic polyp detection with a polyp appearance model," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 2309–2313.

[27] S. F. Sánchez-Peralta, J. B. Pagador, A. Picón, A. Calderón, F. Polo, L. J. Herrera, and J. M. Pagador, "Polyp detection with instance segmentation in colonoscopy images," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2020, pp. 574–584.