Patch-MoE Mamba: A Patch-Ordered Mixture-of-Experts State Space Architecture for Medical Image Segmentation

Diego Adame[1] DIEGO.ADAME01@UTRGV.EDU
Jose A. Nuñez[1] JOSE.NUNEZ01@UTRGV.EDU
Fabian Vazquez[1] FABIAN.VAZQUEZ03@UTRGV.EDU
Huimin Li[2] HUIMIN.LI01@UTRGV.EDU
Erik Enriquez[1] ERIK.ENRIQUEZ01@UTRGV.EDU
DongChul Kim[1] DONGCHUL.KIM@UTRGV.EDU
Haoteng Tang[1] HAOTENG.TANG@UTRGV.EDU
Bin Fu[1] BIN.FU@UTRGV.EDU
Pengfei Gu[1] PENGFEI.GU01@UTRGV.EDU

[1] *Department of Computer Science, University of Texas Rio Grande Valley, Edinburg, TX 78539, USA*

[2] *School of Mathematical and Statistical Sciences, University of Texas Rio Grande Valley, Edinburg, TX 78539, USA*

Editors: Under Review for MIDL 2026

Abstract

Advanced convolutional neural networks (CNNs) and Transformer-based architectures currently achieve state-of-the-art performance in medical image segmentation. However, CNNs have limited capacity to model long-range dependencies, while Transformers incur at least quadratic computational and memory complexity in the number of tokens, which can hinder their deployment in resource-constrained clinical settings and make model training and tuning more demanding. Recently, state space models (SSMs), such as Vision Mamba, have gained attention for their ability to capture global dependencies with linear complexity in sequence length. Despite promising results, existing Mamba-based segmentation networks (e.g., VM-UNetV2) still face two key challenges for medical image segmentation: (1) pixel-wise scanning along fixed directions does not sufficiently preserve or exploit local 2D spatial structure, and (2) feature fusion across scan directions typically relies on simple summation, which fails to adapt to varying object sizes and shapes, leading to inaccurate boundary localization and incomplete object masks. To address these limitations, we propose **Patch-MoE Mamba**, a patch-ordered mixture-of-experts (MoE) state space architecture for medical image segmentation. First, we develop a hierarchical, patch-ordered scanning mechanism that partitions feature maps into local patches and applies directional scanning with multiple patch sizes at different stages, thereby better preserving spatial neighborhoods and capturing multi-scale spatial context. Second, we introduce a new MoE-based fusion module that adaptively combines the output signals from multiple directional Mamba scanners. This module integrates four directional scanners with a learnable concatenation expert and incorporates a residual summation of all directional outputs, which stabilizes expert weight computation and yields more discriminative fused feature representations. Extensive experiments on five public polyp segmentation benchmarks and the ISIC 2017 and 2018 skin lesion segmentation datasets demonstrate the effectiveness and generality of the proposed Patch-MoE Mamba.

Keywords: Vision Mamba; state space models; mixture-of-experts; medical image segmentation

1. Introduction

Medical image segmentation is a fundamental task in computer-assisted diagnosis and treatment planning, as it aims to delineate clinically relevant structures (e.g., organs, lesions, and anatomical regions) from complex imaging data [6; 16]. Advanced convolutional neural networks (CNNs) and Transformer-based architectures currently achieve state-of-the-art (SOTA) performance in medical image segmentation, benefiting from powerful feature extraction and rich contextual modeling [13; 21; 9; 8; 20; 5; 12]. However, CNNs have limited capacity to capture long-range dependencies due to their inherently local receptive fields, while Transformers often require large-scale labeled data and incur substantial computational and memory overhead, which can hinder their deployment in resource-constrained clinical settings and make model training and tuning more demanding.

Recently, Mamba-based architectures have shown strong potential for medical image segmentation due to their ability to capture global dependencies with linear complexity in sequence length [7; 11]. Despite promising results, existing Mamba-based segmentation networks still face two key challenges in medical image segmentation. First, pixel-wise scanning along fixed directions does not sufficiently preserve or exploit local 2D spatial structure. Existing image Mamba models [7; 11; 14; 19] typically treat a 2D feature map as a long 1D token sequence and apply pixel-level raster scanning. While this formulation leverages the sequential modeling strength of state space models, it implicitly destroys local 2D structure: pixels that are adjacent in the scan order may be far apart spatially, and spatially neighboring pixels can be separated by a large sequence distance. As a result, important local patterns such as lesion boundaries, fine surface structures, and local context around small or low-contrast objects are diluted or distorted in the sequence representation. This pixel-based scanning therefore tends to lose spatial coherence, which is particularly problematic in clinical settings where accurate boundary localization and shape preservation are crucial, such as in polyp and skin lesion segmentation (see examples in Fig. 1).

Second, feature fusion across scan directions typically relies on simple summation, which fails to adapt to varying object sizes and shapes, leading to inaccurate boundary localization and incomplete object masks. Current Mamba-based segmentation networks [1; 14; 19] usually rely on a fixed way of aggregating the outputs from multiple scan directions or branches (e.g., simple summation), regardless of the underlying object scale or the complexity of the local region. This uniform aggregation implicitly assumes that all directions and receptive fields are equally informative everywhere, which is rarely true in practice. Regions containing small, subtle lesions may require fine-scale, locally focused modeling, while larger structures and complex backgrounds may benefit more from broader, long-range context. A fixed, hand-crafted fusion thus limits the model's ability to adapt its capacity to different lesion sizes, shapes, and appearances. As a result, objects such as colorectal polyps, which can range from tiny, flat, and low-contrast regions to large, protruding, or highly irregular structures, may be under- or over-segmented (see examples in Fig. 1).

To address these limitations, in this paper we propose **Patch-MoE Mamba**, a new patch-ordered mixture-of-experts (MoE) state space architecture specifically tailored for medical image segmentation. First, we introduce a patch-ordered scanning strategy: instead of scanning individual pixels, we partition the feature maps into local patches and perform directional Mamba scanning in a patch-ordered manner. This patch-level sequence construction explicitly preserves local spatial neighborhoods within each patch while still

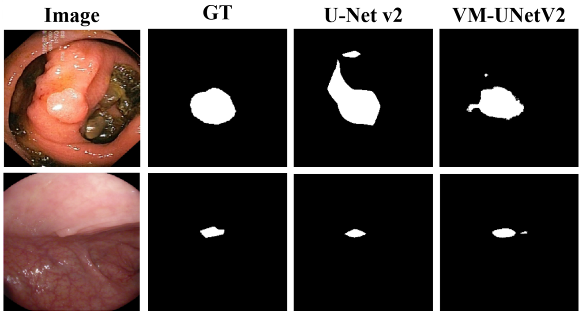

Figure 1: Examples of segmentation results on polyp datasets generated by U-Net v2 [12] and VM-UNetV2 [19], illustrating that small polyps and fine boundary details are not well captured compared to the ground truth (GT).

enabling long-range modeling across patches. By further employing hierarchical patch sizes at different stages, the proposed scanning mechanism captures both fine-scale details and coarse-scale structures, allowing the state space model to better respect the 2D structure of the feature map and enhancing spatial coherence and boundary sensitivity. Second, we replace the fixed summation operation used to aggregate Mamba outputs with an MoE fusion module. Specifically, we treat the outputs of the four directional scanners and their concatenation as five experts, and introduce a spatial-aware gating network to compute spatially varying gating weights, which are used to adaptively fuse these experts. A residual summation of the raw directional outputs is added to the fused result to stabilize training and preserve strong directional signals.

Our experiments on Patch-MoE Mamba using five public polyp segmentation datasets and two public skin lesion segmentation datasets show consistent improvements over strong CNN-, Transformer-, and Mamba-based baselines, demonstrating the effectiveness and generality of Patch-MoE Mamba across diverse medical image segmentation tasks.

2. Methods

2.1. Patch-MoE Mamba Architecture

Fig. 2(a) shows the overall architecture of Patch-MoE Mamba. It follows a U-Net-style design, consisting of a Mamba-based encoder, a semantics and detail infusion (SDI) module, and a decoder. Specifically, we replace the visual state space (VSS) block in VM-UNetV2 [19] with a new Patch-MoE VSS block that integrates a patch-ordered scanning mechanism (Section 2.2) and an adaptive MoE fusion module (Section 2.3) to form the encoder (Fig. 2(b)). We then adopt the SDI module from U-Net v2 [12] (Fig. 2(c)) to enhance the feature maps at each level by infusing semantic information from higher-level features and integrating finer details from lower-level features via the Hadamard product. Finally, we retain the decoder design of VM-UNetV2 [19].

2.2. Patch-Ordered Scanning

Existing Mamba-based models [11; 14; 19] typically treat a 2D feature map as a long 1D token sequence and apply pixel-level raster scanning. As illustrated in Fig. 3(a), standard

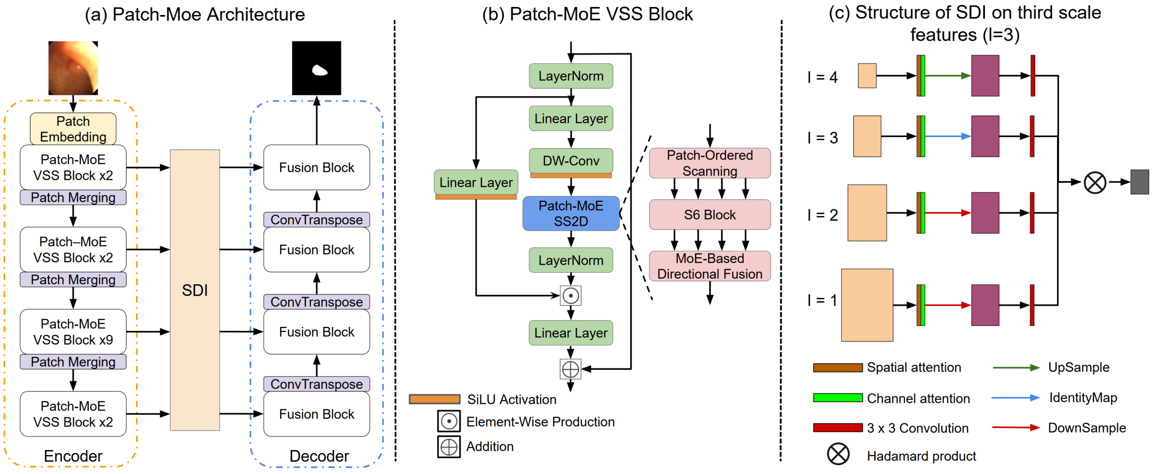

Figure 2: (a) Overview of the proposed Patch-MoE Mamba architecture. (b) Structure of the Patch-MoE Visual State Space (VSS) block. (c) Structure of the Semantics and Detail Infusion (SDI) module. For simplicity, only the refinement of features at the third scale ($l = 3$) is shown.

raster scanning traverses the grid row by row. While this formulation leverages the sequential modeling strength of state space models, it implicitly destroys local 2D structure: pixels that are adjacent in the scan order may be far apart spatially, and spatially neighboring pixels can be separated by a large sequence distance. For example, vertically adjacent pixels (positions 1 and 17 in the 16×16 grid) are separated by 16 steps in the 1D sequence. As a result, important local patterns such as lesion boundaries, fine surface structures, and local context around small or low-contrast objects are diluted or distorted in the sequence representation. This pixel-wise scanning therefore tends to lose spatial coherence, which is undesirable for dense prediction tasks like medical image segmentation.

To preserve spatial coherence during sequence construction, we introduce a *patch-ordered* scanning strategy in place of the raster-based flattening used by standard Vision Mamba modules. Given a feature map $X_l \in \mathbb{R}^{C_l \times H_l \times W_l}$ and a patch size p, we partition the spatial grid into $\lceil H_l/p \rceil \times \lceil W_l/p \rceil$ non-overlapping patches of at most $p \times p$ pixels (Fig. 3(b)). Within each patch, we enumerate all pixel locations (r, c) in row-major order and append their indices to a global index list before proceeding to the next patch. This process yields a permutation vector of length $H_l \times W_l$ that reorders spatial positions but retains every pixel. In other words, the full-resolution feature grid is preserved and no pooling or token reduction occurs; only the visiting order changes. Compared to raster scanning, patch-ordered scanning ensures that all pixels inside a patch are mapped to consecutive positions in the sequence, thereby improving local spatial coherence while still allowing the state space model to capture long-range dependencies across patches.

To further increase spatial locality within contiguous segments of the sequence and to encode multi-scale context, we develop a hierarchical patch-based scanning mechanism (Fig. 3(c)). Instead of using a single patch size, we define a set of patch sizes $\{p^{(1)}, p^{(2)}, \ldots\}$

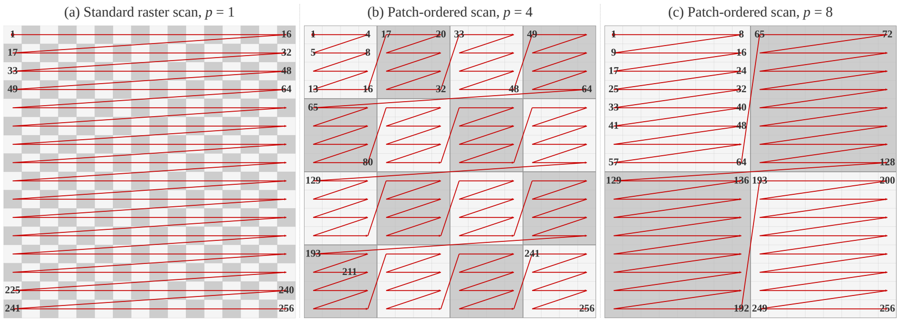

Figure 3: Illustration of the patch-ordered scanning method on a 16×16 grid. (a) Standard raster scan traverses pixels row by row, so vertically adjacent pixels (e.g., positions 1 and 17) are separated by 16 steps in the 1D sequence. (b) With patch size $p = 4$, pixels are grouped into non-overlapping 4×4 patches and scanned patch by patch; all 16 pixels within a patch are processed consecutively before moving to the next patch. (c) With patch size $p = 8$, larger 8×8 patches group 64 pixels each, further increasing spatial locality within contiguous segments of the sequence. For simplicity, only the forward scanning direction is shown.

that includes both small and large patches. Larger patches (e.g., $p^{(2)} = 8$) group more pixels into each contiguous sequence segment, strengthening the modeling of coarser structures and broader regions, while smaller patches (e.g., $p^{(1)} = 4$) enforce fine-grained locality and are well suited for capturing subtle boundary details. For each patch size $p^{(k)}$, we construct a corresponding patch-ordered index sequence as described above. These hierarchical sequences allow the Mamba blocks to attend to both fine-scale and coarse-scale spatial neighborhoods without sacrificing resolution.

Following VM-UNetV2 [19], we employ four scanning directions in each VSS block: (i) forward (left-to-right, top-to-bottom), (ii) reverse, (iii) width–height (WH) forward (top-to-bottom, left-to-right), and (iv) WH reverse. For each direction d, we apply patch-ordered scanning with a possibly distinct patch size $p^{(k_d)}$ drawn from the hierarchical set. This design enables each directional Mamba scanner to operate with its own spatial granularity—e.g., finer patches for horizontally oriented scans and coarser patches for vertically oriented scans—jointly capturing anisotropic and multi-scale spatial structures. The resulting directional sequences are then processed by the state space model and subsequently fused by our MoE-based fusion module (Section 2.3).

2.3. MoE-Based Directional Fusion

In existing Mamba-based segmentation networks [1; 14; 19], feature fusion across scan directions typically relies on simple summation of the directional outputs. This fixed aggregation strategy is agnostic to the underlying object scale and local image complexity. It implicitly assumes that all directions and receptive fields are equally informative at every spatial location, which is rarely true in practice. Regions containing small, subtle lesions may require fine-scale, locally focused modeling, whereas larger structures and cluttered backgrounds may benefit more from broader, long-range context. As a consequence, such hand-crafted fusion can lead to inaccurate boundary localization and incomplete object masks, and it

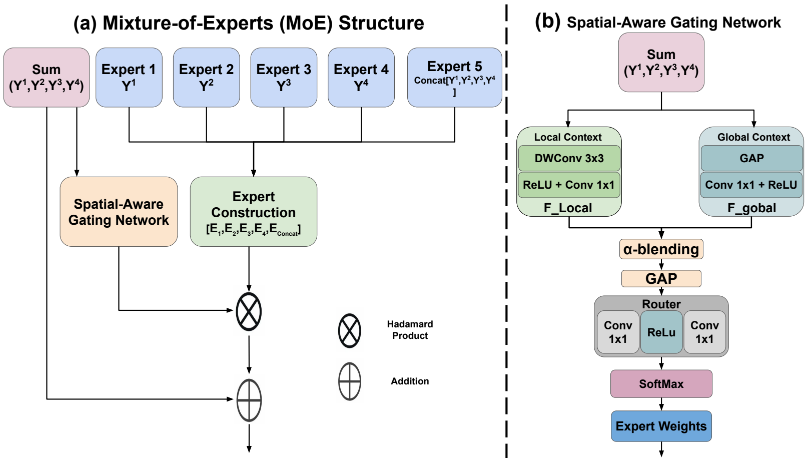

Figure 4: Overview of the proposed MoE-based directional fusion module.

limits the model's ability to adapt its representational capacity to different lesion sizes, shapes, and appearances.

Expert construction. To overcome these limitations, we treat the outputs of the directional scanners as a set of experts and introduce an MoE-based fusion module (Fig. 4). Let $\{Y_l^{(1)}, \dots, Y_l^{(4)}\}$ denote the four directional feature maps at stage l, each in $\mathbb{R}^{C_l \times H_l \times W_l}$. We first normalize each map with Group Normalization to stabilize expert routing and reduce scale discrepancies across directions: $\widehat{Y}_l^{(i)} = \mathrm{GN}(Y_l^{(i)}), \quad i = 1, \dots, 4$. These four normalized maps form the first four experts: $E_1 = \widehat{Y}_l^{(1)},\ E_2 = \widehat{Y}_l^{(2)},\ E_3 = \widehat{Y}_l^{(3)},\ E_4 = \widehat{Y}_l^{(4)}$. In addition to these direction-specific experts, we construct a fifth *concatenation expert* that explicitly models interactions among all directions. We concatenate the four normalized feature maps along the channel dimension and project them back to C_l channels using a 1×1 convolution followed by batch normalization and ReLU: $E_{\text{concat}} = \phi\Big(\mathrm{Conv}_{1\times1}\big(\mathrm{Concat}(\widehat{Y}_l^{(1)}, \dots, \widehat{Y}_l^{(4)})\big)\Big)$, where ϕ denotes batch normalization and ReLU activation. This expert captures cross-directional correlations that cannot be represented by any single directional map alone, providing a more holistic, direction-agnostic view that is especially useful when lesion appearance is not aligned with a particular scanning direction. The final expert stack is $\mathcal{E}_l = [E_1, E_2, E_3, E_4, E_{\text{concat}}]$.

Spatial-aware gating network. To adaptively fuse these experts, we design a spatial-aware gating network (router) that produces spatially varying expert weights (left block in Fig. 4). The router is driven by both local and global context. We first obtain a local context descriptor F_{local} by summing the raw directional outputs and applying a depthwise 3×3 convolution: $F_{\text{local}} = \mathrm{DWConv}_{3\times3}\Big(\sum_{i=1}^{4} Y_l^{(i)}\Big)$. In parallel, we compute a global context descriptor F_{global} via adaptive average pooling to a 1×1 spatial resolution: $F_{\text{global}} = \mathrm{GAP}\Big(\sum_{i=1}^{4} Y_l^{(i)}\Big)$, and broadcast it back to the spatial size $H_l \times W_l$. A learnable scalar $\alpha \in (0, 1)$ then blends these two signals: $F_l = \alpha \cdot F_{\text{local}} + (1 - \alpha) \cdot F_{\text{global}}$. This α-blending allows the router to smoothly trade off between fine local details and coarse global semantics, depending on what is more informative for a given dataset or training stage. A lightweight two-layer 1×1 convolutional router maps F_l to unnormalized logits for the five experts at each spatial location: $\mathbf{z}_l = \text{Router}(F_l) \in \mathbb{R}^{5 \times H_l \times W_l}$. Applying a softmax over the expert dimension yields spatially conditioned expert weights: $\mathbf{w}_l = \text{Softmax}(\mathbf{z}_l), \quad \sum_{e=1}^{5} w_{l,e}(h, w) = 1, \; \forall (h, w)$. The routed fusion is then computed as a weighted sum of experts: $\widetilde{Y}_l = \sum_{e=1}^{5} w_{l,e} \odot E_e$, where $\odot$ denotes element-wise multiplication and $e \in \{1, 2, 3, 4, \text{concat}\}$ indexes the experts.

Residual stabilization. To prevent routing degeneracy and preserve the strong directional signals, we introduce a residual bypass that adds back the raw directional outputs: $Z_l = \widetilde{Y}_l + \sum_{i=1}^{4} Y_l^{(i)}$. This residual path ensures that the encoder retains a robust baseline response even when the gating network is not yet well trained, while the MoE focuses on refining the relative importance of each direction and the concatenation expert. The fused output Z_l is then forwarded to the next encoder stage and/or skip-connection path, providing the SDI decoder with a feature representation that is spatially coherent, directionally enriched, and adaptively fused according to lesion scale and boundary complexity.

Patch-MoE VSS block. Combining the patch-ordered scanning in Section 2.2 with the MoE-based fusion in Section 2.3, we obtain the Patch-MoE VSS block that replaces the original VSS block in VM-UNetV2 [19]. Given an input feature map $X_l \in \mathbb{R}^{C_l \times H_l \times W_l}$, we first apply four directional patch-ordered scans (forward, reverse, WH forward, WH reverse), each with its associated patch size $p^{(k_d)}$, to generate four 1D sequences. Each sequence is processed by a shared-architecture Mamba layer (state space model) and then reshaped back to the 2D grid, yielding the directional feature maps $\{Y_l^{(1)}, \ldots, Y_l^{(4)}\}$. These maps are passed through the expert construction block to form the expert stack $\mathcal{E}_l = [E_1, E_2, E_3, E_4, E_{\text{concat}}]$, and the spatial-aware gating network computes the routing weights that produce the routed fusion $\widetilde{Y}_l$. Finally, we add the residual sum of the raw directional outputs to obtain the block output $Z_l = \widetilde{Y}_l + \sum_{i=1}^{4} Y_l^{(i)}$, which serves as the output of the Patch-MoE VSS block and is forwarded to the next encoder stage or the skip-connection path. In this way, the Patch-MoE VSS block preserves the overall VSS design while replacing fixed summation with adaptive, context-aware expert fusion.

3. Experiments and Results

3.1. Datasets

Polyp segmentation. We conduct experiments on five public polyp segmentation datasets: Kvasir-SEG [10], ClinicDB [2], ColonDB [16], ETIS [15], and CVC-300 [18]. For a fair comparison, we follow the train/test protocol in [12]. Specifically, the training set consists of 900 images from Kvasir-SEG and 550 images from ClinicDB. The test set includes all images from CVC-300 (60), ColonDB (380), ETIS (196), as well as 100 images from Kvasir-SEG and 62 images from ClinicDB.

Skin lesion segmentation. We also evaluate on two skin lesion segmentation datasets: ISIC 2017 [4], which comprises 2,150 dermoscopy images, and ISIC 2018 [3; 17], which contains 2,694 dermoscopy images. For a fair comparison, we follow the train/test split strategy outlined in [12].

Table 1: Quantitative comparison of different methods on five polyp segmentation datasets. The best results are highlighted in bold; the same in the other tables. Each experiment is run 5 times with different random seeds, and the mean and standard deviation across runs are reported.

Datasets	Methods	DSC (%) ↑	IoU (%) ↑	MAE ↓
Kvasir-SEG	U-Net [13]	83.19 ± 0.82	75.58 ± 1.04	0.0472 ± 0.0024
	U-Net v2 [12]	84.39 ± 6.09	76.41 ± 7.56	0.0363 ± 0.0018
	VM-UNet [14]	89.90 ± 0.98	84.00 ± 1.41	0.0294 ± 0.0034
	VM-UNetV2 [19]	90.82 ± 0.25	85.30 ± 0.29	0.0260 ± 0.0020
	Patch-MoE Mamba (Ours)	**90.90 ± 0.35**	**85.32 ± 0.44**	**0.0258 ± 0.0015**
ClinicDB	U-Net [13]	85.68 ± 1.34	79.58 ± 1.25	0.0171 ± 0.0018
	U-Net v2 [12]	86.00 ± 0.79	79.09 ± 1.10	0.0168 ± 0.0028
	VM-UNet [14]	88.46 ± 0.89	82.74 ± 0.92	0.0156 ± 0.0032
	VM-UNetV2 [19]	90.52 ± 0.79	85.34 ± 0.68	0.0120 ± 0.0028
	Patch-MoE Mamba (Ours)	**91.32 ± 0.39**	**86.05 ± 0.47**	**0.0104 ± 0.0011**
ColonDB	U-Net [13]	62.48 ± 2.36	53.99 ± 2.34	0.0488 ± 0.0025
	U-Net v2 [12]	72.40 ± 1.08	62.42 ± 1.09	0.0430 ± 0.0035
	VM-UNet [14]	75.40 ± 1.01	67.80 ± 1.16	0.0356 ± 0.0020
	VM-UNetV2 [19]	76.62 ± 1.28	68.64 ± 1.02	0.0336 ± 0.0022
	Patch-MoE Mamba (Ours)	**77.94 ± 1.60**	**69.65 ± 1.31**	**0.0314 ± 0.0019**
ETIS	U-Net [13]	42.82 ± 1.38	36.36 ± 1.22	0.0365 ± 0.0039
	U-Net v2 [12]	61.43 ± 2.12	51.70 ± 2.24	0.0382 ± 0.0072
	VM-UNet [14]	70.56 ± 2.27	62.40 ± 2.19	0.0240 ± 0.0094
	VM-UNetV2 [19]	72.56 ± 1.55	63.68 ± 1.65	**0.0180 ± 0.0016**
	Patch-MoE Mamba (Ours)	**74.04 ± 0.78**	**64.86 ± 0.94**	0.0196 ± 0.0006
CVC-300	U-Net [13]	77.67 ± 2.00	69.70 ± 1.63	0.0149 ± 0.0016
	U-Net v2 [12]	83.46 ± 1.16	74.99 ± 1.08	0.0128 ± 0.0019
	VM-UNet [14]	86.72 ± 1.25	**80.42 ± 1.32**	0.0098 ± 0.0029
	VM-UNetV2 [19]	86.80 ± 1.21	79.50 ± 1.10	0.0086 ± 0.0015
	Patch-MoE Mamba (Ours)	**87.31 ± 0.42**	79.91 ± 0.77	**0.0078 ± 0.0008**

Table 2: Quantitative comparison of different methods on ISIC 2017 and ISIC 2018 datasets.

Datasets	Methods	DSC (%) ↑	IoU (%) ↑	MAE
ISIC 2017	U-Net [13]	87.48 ± 0.19	79.89 ± 0.24	0.0413 ± 0.0011
	U-Net v2 [12]	87.69 ± 0.19	79.94 ± 0.63	0.0383 ± 0.0018
	VM-UNet [14]	90.03 ± 0.11	83.35 ± 0.24	0.0317 ± 0.0009
	VM-UNetV2 [19]	90.23 ± 0.77	83.59 ± 0.91	0.0310 ± 0.0023
	Patch-MoE Mamba (Ours)	**90.85 ± 0.93**	**84.45 ± 1.23**	**0.0293 ± 0.0026**
ISIC 2018	U-Net [13]	86.82 ± 0.33	78.64 ± 0.38	0.0660 ± 0.0025
	U-Net v2 [12]	88.14 ± 0.40	80.24 ± 0.52	0.0528 ± 0.0016
	VM-UNet [14]	87.42 ± 0.47	79.66 ± 0.65	0.0558 ± 0.0017
	VM-UNetV2 [19]	88.36 ± 0.33	80.90 ± 0.40	0.0550 ± 0.0021
	Patch-MoE Mamba (Ours)	**89.34 ± 0.39**	**82.28 ± 0.58**	**0.0496 ± 0.0023**

3.2. Experimental Results

Table 1 presents a quantitative comparison of Patch-MoE Mamba against U-Net, U-Net v2, VM-UNet, and VM-UNetV2 on the five public polyp segmentation datasets. From these results, we make the following observations. First, Patch-MoE Mamba achieves the best performance on all five datasets in terms of Dice score, while also producing highly competitive IoU and MAE values. This demonstrates that the proposed patch-ordered scanning and adaptive MoE fusion lead to consistent performance gains across diverse colonoscopic benchmarks. More importantly, the largest improvement over the previous SOTA model VM-UNetV2 is observed on the most challenging dataset, ETIS. Specifically, Patch-MoE Mamba improves the Dice score from 72.56 to 74.04, corresponding to an absolute gain of

Table 3: Ablation study on different components of Patch-MoE Mamba.

Method	Kvasir-SEG	ClinicDB	ColonDB	ETIS	CVC-300	Average
VM-UNetV2 [19]	90.82	90.52	76.62	72.56	86.80	83.46
VM-UNetV2 w/ Patch-Ordered Scanning	**91.14**	91.12	76.68	73.76	**87.40**	84.02
VM-UNetV2 w/ Patch-Ordered Scanning + MoE Fusion (Ours)	90.90	**91.32**	**77.94**	**74.04**	87.31	**84.30**

Table 4: Ablation study on the components of the MoE-based directional fusion module.

Method	Kvasir-SEG	ClinicDB	ColonDB	ETIS	CVC-300	Average
Patch-MoE w/o Concat Expert	90.70	90.80	76.90	73.50	86.00	83.58
Patch-MoE w/o Residual Addition	**90.90**	90.90	77.10	73.10	86.60	83.72
Patch-MoE Mamba (Ours)	**90.90**	**91.32**	**77.94**	**74.04**	**87.31**	**84.30**

1.48 Dice points. This dataset contains low-contrast images, small-scale polyps, and highly irregular lesion boundaries, making it particularly sensitive to spatial modeling and boundary localization. The strong improvement on ETIS verifies the effectiveness of preserving local spatial structure via patch-ordered sequence construction and adaptively weighting multi-directional scan responses through MoE fusion.

Table 2 further reports performance on the ISIC 2017 and ISIC 2018 skin lesion segmentation datasets. Patch-MoE Mamba achieves the best results on both benchmarks, confirming that the proposed architecture generalizes beyond colonoscopic imagery to a different medical imaging modality with distinct appearance characteristics.

Fig. 5 illustrates several qualitative segmentation examples from the seven datasets. From these visual results, we observe that Patch-MoE Mamba produces segmentation masks with cleaner boundaries and fewer false positive predictions compared with the baseline methods. In particular, our method more effectively suppresses spurious activations in background regions (as shown on ColonDB and ETIS) and produces sharper boundary delineation (as shown on Kvasir-SEG), yielding predictions that are closer to the GT.

3.3. Ablation Study

Table 3 evaluates the contributions of the proposed patch-ordered scanning strategy and the MoE fusion module. Starting from the VM-UNetV2 baseline, incorporating patch-ordered scanning alone improves the average Dice score from 83.46 to 84.02, demonstrating that reorganizing the token sequence into spatially coherent patches enhances feature representation. Further integrating the MoE fusion module yields an additional performance gain, raising the average Dice score to 84.30. This confirms the effectiveness of the proposed MoE-based directional fusion.

Table 4 examines the individual contributions of the MoE fusion components. Removing the concatenation expert leads to a noticeable performance drop, reducing the average Dice score from 84.30 to 83.58. This indicates that explicit cross-directional feature interaction plays a crucial role in enhancing the representational capacity of the fusion module. Removing the residual summation results in a smaller degradation to 83.72, showing that the residual path primarily contributes to training stability rather than serving as the dominant source of performance improvement.

Table 5: Ablation study of different patch sizes at different stages.

Patch Sizes	Kvasir-SEG	ClinicDB	ColonDB	ETIS	CVC-300	Average
8844/8844/8844/8844	90.70	90.00	75.67	71.58	86.41	82.87
8844/8844/8844/1111	90.61	90.02	75.97	71.43	85.60	82.73
8844/8844/1111/1111	90.47	89.95	76.32	71.55	86.54	82.97
8844/1111/1111/1111	**91.13**	**91.10**	**76.69**	**73.74**	**87.41**	**84.01**

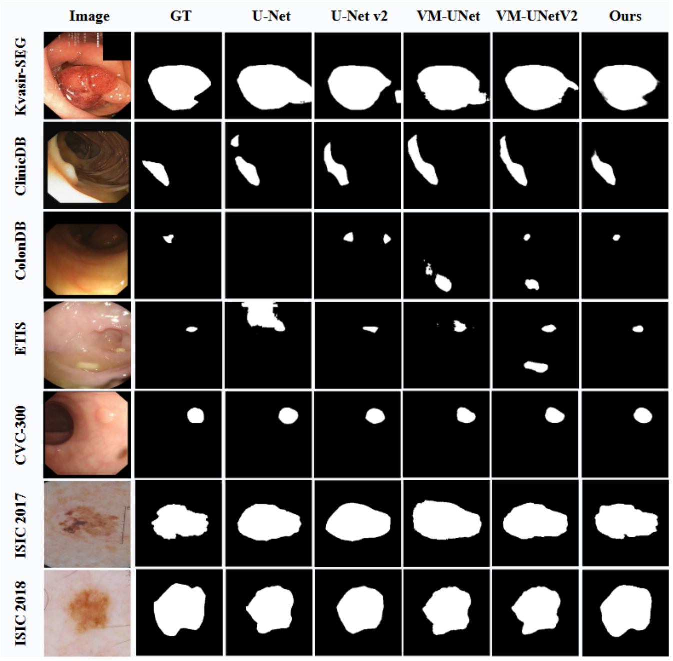

Figure 5: Visual examples of segmentations results.

Table 5 investigates the influence of different patch-size configurations. Among all tested settings, the configuration 8844/1111/1111/1111 for stages 1 to 4 achieves the highest average Dice score of 84.01. This result suggests that using coarse-grained scanning in shallow layers and fine-grained scanning in deeper layers enables more effective multi-scale feature modeling, balancing global context aggregation with precise boundary localization.

4. Conclusions

In this paper, we presented **Patch-MoE Mamba**, a patch-ordered mixture-of-experts state space architecture for medical image segmentation. We first designed a hierarchical, patch-ordered scanning mechanism that partitions feature maps into local patches and applies directional scanning with multiple patch sizes at different stages, thereby preserving spatial neighborhoods while capturing multi-scale spatial context. We then introduced an MoE-based fusion module that adaptively combines the outputs of multiple directional Mamba scanners and a learnable concatenation expert, together with a residual summation of all directional outputs, leading to more stable expert weights and more discriminative fused features. Extensive experiments on five public polyp segmentation benchmarks and the ISIC 2017 and 2018 skin lesion segmentation datasets demonstrated the effectiveness and generality of the proposed Patch-MoE Mamba.

Acknowledgments

This research was supported in part by NSF grant CCF-2523787. We also acknowledge the UTRGV High Performance Computing Resource, supported by NSF grants 2018900 and IIS-2334389, and DoD grant W911NF2110169.

References

[1] Diego Adame, Jose A Nunez, Fabian Vazquez, Nayeli Gurrola, Huimin Li, Haoteng Tang, Bin Fu, and Pengfei Gu. Topo-VM-UNetV2: Encoding topology into vision Mamba UNet for polyp segmentation. In *Proceedings of the IEEE 38th International Symposium on Computer-Based Medical Systems*, pages 258–263, 2025.

[2] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[3] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael 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.

[4] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald 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 *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging*, pages 168–172, 2018.

[5] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-PVT: Polyp segmentation with pyramid vision Transformers. *arXiv preprint arXiv:2108.06932*, 2021.

[6] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 263–273, 2020.

[7] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[8] Pengfei Gu, Hao Zheng, Yizhe Zhang, Chaoli Wang, and Danny Z Chen. kCBAC-Net: Deeply supervised complete bipartite networks with asymmetric convolutions for medical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 337–347, 2021.

[9] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[10] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-SEG: A segmented polyp dataset. In *Proceedings of the International Conference on Multimedia Modeling*, pages 451–462, 2020.

[11] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. *Advances in Neural Information Processing Systems*, 37:103031–103063, 2024.

[12] Yaopeng Peng, Danny Z Chen, and Milan Sonka. U-Net v2: Rethinking the skip connections of U-Net for medical image segmentation. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, pages 1–5, 2025.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 234–241, 2015.

[14] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. VM-UNet: Vision Mamba U-Net for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

[15] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Journal of CARS*, 9:283–293, 2014.

[16] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging*, 35(2):630–644, 2015.

[17] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

[18] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, Aaron Courville, et al. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of Healthcare Engineering*, 2017.

[19] Mingya Zhang, Yue Yu, Sun Jin, Limei Gu, Tingsheng Ling, and Xianping Tao. VM-UNetV2: rethinking vision Mamba UNet for medical image segmentation. In *Proceedings of the International Symposium on Bioinformatics Research and Applications*, pages 335–346, 2024.

[20] Yundong Zhang, Huiye Liu, and Qiang Hu. TransFuse: Fusing Transformers and CNNs for medical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions*, pages 14–24, 2021.

[21] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. U-Net++: A nested U-Net architecture for medical image segmentation. In *Proceedings of the International Workshop on Deep Learning in Medical Image Analysis*, pages 3–11, 2018.

Table 6: Comparison of the computational complexity of different models.

Model	Input size	Params (M) ↓	FLOPs (G) ↓
U-Net v2 [12]	(3, 256, 256)	25.15	5.58
VM-UNetV2 [19]	(3, 256, 256)	**22.77**	**5.31**
Patch-MoE w/o Concat Expert	(3, 256, 256)	28.44	10.03
Patch-MoE w/o Residual Addition	(3, 256, 256)	70.06	28.18
Patch-MoE Mamba (Ours)	(3, 256, 256)	70.06	28.18

Appendix A. Experimental Setup

All experiments are conducted using PyTorch. The model is trained on an NVIDIA Tesla A100 GPU (80 GB memory) using the AdamW optimizer with an initial learning rate of 1×10^{-3} and a batch size of 80. Following [19], we resize all images to 256×256. A cosine annealing learning rate schedule is employed with a minimum learning rate of 1×10^{-5} and a cycle length of 50 epochs. Training is performed for 300 epochs. Following prior work, the Vision Mamba encoder in VM-UNetV2 is initialized with pretrained VMamba-S [11] weights. Standard data augmentation (e.g., random flipping and random rotation) is applied to mitigate overfitting. We report Dice similarity coefficient (DSC), intersection over union (IoU), and mean absolute error (MAE). Each experiment is run 5 times with different random seeds, and we report the mean and standard deviation across runs.

Appendix B. Computational Complexity

We report the number of parameters and FLOPs for all models. As shown in Table 6, the introduction of patch-ordered scanning does not change the overall computational order, as it only reorganizes the token traversal sequence while preserving the total number of tokens. In contrast, the concatenation expert substantially increases the computational cost. When the concatenation expert is enabled, the parameter count rises to 70.06M and the FLOPs increase to 28.18 GFLOPs. This is expected, as the concatenation expert explicitly aggregates the four directional feature maps along the channel dimension and applies an additional 1×1 projection layer, leading to denser channel interactions and higher computational overhead.