DM-VMUNet: Dual-Context Perception and Multi-Scale Feature Alignment in Vision Mamba for Medical Image Segmentation

1st Tian Zhou
School of Computer Science and Technology, Xinjiang University
Urumqi, China
19103893896@163.com

2nd Yehang Li
School of Computer Science and Technology, Xinjiang University
Urumqi, China
107556525382@stu.xju.edu.cn

3rd Hui Zhao*
School of Computer Science and Technology, Xinjiang University
Urumqi, China
zhaohui@xju.edu.cn

***Abstract*—Accurate segmentation of colorectal polyps and skin lesions is critical for clinical diagnosis. However, morphological variability and blurred lesion boundaries pose significant challenges to automated segmentation. Although the Mamba architecture excels in long-range modeling, its inherent 1D scanning mechanism often compromises 2D local geometric fidelity. Furthermore, conventional skip connections fail to adequately address semantic misalignment across different feature levels. To overcome these limitations, we propose DM-VMUNet, a hybrid architecture designed to integrate global context with local precision. Specifically, we introduce a Dual-Context Large-Small Convolution (DCLSConv) module. By leveraging multi-scale large-kernel perception and dynamic small-kernel aggregation, DCLSConv is designed to decouple receptive field expansion from computational cost, facilitating adaptive sharpening of lesion boundaries.Additionally, we propose a Multi-Scale Feature Alignment (MFA) module to semantically align encoder-decoder features prior to fusion. Extensive experiments on the Kvasir-SEG, ISIC 2017, and ISIC 2018 datasets demonstrate that DM-VMUNet achieves Dice coefficients of 90.75%, 89.87%, and 90.25%, respectively.These results validate the effectiveness of our approach, particularly its robust capability in delineating complex lesion boundaries compared to representative methods.**

***Index Terms*—Medical Image Segmentation, Deep Learning, Vision Mamba, feature fusion**

I. Introduction

Medical image segmentation is a cornerstone of computer-aided diagnosis systems, essential for clinical tasks such as colorectal polyp detection and skin lesion analysis [1] [2] [3]. However, significant variability in lesion morphology and size, coupled with blurred boundaries, challenges automated segmentation. While CNNs [4]offer strong local perception, they lack global receptive fields; conversely, Vision Transformers (ViTs) [5] [6]excel at global modeling but often incur prohibitive computational costs. The Mamba architecture, distinguished by global modeling capabilities with linear complexity, has emerged as a compelling paradigm for balancing performance and efficiency [7] [8] [9]. Despite the success of Mamba-based methods in modeling long-range dependencies, two key limitations persist in handling medical lesions with complex geometries and ambiguous boundaries. First, an inherent conflict exists between 1D sequence modeling and the preservation of 2D local geometric features. Existing VM-UNet-like methods [10] typically flatten 2D images into 1D sequences for scanning. While this effectively captures global context, it inevitably disrupts the 2D spatial continuity of local regions. Although current Mamba models can localize polyp regions [11] they often fail to capture jagged structures or minute protrusions due to insufficient fine-grained texture analysis, yielding over-smoothed predictions that lack critical diagnostic details.

Second, semantic misalignment remains a critical bottleneck in classic U-shaped architectures. Skip connections typically concatenate high-resolution features directly from the encoder with upsampled features from the decoder. However, as shallow features often contain substantial background noise dominated by low-level textural information, indiscriminate fusion without alignment mechanisms introduces noise into the decoder. This causes feature space confusion and ambiguity during spatial detail recovery, limiting the model's ability to localize small lesions [12].

To address these challenges, we introduce DM-VMUNet. Building upon the powerful global modeling capability of Mamba, we aim to enhance geometric fidelity for complex boundaries and optimize cross-level fusion quality by introducing explicit multi-scale spatial perception and feature alignment mechanisms. Our main contributions are as follows:

- To mitigate Mamba's limited local perception, we propose the Dual-Context Large-Small Convolution. Inspired by the human visual system's mechanism [13], this module integrates multi-scale large-kernel depth-wise convolution with dynamic weighted small-kernel aggregation.It compensates for the lack of 2D local inductive bias in State Space Models, significantly enhancing geometric fidelity at complex lesion boundaries.
- To resolve semantic misalignment in skip connections, we developed the Multi-Scale Feature Alignment module.

*Corresponding author.

This article is supported by the Key Research and Development Program of Xinjiang Uygur Autonomous Region (Project No. 2023B01032)

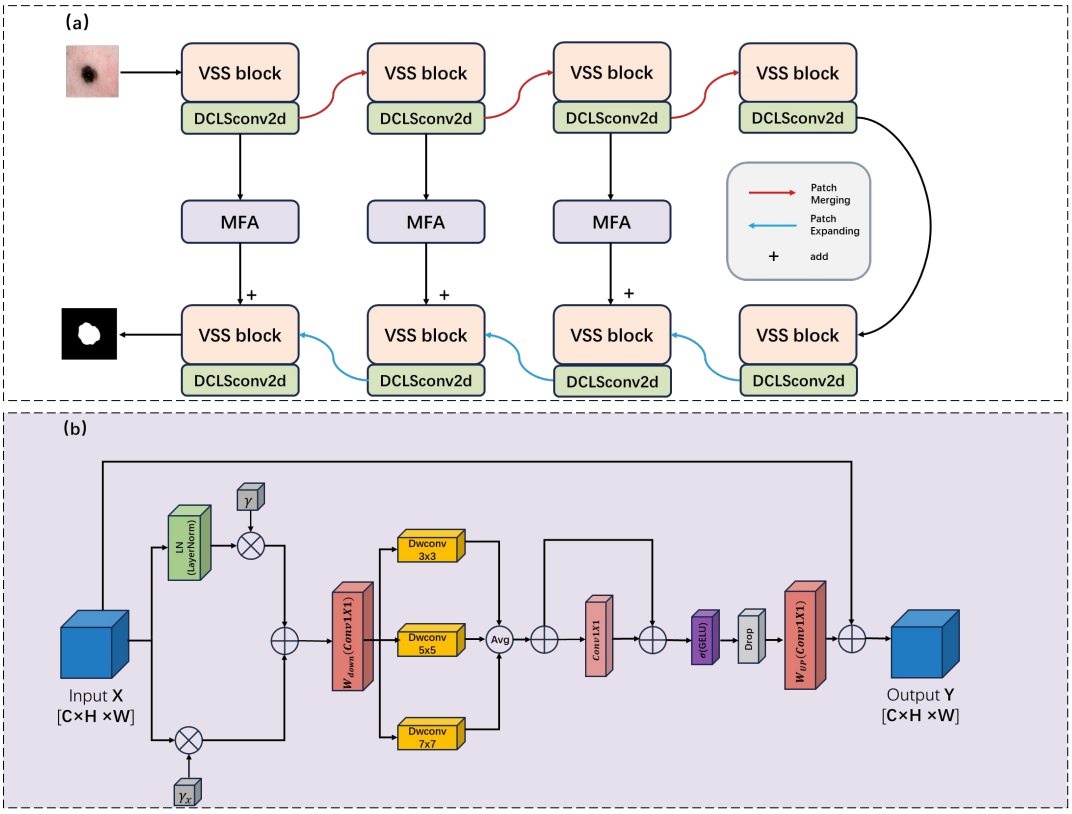

Fig. 1. a. The overall architecture of DM-VMUNet. b. The detailed structure of the MFA module.

Following an "align-then-fuse" philosophy, MFA introduces adaptive affine transformations and multi-granular spatial perception prior to fusion. Functioning as a semantic bridge, MFA dynamically calibrates encoder features to accommodate significant target scale variations in the ISIC and Kvasir datasets, ensuring robust segmentation.

- DM-VMUNet achieves state-of-the-art performance across multiple medical image segmentation benchmarks. By synergizing the linear global modeling advantage of Mamba with the local geometric refinement of DCLSConv and the feature alignment mechanism of MFA, our model demonstrates superior accuracy and robustness.

II. Related Work

A. CNNs and Vision Transformers

For years, U-Net and its variants [14] [15]have served as the backbone of medical image segmentation, leveraging encoder-decoder architectures with skip connections. Subsequently, to overcome the restricted receptive fields of CNNs, Vision Transformers were introduced. Approaches such as Swin-UNet [16] utilize self-attention mechanisms to capture long-range dependencies. However, while pure CNN models are limited by insufficient global context awareness, ViTs are hindered by quadratic computational complexity and the potential loss of fine-grained details in tiny lesions, creating a difficult trade-off between efficiency and accuracy.

B. Vision Mamba for Medical Image Segmentation

To circumvent the computational bottlenecks of Transformers, the Mamba architecture has gained prominence for its linear-complexity global modeling capabilities. VM-UNet pioneered the integration of Mamba modules into U-shaped networks, utilizing the 2D Selective Scan (SS2D) mechanism to achieve global context awareness while maintaining low computational costs. Subsequent works, such as Swin-Umamba [17] and H-Vmunet [18], have further investigated its application across various medical tasks. However, existing Mamba-based methods typically rely on flattening 2D images into 1D sequences. This dimensionality reduction inherently disrupts the intrinsic 2D spatial continuity of anatomical structures. Consequently, when processing complex geometric shapes or low-contrast edges, these models often exhibit a loss of local geometric fidelity. Furthermore, most approaches directly adopt traditional simple skip connections, neglecting the potential semantic misalignment between Mamba encoder features and decoder features.

C. Large-Kernel Convolutions and Feature Fusion

To enhance the perception of local details, large-kernel convolutions [19] [20]have gained prominence in recent years.

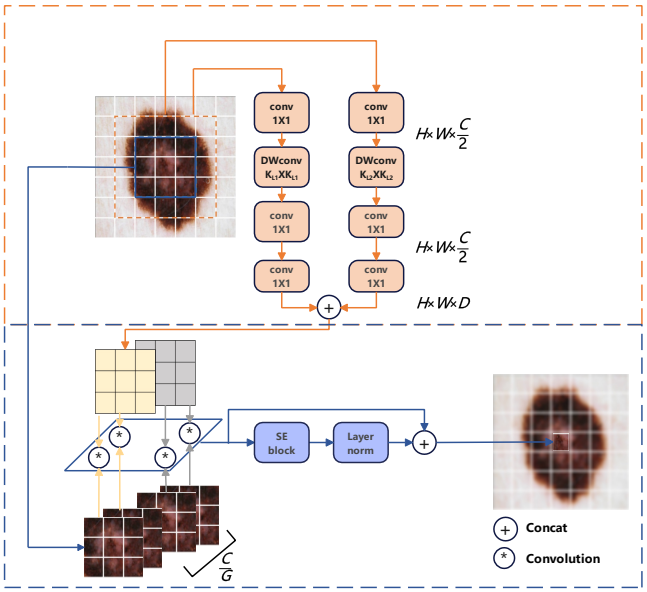

Fig. 2. Detailed architecture of the DCLSConv module, consisting of three primary components: multi-scale large-kernel perception, dynamic small-kernel aggregation, and channel attention mechanism.

LSNet [13] introduced a "large-kernel perception, small-kernel aggregation" paradigm to efficiently balance receptive field expansion and computational cost. However, the original LSNet relies on a fixed large kernel size, limiting its adaptability to the significant scale variations typical of medical lesions. In terms of feature fusion, methods like Attention U-Net [21] and UNet 3+ [22] have optimized skip connections. However, these improvements mostly focus on feature selection or dense connectivity, often overlooking the critical need for "semantic alignment" prior to fusion. This misalignment is particularly acute in Mamba architectures, where the spatial disparity between the encoder's 1D scanned features and the decoder's upsampled features exacerbates semantic ambiguity during direct fusion.

III. Methodology

A. Architecture Overview

The proposed DM-VMUNet introduces a Dual-Context Large-Small Convolution (DCLS) mechanism and a Multi-scale Feature Adaptive (MFA) module to enhance local modeling capabilities and cross-level feature fusion. The overall architecture, illustrated in Fig. 1(a), adopts a U-shaped structure comprising an encoder, a decoder, and skip connections, with a total of 4 stages. The channel dimensions for these four stages are set to [96, 192, 384, 768]. As shown in Fig. 1 (b), DM-VMUNet utilizes MFA in the skip connections to achieve cross-level semantic alignment and improve the quality of feature fusion.

B. Dual-Context Large-Small Convolution (DCLS)

In existing networks, convolution and self-attention typically undertake the roles of local and global modeling, respectively. However, small-kernel convolutions with fixed weights struggle to adapt to complex contexts, while global attention incurs significant computational overhead. Inspired by the "See Large, Focus Small" mechanism of the human visual system, LSNet proposed the Large-Small (LS) convolution, which decouples large-kernel perception from small-kernel dynamic aggregation to balance receptive field and efficiency. Building upon this, we design the DCLSConv specifically to address the large scale variations inherent in medical images. unlike the original LSConv, which employs a single-sized large kernel, DCLSConv incorporates a multi-scale large-kernel perception module and introduces channel attention for feature recalibration. This module decouples long-range spatial context capture from local fine-grained feature fusion. The overall workflow, comprising three components—Parallel Multi-Scale Large Kernel Perception (MultiScale LKP), Small Kernel Aggregation (SKA), and Channel Attention—is illustrated in Fig. 2. By leveraging large-view perception to collect comprehensive context and model relationships, and small-view aggregation for efficient fusion in highly correlated environments, DCLSConv facilitates detailed visual representation.

C. Multi-Scale Large Kernel Perception (MultiScale LKP)

To accommodate the significant scale variations in medical images, we employ multi-path parallel large-kernel depth-wise convolutions to extract contextual information. Let the input feature map be $X \in \mathbb{R}^{C\times H\times W}$, where $H \times W$ denotes the spatial resolution and C is the number of channels. We first utilize a point-wise convolution (PW) to project the tokens to $D = C/2$ to reduce computational costs and maintain model lightweightness. For the compressed features x_i, we adopt dual-context multi-scale large kernels, specifically depth-wise (DW) convolutions with kernel sizes $K_{Lm}, m \in \{1, 2\}$, to efficiently capture large-range spatial context information $N_{KLm}(x_i)$. Here, N_{KLm} represents the peripheral region of size K_{Lm} centered at x_i, and P denotes perception, which includes extracting contextual information and capturing relationships between tokens. Large-kernel depth-wise convolutions effectively expand the receptive field and enhance context awareness at a minimal cost. Subsequently, we use point-wise convolution to model the spatial relationships between tokens, generating context-adaptive weights $W_{im} \in \mathbb{R}^{H\times W\times D}$ for the aggregation step. Finally, these are concatenated to obtain W_i. The entire process can be formulated as:

$$W_{im} = \mathrm{PW}(\mathrm{DW}_{K_{Lm}\times K_{Lm}}(\mathrm{PW}(x_i))) \tag{1}$$

$$W_i = \mathrm{PW}(\mathrm{Concat}(W_{i1}, W_{i2})) \tag{2}$$

D. Small Kernel Aggregation (SKA)

After capturing extensive spatial context and generating adaptive weights W_i via the MultiScale LKP module, we utilize the Small Kernel Aggregation (SKA) mechanism to inject these global priors into local fine-grained features. SKA aims to dynamically guide the fusion of local features using the generated weights, thereby achieving "small-view focus under wide-view guidance."

Specifically, to reduce computational complexity, we split the feature map X along the channel dimension into G groups (set to $G = 8$ in our experiments), with each group containing C/G channels. Let x_{ic} represent the feature map of the c-th sub-channel belonging to the i-th group. For the local neighborhood N_{Ks} at each position, we perform dynamic convolutional aggregation using the corresponding large-kernel perception weight W_i. This process no longer relies on static convolution kernels but adaptively adjusts the aggregation intensity based on the image content. The process is formally defined as:

$$y_{ic} = \mathcal{A}(W_i, N_{Ks}(x_{ic})) = W_i \circledast N_{Ks}(x_{ic}) \tag{3}$$

where $\mathcal{A}$ denotes aggregation and $\circledast$ represents the dynamic convolution operation. In this manner, the model can precisely recover minute lesion details while suppressing irrelevant background noise.

Subsequently, we concatenate the locally aggregated features y_{ic} from all channels along the channel dimension to reconstruct the complete feature tensor Y_{agg}:

$$Y_{agg} = \text{Concat}(y_{i1}, y_{i2}, \ldots, y_{ic}) \tag{4}$$

To further enhance the model's sensitivity to critical anatomical structures, we introduce a Channel Attention (SE) mechanism for feature recalibration on top of the SKA spatial aggregation. Finally, the output of the DCLSConv module is given by a residual connection:

$$\text{Out} = X + \text{BN}(\text{SE}(Y_{agg})) \tag{5}$$

E. Multi-Scale Feature Alignment (MFA) in Skip Connections

In the standard U-Net architecture, simple skip connections typically concatenate shallow low-level features directly with deep high-level features. However, a significant semantic discrepancy exists between these two representations. To address this feature misalignment issue and enhance the model's ability to capture multi-scale lesions, we propose the Multi-Scale Feature Alignment module.

MFA is designed as a residual block with a bottleneck structure. Given input features $X \in \mathbb{R}^{C\times H\times W}$, the computational flow of this module can be formalized into the following four stages:

First, to mitigate the distribution variance between encoder and decoder features, we introduce an enhanced affine transformation mechanism. Unlike standard LayerNorm, we introduce an additional learnable parameter γ_x to preserve the identity information of the original features:

$$X_n = \text{LN}(X) \odot \gamma + X \odot \gamma_x \tag{6}$$

where LN denotes Layer Normalization, $\odot$ represents element-wise multiplication, and γ and γ_x are learnable scaling vectors.

To capture rich spatial context at a low computational cost, we map the features to a lower-dimensional space via a projection matrix W_{down}, followed by multi-branch depth-wise convolutions to extract features with different receptive fields in parallel:

$$X_{agg} = W_{down}(X_n) + \sum_{k\in\{3,5,7\}} \text{DW}_k(W_{down}(X_n)) \tag{7}$$

Here, we employ a set of convolutions with kernel sizes $k \in \{3, 5, 7\}$. Note that a first residual connection is introduced here to facilitate gradient propagation across the bottleneck layer.

Following multi-scale spatial aggregation, we further introduce inter-channel information interaction via point-wise convolution:

$$X_{out} = X_{agg} + \text{Conv}_{1\times1}(X_{agg}) \tag{8}$$

Finally, the aggregated features undergo non-linear activation and regularization, are restored to the original channel dimension via an up-projection W_{up}, and are fused with the original input through a global residual connection to form the final output Y:

$$Y = X + W_{up}(\text{Dropout}(\sigma(X_{out}))) \tag{9}$$

where σ denotes the GELU activation function.

IV. Experiments

A. Datasets

To evaluate the efficacy of DM-VMUNet, we conducted extensive experiments on three public medical image segmentation benchmarks: one colorectal polyp dataset and two skin lesion datasets.

- **Kvasir-SEG Dataset:** This dataset comprises 1,000 high-resolution polyp images obtained from real colonoscopy examinations [11], paired with pixel-level annotation masks. The dataset encompasses polyps exhibiting diverse sizes, morphologies, and textures. We randomly partitioned the dataset into training (700 images) and testing (300 images) sets to assess model robustness.
- **ISIC Datasets:** To evaluate generalization capability on skin lesion segmentation, we employed two public datasets from the International Skin Imaging Collaboration challenge: ISIC 2017 [23] and ISIC 2018 [24]. These datasets contain 2,150 and 2,694 dermoscopy images, respectively, with pixel-level gold standard annotations. To ensure fair comparison, we adhered strictly to the data splitting scheme from VM-UNet, utilizing identical training and testing samples.

B. Implementation Details

All models were implemented in PyTorch and executed on a single NVIDIA GeForce RTX 4090 GPU for both training and inference. To manage GPU memory constraints and improve computational efficiency, input images and masks from all datasets were uniformly resized to 256×256 resolution. To bolster generalization and prevent overfitting, we employed online data augmentation strategies during training, including random horizontal/vertical flipping and random rotation.

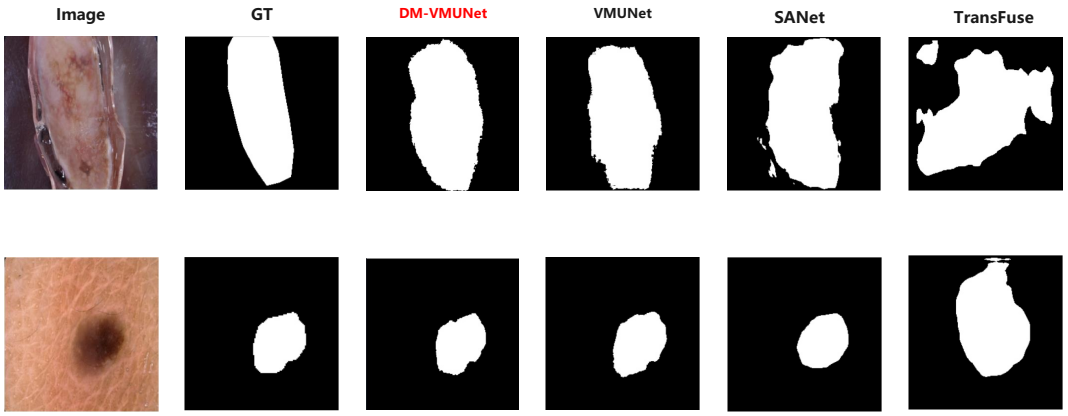

Fig. 3. Visual Comparisons on the ISIC18 Dataset. GT stands for ground truth.

We utilized the AdamW optimizer with an initial learning rate of 1×10^{-3}, a weight decay of 1×10^{-2}, and default momentum parameters. Models were trained for 300 epochs with a batch size of 32. Addressing class imbalance, we employed a composite loss function combining Binary Cross-Entropy (BCE) Loss and Dice Loss to simultaneously optimize pixel-level classification accuracy and regional overlap.

C. Evaluation Metrics

We utilized Mean Intersection over Union (mIoU), Dice Similarity Coefficient (DSC), Specificity (Spe), and Sensitivity (Sen) for quantitative evaluation. mIoU measures the ratio of intersection to union between the predicted region and ground truth. The Dice coefficient (DSC) quantifies segmentation accuracy by measuring the overlap between the prediction and ground truth. Accuracy denotes the proportion of correctly predicted pixels, while Specificity measures the model's ability to identify negative examples.

D. Comparative Experiments

To demonstrate the efficacy of our proposed model, we compared it against several widely adopted and state-of-the-art medical image segmentation models.

- **Kvasir-SEG Results:** As detailed in Table I, our model achieved superior performance in both mIoU and DSC metrics, reaching 82.73% and 90.75%, respectively.

TABLE I
QUANTITATIVE COMPARISON ON THE KVASIR-SEG DATASET

Methods	mIoU (%)	DSC (%)	Spe (%)	Sen (%)
U-Net [4]	65.50	75.80	85.10	83.60
UNet++ [14]	73.90	81.50	80.20	86.50
SwinUnet [16]	78.92	84.11	93.30	91.10
AttnUNet [21]	67.60	77.40	89.50	83.90
VM-UNet [10]	80.32	89.09	98.49	87.21
XNet [25]	80.80	85.80	86.90	**92.41**
SegDINO [26]	80.64	87.65	97.48	88.95
DM-VMUNet	**82.73**	**90.75**	**98.54**	91.47

- **ISIC Results:** Table II and Table III present comparative results for skin lesion datasets. Our method achieved DSC scores of 90.25% on ISIC 2018 and 89.87% on ISIC 2017, surpassing state-of-the-art competitors.

TABLE II
QUANTITATIVE COMPARISON ON THE ISIC 2018 DATASET

Methods	mIoU (%)	DSC (%)	Spe (%)	Sen (%)
U-Net [4]	77.86	87.55	96.69	85.86
UNet++ [14]	78.31	87.83	95.75	88.65
Att-UNet [21]	78.43	87.91	96.23	87.60
UTNetV2 [27]	78.97	88.25	96.48	87.60
SANet [28]	79.52	88.59	95.97	89.46
TransFuse [29]	80.63	89.27	95.74	91.28
MALUNet [30]	80.25	89.04	96.19	89.74
VM-UNet [10]	81.35	89.71	96.13	91.12
VMUNetV2 [31]	81.37	89.73	**97.67**	**91.89**
DM-VMUNet	**81.89**	**90.25**	96.33	91.23

TABLE III
QUANTITATIVE COMPARISON ON THE ISIC 2017 DATASET

Methods	mIoU (%)	DSC (%)	Spe (%)	Sen (%)
U-Net [4]	76.98	86.99	97.43	86.82
UTNetV2 [27]	77.35	87.23	98.05	84.85
TransFuse [29]	79.21	88.40	97.98	87.14
MALUNet [30]	78.78	88.13	**98.47**	84.78
VM-UNet [10]	80.23	89.03	97.58	**89.90**
DM-VMUNet	**81.47**	**89.87**	98.32	88.09

We evaluated the statistical significance of our model in comparison with VMUNet using the Wilcoxon signed-rank test. The mIoU improvements are statistically significant on three datasets, with p-values of 0.0328 on Kvasir-SEG, 0.0473 on ISIC 2018 and 0.0487 on ISIC 2017, respectively.

E. Ablation Studies

To validate our contributions, we conducted ablation experiments on the Kvasir-SEG and ISIC 2017 datasets.The quantitative results are detailed in Table IV and Table V. As

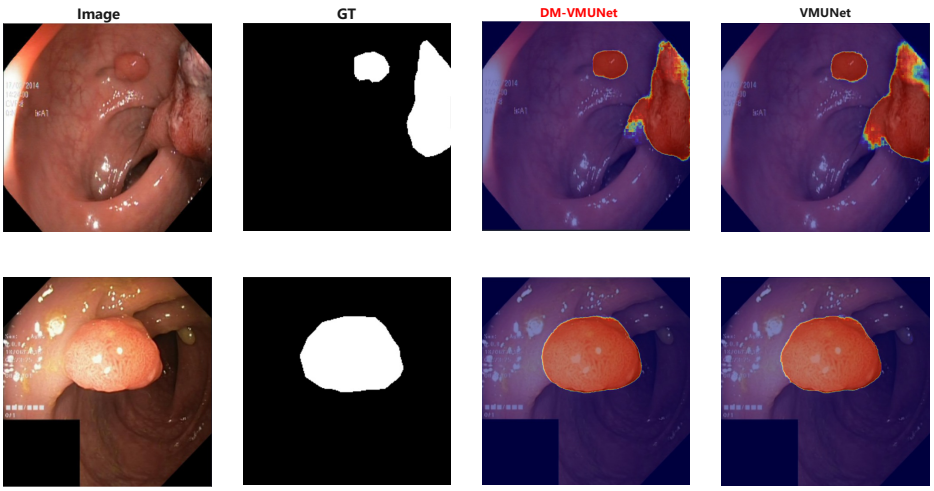

Fig. 4. Visual comparisons of colorectal polyp segmentation and corresponding attention map overlays on the Kvasir-SEG dataset.GT stands for ground truth.

presented in the ablation results, incorporating DCLSConv significantly boosts performance, suggesting that this module effectively mitigates the Mamba architecture's limitations in 2D local spatial perception and contributes substantially to fine-grained lesion boundary segmentation. Similarly, the isolated introduction of the MFA module yields performance gains, validating its effectiveness in alleviating semantic misalignment and adapting to multi-scale lesions. The full DM-VMUNet architecture outperforms configurations with individual modules across all metrics. This confirms the synergistic effect of DCLSConv and MFA, achieving an optimal balance between local detail refinement and cross-level feature fusion.

TABLE IV
ABLATION STUDY ON THE KVASIR-SEG DATASET

Methods	mIoU (%)	DSC (%)	Spe (%)	Sen (%)
VM-UNet + DCLS	81.73	89.95	98.14	89.54
VM-UNet + MFA	81.47	89.79	98.29	88.63
DM-VMUNet (Ours)	**82.73**	**90.75**	**98.54**	**91.47**

TABLE V
ABLATION STUDY ON THE ISIC 2017 DATASET

Methods	mIoU (%)	DSC (%)	Spe (%)	Sen (%)
VM-UNet + DCLS	81.10	89.56	98.24	88.02
VM-UNet + MFA	80.45	89.16	97.29	86.83
DM-VMUNet (Ours)	**81.47**	**89.87**	**98.32**	**88.09**

Furthermore, in order to investigate the impact of different kernel sizes in the DCLSConv module and provide a theoretical rationale for our selection, we conduct an ablation study comparing three configurations: a symmetric small kernel (7×7), an asymmetric kernel (7×11), and a symmetric large kernel (11×11).Notably, to isolate the specific influence of kernel dimensions, the Multi-scale Feature Adaptive Fusion (MFA) module is excluded from this particular evaluation. As shown in Table VI, the 7×7 setting achieves standard baseline performance but struggles to fully capture irregularly shaped or elongated lesions due to a limited receptive field. Conversely, scaling the kernel up to 11×11 incurs a significantly higher computational overhead without bringing further performance gains. This performance saturation is likely due to the incorporation of irrelevant background noise when the receptive field becomes excessively large.

TABLE VI
ABLATION STUDY ON THE KERNEL SIZE OF DCLSCONV ON THE ISIC 2018 DATASET

Kernel Size	mIoU (%)	DSC (%)	Spe (%)	Sen (%)
7×7	81.51	89.81	96.18	90.93
7×11 **(Ours)**	**81.66**	**89.91**	**96.21**	**91.30**
11×11	81.36	89.73	96.15	90.23

The asymmetric configuration (7×11) yields the best trade-off, achieving the highest Dice and mIoU scores. This indicates that asymmetric receptive fields are inherently better suited for capturing the anisotropic and elongated morphological features typical of medical lesions, while simultaneously maintaining high computational efficiency.

F. Model Complexity and Efficiency Analysis

To further evaluate the practical utility of DM-VMUNet, we analyze its computational efficiency in terms of Params and FLOPs compared to VM-UNet.VM-UNet involves 27.43M parameters and 4.11G FLOPs. In comparison, our DM-VMUNet,which is configured without the MFA module here to isolate the impact of DCLSConv,introduces only a marginal increase, with 30.37 M parameters and 5.3G FLOPs.Although these additions lead to a modest rise in computational overhead, the corresponding improvements in segmentation performance are substantial. This favorable trade-off underscores

the superior efficiency of the DCLSConv module, which effectively balances model complexity and representative power.

G. Qualitative Analysis

Complementing the quantitative evaluation, we provide qualitative analysis to visually demonstrate the performance advantages of DM-VMUNet. Fig. 3 illustrates a visual comparison of skin lesion segmentation on the ISIC 2018 dataset. Observations indicate that in cases with ambiguous boundaries or complex morphologies, segmentation masks generated by DM-VMUNet demonstrate superior alignment with the Ground Truth (GT), effectively avoiding over-smoothing or under-segmentation issues common in other models.

Furthermore, Fig. 4 presents segmentation heatmaps and results for colorectal polyps on the Kvasir-SEG dataset. The heatmaps visually highlight the model's attentional focus on specific regions. Specifically, compared to baseline approaches, the heatmaps of DM-VMUNet exhibit highly concentrated and continuous gradient activations precisely along the topological edges of the polyps, with significantly less attention dispersed into the surrounding healthy tissue. This distinct thermal pattern directly reflects the internal mechanism of the DCLSConv module: the multi-scale large-kernel branches establish a broad structural context to locate the primary lesion area, while the dynamic small-kernel aggregation concentrates on extracting high-frequency edge information. By doing so, DCLSConv adaptively sharpens the focus on boundary transitions, validating its capacity to capture local geometric details. Meanwhile, the MFA module ensures accurate segmentation when handling polyps with significant scale variations. This visual evidence further corroborates the effectiveness of our proposed method.

V. Conclusion

In this paper, we introduced DM-VMUNet, a medical image segmentation framework optimized for complex geometric morphologies and ambiguous boundaries. This framework resolves the inherent limitations of Mamba-based 1D scanning in preserving 2D local geometric fidelity and achieving effective cross-level fusion. Specifically, we integrated the DCLSConv module, which explicitly aggregates broad context via multi-scale large kernels and injects global priors into local details through dynamic small-kernel aggregation, thereby enhancing boundary geometric fidelity. Concurrently, we developed the MFA module to perform semantic alignment and structural information screening within skip connections prior to fusion, effectively mitigating semantic misalignment between the encoder and decoder. Extensive experiments demonstrate that our method achieves consistent performance gains across benchmarks such as Kvasir-SEG and ISIC 2017/2018. Future work will focus on incorporating boundary-aware losses or uncertainty modeling to further improve prediction accuracy for non-uniform boundaries. Additionally, extending our framework to 3D volumetric data and cross-domain scenarios represents a promising direction for further research.

References

[1] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET image processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[2] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical image analysis*, vol. 63, p. 101693, 2020.

[3] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-PVT: Polyp segmentation with pyramid vision transformers," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[5] 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," *arXiv preprint arXiv:2102.04306*, 2021.

[6] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[7] Z. Xing, T. Ye, Y. Yang, D. Cai, B. Gai, X.-J. Wu, F. Gao, and L. Zhu, "Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation," *IEEE Transactions on Medical Imaging*, 2025.

[8] A. Chang, J. Zeng, R. Huang, and D. Ni, "Em-net: Efficient channel and frequency learning with mamba for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 266–275.

[9] R. Wu, Y. Liu, and P. Liang, "UltraLight VM-UNet: Parallel vision mamba significantly reduces parameters for medical image segmentation," *arXiv preprint arXiv:2403.20035*, 2024.

[10] J. Ruan, J. Li, and S. Xiang, "Vm-unet: Vision mamba unet for medical image segmentation," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[11] D. Jha, P. Smedsrud, M. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. Johansen, "Kvasir-seg: A segmented polyp dataset. multimedia modeling. mmm 2020," *Lecture Notes in Computer Science*, vol. 11962, 2019.

[12] A. Lou, S. Guan, H. Ko, and M. H. Loew, "CaraNet: Context axial reverse attention for segmentation of small medical objects," in *Medical Imaging 2022: Image Processing*, vol. 12032. SPIE, 2022, pp. 81–92.

[13] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "Lsnet: See large, focus small," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9718–9729.

[14] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *International workshop on deep learning in medical image analysis*. Springer, 2018, pp. 3–11.

[15] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," in *2018 9th international conference on information technology in medicine and education (ITME)*. IEEE, 2018, pp. 327–331.

[16] 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 *European conference on computer vision*. Springer, 2022, pp. 205–218.

[17] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.

[18] R. Wu, Y. Liu, P. Liang, and Q. Chang, "H-vmunet: High-order vision mamba unet for medical image segmentation," *Neurocomputing*, vol. 624, p. 129447, 2025.

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[20] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 963–11 975.

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

[22] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. Ieee, 2020, pp. 1055–1059.

[23] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[24] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," *arXiv preprint arXiv:1902.03368*, 2019.

[25] Y. Zhou, J. Huang, C. Wang, L. Song, and G. Yang, "Xnet: Wavelet-based low and high frequency fusion networks for fully-and semi-supervised semantic segmentation of biomedical images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 21 085–21 096.

[26] S. Yang, H. Wang, Z. Xing, S. Chen, and L. Zhu, "Segdino: An efficient design for medical and natural image segmentation with dino-v3," *arXiv preprint arXiv:2509.00833*, 2025.

[27] Y. Gao, M. Zhou, D. Liu, Z. Yan, S. Zhang, and D. N. Metaxas, "A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark," *arXiv preprint arXiv:2203.00131*, 2022.

[28] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "Shallow attention network for polyp segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 699–708.

[29] Y. Zhang, H. Liu, and Q. Hu, "Transfuse: Fusing transformers and cnns for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 14–24.

[30] J. Ruan, S. Xiang, M. Xie, T. Liu, and Y. Fu, "Malunet: A multi-attention and light-weight unet for skin lesion segmentation," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 1150–1156.

[31] M. Zhang, Y. Yu, S. Jin, L. Gu, T. Ling, and X. Tao, "Vm-unet-v2: rethinking vision mamba unet for medical image segmentation," in *International symposium on bioinformatics research and applications*. Springer, 2024, pp. 335–346.