MambaMIC: An Efficient Baseline for Microscopic Image Classification with State Space Models

Shun Zou[1,5*], Zhuo Zhang[2*], Yi Zou[3], Guangwei Gao[4,5†]
[1]Nanjing Agricultural University [2]National University of Defense Technology
[3]Xiangtan University [4]Nanjing University of Posts and Telecommunications [5]Soochow University
zs@stu.njau.edu.cn, zhangzhuo@nudt.edu.cn, 202205570112@smail.xtu.edu.cn, csgwgao@njupt.edu.cn

2025 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3315-9495-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICME59968.2025.11209808

***Abstract*—In recent years, CNN and Transformer-based methods have made significant progress in Microscopic Image Classification (MIC). However, existing approaches still face the dilemma between global modeling and efficient computation. While the Selective State Space Model (SSM) can simulate long-range dependencies with linear complexity, it still encounters challenges in MIC, such as local pixel forgetting, channel redundancy, and lack of local perception. To address these issues, we propose a simple yet efficient vision backbone for MIC tasks, named MambaMIC. Specifically, we introduce a Local-Global dual-branch aggregation module: the MambaMIC Block, designed to effectively capture and fuse local connectivity and global dependencies. In the local branch, we use local convolutions to capture pixel similarity, mitigating local pixel forgetting and enhancing perception. In the global branch, SSM extracts global dependencies, while Locally Aware Enhanced Filter reduces channel redundancy and local pixel forgetting. Additionally, we design a Feature Modulation Interaction Aggregation Module for deep feature interaction and key feature re-localization. Extensive benchmarking shows that MambaMIC achieves state-of-the-art performance across five datasets. code is available at https://zs1314.github.io/MambaMIC.**

***Index Terms*—Microscopic Image Classification, State Space Model, Mamba, Local Perception Enhancement**

I. Introduction

Microscopic imaging technology plays a crucial role in the medical field and is an indispensable tool in modern medical research and clinical diagnosis [1]. By classifying microscopic images, medical researchers can observe the structural and dynamic changes at the tissue, cellular, and molecular levels, leading to a deeper understanding of disease mechanisms [2].

In recent years, inspired by the success of deep learning in various vision tasks, many studies have developed different network architectures based on Convolutional Neural Networks (CNNs) and applied them to MIC [3]–[5]. Although convolution operations can effectively model local connectivity, their inherent characteristics, such as limited local receptive fields, hinder the extraction of long-range dependencies, often resulting in insufficient semantic context extraction and incomplete feature representation. Fortunately, inspired by the Transformer in natural language processing and advanced vision tasks [6], Transformer-based architectures have been developed for MIC tasks [3], [12], [13]. Thanks to the self-attention mechanism, they can effectively model global dependencies, alleviating the limitations of CNN models. However, Transformer-based methods still face a significant challenge: they exhibit high quadratic complexity when modeling long sequences, leading to substantial computational overhead. This disregards the computational constraints in real-world medical environments and fails to meet the need for low-parameter, low-computation models in mobile MIC [14]. While some studies have adopted efficient attention techniques, such as mobile window attention [15], [16], these approaches still fail to fully exploit the information within each patch, often sacrificing the global receptive field and not fundamentally resolving the trade-off between global dependency modeling and efficient computation.

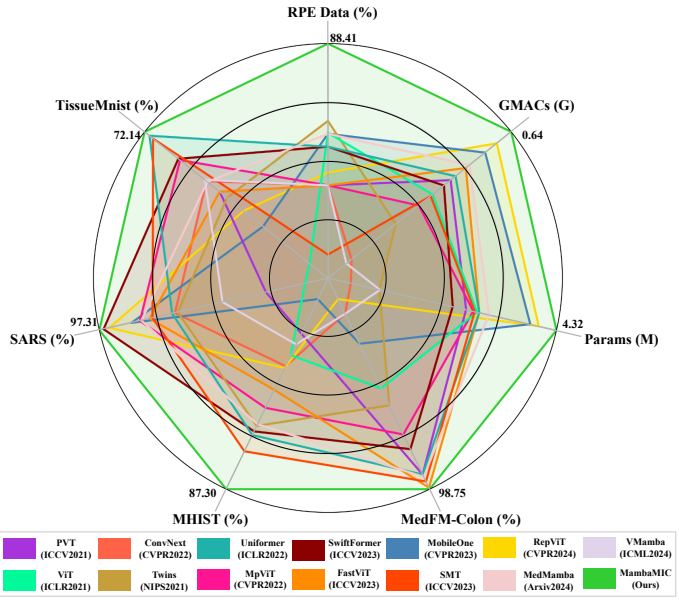

Fig. 1: A seven-dimensional radar map of the Overall Accuracy of RPE Data [7], TissueMnist [8], SARS [9], MHIST [10], MedFM-Colon [11], along with Params and GMACs.

Recently, State Space Models (SSM) [17] have attracted significant attention from researchers. Building upon classical SSM research, modern SSMs like Mamba [18] not only establish long-range dependencies but also exhibit linear complexity with respect to input size, making Mamba a strong competitor to CNNs and Transformers in lightweight MIC tasks. However, Mamba still faces three major challenges when applied to MIC: (1) Since Mamba processes the flattened

*Equal Contribution, †Corresponding author.

This work was supported in part by the Open Fund Project of Provincial Key Laboratory for Computer Information Processing Technology (Soochow University) under Grant KJS2274.

1D image sequence in a recursive manner, it may cause adjacent pixels in the 2D space to be far apart in the flattened sequence, leading to local pixel forgetting. In MIC tasks, adjacent pixels often have strong relationships, and the loss of this relationship results in the loss of semantic information in multiple local blocks, creating a cumulative effect and leading to a disastrous pixel forgetting phenomenon; (2) Due to the need to memorize long-range dependencies in Visual State Space Models (VSSM), the number of hidden states in the state space equations becomes very large, which not only leads to channel information redundancy and increased computational burden, but also generates a significant amount of irrelevant interference, thereby hindering the representation learning of critical channel information; (3) Unlike other vision tasks with clearly defined target features, MIC requires not only the capture of global context but also a focus on local fine-grained features.

To address the above issues, we introduce MambaMIC, a simple yet highly effective baseline model. Its core idea is to fully leverage the local feature extraction advantages of CNNs and the global modeling strengths of Mamba, while maintaining linear complexity and a low computational burden. Specifically, the core component of MambaMIC is the MambaMIC Block, which adopts a Local-Global dual-branch architecture designed to effectively extract and aggregate both local invariant features and long-range dependency characteristics. In the Local branch, we use local convolutional designs that, on the one hand, capture fine-grained local features, providing local connectivity and enhancing local perception ability, and on the other hand, alleviate the local pixel forgetting problem faced by vanilla Mamba [18] when dealing with 2D images (see Fig. 2). The Global branch, composed of multiple parallel Residual Efficient Vision State Space Modules (REVSSM), mitigates information blocking caused by the exponential increase in hidden state numbers as the number of channels grows, thanks to the parallel mechanism. Moreover, in the REVSSM, we introduce the Locally Aware Enhanced Filter (LAEF), which employs a sophisticated channel selection and pruning mechanism to enhance the local perception capability of VSSM, promote context expert information interaction and flow, and reduce channel redundancy caused by excessive hidden states, allowing the most valuable information to circulate globally. Simultaneously, LAEF and the local convolutions of the Local branch form a complementary flow, enhancing local pixel blocks through both parallel and serial paradigms. Additionally, we observe a non-negligible feature gap between the Local and Global branches. Simple addition or concatenation inevitably leads to the loss of valuable information, limiting performance improvements. Therefore, to further promote feature fusion and information interaction within the paradigm, we propose the Feature Modulation Interaction Aggregation Module (FMIAM). FMIAM achieves deep fusion and interaction by adaptively weighting the corresponding branches, and we also incorporate a simplified channel attention mechanism to recalibrate and localize channel features, filtering out irrelevant features and enhancing the representation of key features. Finally, comprehensive experiments demonstrate that MambaMIC achieves the optimal performance-parameter trade-off, making it a true "jack-of-all-trades," as shown in Fig. 1. In summary, our main contributions can be summarized as follows:

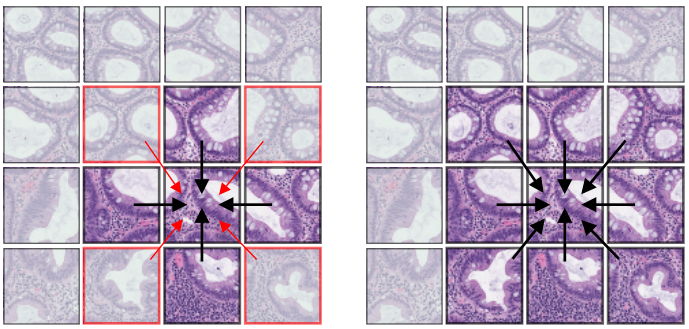

Fig. 2: In Mamba's one-dimensional recursive image processing, local pixels (highlighted in red) are easily forgotten in the flattened sequence. However, enhancing local perception effectively captures pixel relationships.

- We are the first to apply State Space Models (SSM) to MIC through extensive experiments, leading to the proposal of MambaMIC, a simple yet effective alternative to CNN- and Transformer-based methods.
- We introduce a simple and efficient dual-branch architecture, the MambaMIC Block, consisting of a local branch and a global branch. Specifically, we develop the Residual Efficient Vision State Space Module as the core of the global branch and enhance local perception using the Locally Aware Enhanced Filter, promoting the interaction and flow of contextual channel information while reducing channel redundancy caused by excessive hidden states. Additionally, we introduce the Feature Modulation Interaction Aggregation Module to effectively bridge the semantic gap between different types of features and better aggregate diverse information.
- Extensive experiments on five datasets demonstrate that our MambaMIC outperforms other state-of-the-art methods, providing a new benchmark and reference for MIC.

II. Method

A. Overall Pipeline

Fig. 3 illustrates the overall architecture of MambaMIC. Consistent with previous general visual backbones [19]–[21], MambaMIC is divided into four stages, each consisting of several stacked MambaMIC Blocks. Additionally, each stage is preceded by an Embedding or Merging layer for spatial downsampling and channel expansion. A global average pooling layer is applied to the final output, which is then fed into a linear classification head. The MambaMIC Block is the core component of MambaMIC. Specifically, as shown in Fig. 3, it adopts a Local-Global dual-branch interactive aggregation architecture. In the global branch, we introduce four parallel Residual Efficient Vision State Space Modules (REVSSM). The parallel mechanism alleviates the computational burden caused by the exponential increase in the number of states with the number of channels in the visual state space, while also

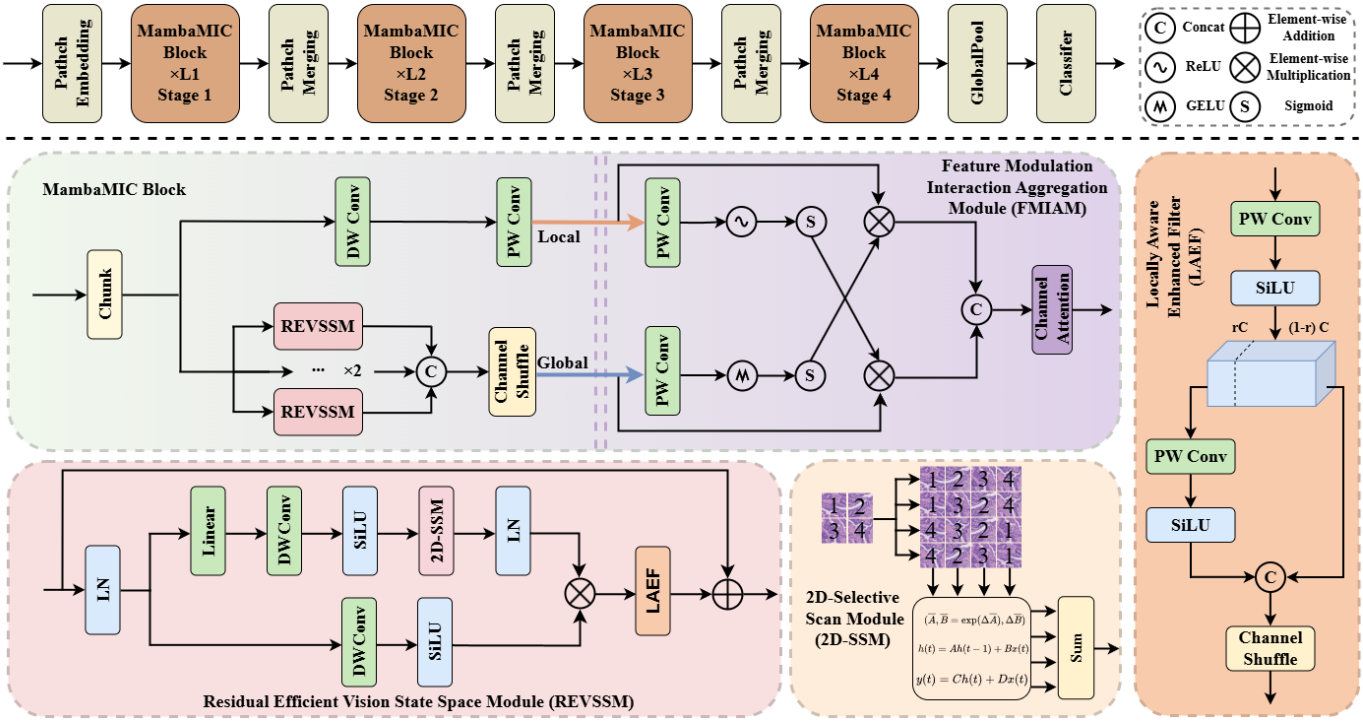

Fig. 3: The overall architecture of the proposed MambaMIC.

promoting the interaction of contextual channel information, compensating for the lack of global relationship modeling at the channel level in VSSM. Additionally, we observed that since SSM processes the flattened feature map into a 1D sequence of tokens, the proximity of adjacent pixels in the sequence is highly influenced by the flattening strategy. For instance, when using the four-direction unfolding strategy from [21], only four nearest adjacent pixels are available for each anchor, meaning some spatially neighboring pixels in the 2D feature map are far apart in the 1D token sequence. This long-distance separation leads to local pixel forgetting, where the relationships between pixels gradually diminish (see Fig. 2). To address the missing local information in Mamba, we introduce additional local convolutions in the local branch to help restore pixel neighborhood similarity. Furthermore, we develop the Feature Modulation Interaction Aggregation Module (FMIAM) to reduce the knowledge gap between the local and global branches, fully fuse internal paradigm features, and implement channel recalibration via a simplified channel attention mechanism. Mathematically, the entire process of the MambaMIC Block can be expressed as follows:

$$F', F'' = Chunk(F), \tag{1}$$

$$F_L = PW(DW(F')), \tag{2}$$

$$F_1, F_2, F_3, F_4 = Split(F''), \tag{3}$$

$$\hat{F}_i = REVSSM(F_i), i \in [1, 2, 3, 4], \tag{4}$$

$$F_G = Shuffle([\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4]), \tag{5}$$

$$\hat{F} = FMIAM(F_L, F_G). \tag{6}$$

where Chunk(·) represents channel splitting, DW(·) denotes Depthwise Separable Convolution, PW(·) refers to Point-Wise Convolution, Shuffle(·) indicates Channel Shuffle, [,] represents channel concatenation, REVSSM(·) stands for Residual Efficient Vision State Space Module, and FMIAM(·) denotes Feature Modulation Interaction Aggregation Module.

B. Residual Efficient Vision State Space Module

Fig. 3 illustrates the Residual Efficient Vision State Space Module (REVSSM), which utilizes state space equations to capture long-range dependencies. Specifically, given the input feature $X \in \mathbb{R}^{H \times W \times C}$, after layer normalization, the features pass through two parallel branches. In the first branch, the feature channels undergo Depthwise Separable Convolution (DWConv) and are expanded to λC (where λ is the channel expansion factor), followed by processing with the SiLU activation function. In the second branch, the feature channels are first expanded to λC via a linear layer, then processed with DWConv, SiLU activation, a 2D-selective scanning module (2D-SSM), and a LayerNorm layer. The features from both branches are then roughly aggregated through element-wise multiplication. Finally, the rough-aggregated features pass through the Locally Aware Enhanced Filter (LAEF) to enhance local information perception and capture pixel neighborhood similarity, while mitigating channel information redundancy. Residual connections are introduced, resulting in the output $\hat{X}$. The process is formally expressed as follows:

$$X_1 = LN(2D\text{-}SSM(SiLU(DW(Linear(X))))), \tag{7}$$

$$X_2 = SiLU(DW(LN(X))), \tag{8}$$

$$\hat{X} = X + LAEF(X_1 \odot X_2), \tag{9}$$

where Linear(·) represents processing using a linear layer, LN(·) denotes the layer normalization process, SiLU(·) is the SiLU activation function, 2D-SSM(·) denotes the 2D Selective Scanning Module, $\odot$ represents element-wise multiplication, and LAEF(·) denotes the Locally Aware Enhanced Filter.

TABLE I: Comparison with state-of-the-art methods on five public datasets [7]–[11]. The best results are highlighted in **bold fonts**. " ↑ "and " ↓ " indicate that larger or smaller is better.

	Method	Year	GMACs↓	Params↓	RPE Data [7]		MHIST [10]		SARS [9]		TissueMnist [8]		FM-Colon [11]	
					OA↑	AUC↑	OA↑	AUC↑	OA↑	AUC↑	OA↑	AUC↑	OA↑	AUC↑
CNNs	ConvNext-tiny [22]	CVPR2022	4.49	28.69	86.06	97.96	77.34	84.24	96.88	99.46	66.47	92.07	94.56	98.71
	ConvNext-small [22]	CVPR2022	8.73	50.22	85.44	98.02	76.55	84.12	96.77	99.42	69.22	93.21	87.36	95.05
	RepViT-m1_0 [23]	CVPR2024	1.13	6.85	87.60	98.15	69.10	71.32	96.62	99.42	67.54	92.45	97.45	99.63
	RepViT-m1_1 [23]	CVPR2024	1.37	8.29	85.71	98.06	76.70	80.90	97.24	99.56	67.47	92.41	86.41	93.60
	MobileOne-s0 [24]	CVPR2023	1.10	5.29	86.79	98.28	67.83	60.04	96.03	99.17	63.20	90.34	89.26	95.97
	MobileOne-s2 [24]	CVPR2023	1.35	7.88	85.71	98.16	70.68	73.07	96.61	99.33	62.66	90.11	87.76	94.62
	MobileOne-s3 [24]	CVPR2023	1.96	10.17	86.52	98.24	70.69	73.44	97.08	99.47	66.55	92.05	89.31	95.78
ViTs	PVT-smal [25]	ICCV2021	3.71	24.49	85.44	98.13	73.69	80.11	96.12	99.16	68.60	92.88	97.80	99.81
	PVT-medium [25]	ICCV2021	6.49	44.21	87.33	98.19	80.82	87.72	96.61	99.31	69.17	93.14	97.25	99.57
	MpViT-tiny [26]	CVPR2022	1.84	5.84	87.87	98.11	81.30	87.37	96.93	99.43	70.63	93.79	98.05	99.83
	MpViT-small [26]	CVPR2022	5.32	22.89	85.44	97.95	80.19	86.57	97.02	99.52	70.47	93.73	95.20	98.51
	Twins-small [27]	NIPS2021	3.71	24.11	88.14	98.24	78.76	86.15	96.61	99.40	67.80	82.55	97.45	99.65
	Twins-base [27]	NIPS2021	6.49	43.83	86.79	98.40	81.77	88.93	96.75	99.43	68.25	92.82	93.31	98.29
	ViT-tiny [28]	ICLR2021	1.26	5.71	86.52	97.50	74.33	79.33	96.01	99.18	58.50	87.80	96.15	99.38
	ViT-small [28]	ICLR2021	4.62	22.04	86.54	97.76	75.59	80.94	95.84	99.20	64.36	90.94	92.21	97.63
	ViT-base [28]	ICLR2021	17.6	86.54	86.24	97.25	75.44	82.59	96.43	99.37	65.86	91.73	94.61	94.60
Hybrid-CNN-ViT	FastViT-sa24 [29]	ICCV2023	2.94	21.55	85.44	98.07	78.61	84.04	96.95	99.43	68.60	92.92	98.65	99.89
	FastViT-ma36 [29]	ICCV2023	6.07	44.07	88.14	98.08	81.93	85.55	96.95	99.45	68.79	93.00	94.31	98.43
	SwiftFormer-S [19]	ICCV2023	1.01	5.64	85.44	98.04	81.62	90.16	96.88	99.45	69.57	93.40	94.91	98.77
	SwiftFormer-L1 [19]	ICCV2023	1.62	11.29	86.52	98.02	81.93	88.47	97.21	99.54	70.59	93.66	95.45	98.92
	SwiftFormer-L3 [19]	ICCV2023	4.05	27.47	86.25	98.04	82.25	89.95	97.28	99.53	70.55	93.58	96.15	99.27
	Uniformer-small [30]	ICLR2022	3.46	21.55	86.25	88.13	82.57	89.96	96.79	99.45	71.91	94.27	97.80	99.71
	Uniformer-base [30]	ICLR2022	7.81	49.78	88.14	83.41	81.77	89.08	97.22	99.49	72.06	94.08	95.80	99.15
	SMT-s [20]	ICCV2023	4.72	22.55	83.99	83.73	83.99	89.73	96.93	99.46	71.74	94.33	98.25	99.84
	SMT-b [20]	ICCV2023	7.81	32.04	87.87	98.10	86.05	91.04	96.90	99.45	69.27	93.36	97.85	99.73
Mambas	Medmamba-s [31]	Arxiv2024	2.75	19.39	86.52	98.17	81.62	87.04	97.01	99.23	69.18	93.12	97.95	99.85
	Medmamba-b [31]	Arxiv2024	6.16	40.88	86.79	98.02	77.97	85.20	97.30	99.51	69.11	93.18	95.70	98.94
	VMamba-t [21]	ICML2024	4.4	22.1	85.71	97.80	77.34	83.32	95.92	99.28	69.23	93.13	92.66	97.84
	VMamba-s [21]	ICML2024	9.0	43.7	85.44	97.79	74.64	81.27	96.43	99.37	69.30	93.20	87.36	94.77
Ours	MambaMIC-t	-	**0.64**	**4.32**	**88.41**	**98.29**	87.30	93.49	**97.31**	**99.60**	72.14	94.31	**98.75**	**99.90**
	MambaMIC-s	-	0.76	4.97	87.06	98.24	86.05	92.31	97.13	99.49	**73.01**	**94.40**	98.00	99.78
	MambaMIC-b	-	1.59	8.39	87.33	98.17	**87.80**	**94.17**	97.14	99.48	70.53	93.50	96.85	99.64

Locally Aware Enhanced Filter. Due to the flattening characteristic of vanilla Mamba's feature operations [21], it leads to pixel forgetting within local regions when handling 2D images. This induces an accumulation effect of pixel information loss, causing incoherence in key semantic information and severely impacting the correct understanding of the image. To address this, we previously introduced a local branch that uses local convolutions to enhance pixel similarity. However, we believe this improvement is still insufficient to fully solve the problem. Moreover, we found that Mamba requires the memory of long-sequence dependencies, which causes the number of hidden states in the state-space equations to accumulate significantly. This results in redundant information, where irrelevant information not only adds extra computational burden but also interferes with the representation learning of key features, preventing effective flow of crucial expert information. To address this, we developed the Locally Aware Enhanced Filter (LAEF) in the Visual State Space Model (VSSM), which employs a graceful channel routing and local enhancement mechanism to further resolve these challenges. Fig. 3 shows the architecture of LAEF. Specifically, given the input feature $X \in \mathbb{R}^{H\times W\times C}$, the input is first embedded into a lower-dimensional space through Point-Wise Convolution (PWConv), followed by a SiLU activation function to obtain X'. Then, X' is split along the channel dimension into two groups: one for local information perception and the other for retained information. The number of channels in the local information perception group is set to rC, while the number of channels in the retained information group is set to $(1-r)C$, where r is the partial channel ratio (the specific setting of r will be elaborated in the experiments section). In the local information perception group, the divided features are sequentially processed by PWConv and SiLU activation functions to enhance local information. Finally, the locally enhanced features are concatenated with the retained features, followed by channel shuffle. Formally, the above process is defined as follows:

$$X' = SiLU(PW(X)), \tag{10}$$

$$X_L = X'|_{:,:,[1,rC]}, \quad X_R = X'|_{:,:,[rC+1,C]}, \tag{11}$$

$$\hat{X} = Shuffle(SiLU(PW(X_L)), X_R). \tag{12}$$

C. Feature Modulation Interaction Aggregation Module

Although we capture sufficient representation information through the local and global branches, effectively integrating the information from both branches becomes a critical challenge. In fact, an intuitive observation is that there exists an uncertain knowledge gap between the convolution-based CNN local features and the SSM-based global features. Therefore, simply adding or concatenating these features does not fully exploit their potential. To address this, we introduce the

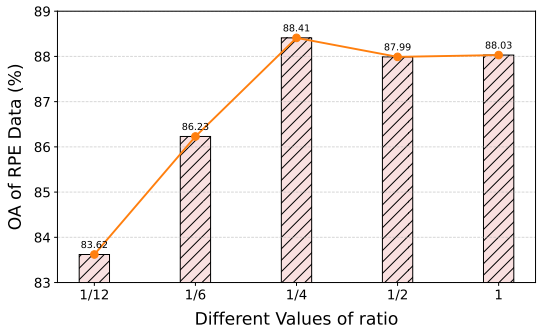

Fig. 4: Ablation analysis of different values of partial ratio.

TABLE II: Quantitative comparisons with different combinations of the LAEF and FMIAM.

LAEF	FMIAM	Params	RPE Data [7]			MHIST [10]		
			OA	Pre	AUC	OA	Pre	AUC
×	×	4.86	87.14	88.21	97.32	84.79	84.56	91.10
✓	×	4.30	87.87	89.24	98.28	86.21	85.99	89.51
×	✓	4.89	87.60	89.28	98.15	85.74	85.92	91.83
✓	✓	4.32	88.41	89.68	98.29	87.30	87.67	93.49

FMIAM, which operates between the two branches to reweight the features, enabling deep interaction and aggregation of the information. Specifically, we first compute the globally weighted features and locally weighted features, then modulate and reweight the features from both branches, and finally introduce a simplified channel attention mechanism [32] to reduce interference from irrelevant information and achieve key feature localization. The above process is expressed as follows:

$$W_L = Sigmoid(\varphi(PW(F_{Local}))), \tag{13}$$

$$W_G = Sigmoid(\varrho(PW(F_{Global}))), \tag{14}$$

$$W_{fusion} = CA([W_G \odot F_{Local}, W_L \odot F_{Global}]). \tag{15}$$

where Sigmoid$(\cdot)$ denotes the Sigmoid activation function, $\varphi(\cdot)$ represents the ReLU activation function, $\varrho(\cdot)$ represents the GELU activation function, and CA$(\cdot)$ denotes the channel attention mechanism.

III. Experiments

A. Experimental settings

Datasets. To validate the model's performance, we selected five public medical image datasets: the Retinal Pigment Epithelium Cell dataset (RPE Data) [7], minimalist histopathology image analysis dataset (MHIST) [10], the Malaria Cell Image dataset (SARS) [9], TissueMNIST [8], and MedFM-Colon [11]. The RPE dataset comprises 1,862 images of retinal pigment epithelium cells, classified into four categories. The SARS dataset includes 27,558 images of malaria cells, divided into two categories. MHIST contains 3,152 images of colorectal polyps for binary classification. The MedMF-Colon dataset classifies tumor tissues in colonoscopy pathology slides, consisting of 10,009 pathological tissue patches from 396 colon cancer patients. TissueMNIST features 236,386 images of human kidney cortex cells, categorized into eight groups. All datasets were split into training, validation, and test sets in a 6:2:2 ratio.

TABLE III: Effect of the parallel VSSM mechanism.

Parallel	Params	RPE Data [7]			MHIST [10]		
		OA	Pre	AUC	OA	Pre	AUC
×	8.83	87.60	88.71	98.28	86.05	85.83	91.30
✓	4.32	88.41	89.68	98.29	87.30	87.67	93.49

TABLE IV: Ablation studies on alternatives to the LAEF.

Method	Params	RPE Data [7]			MHIST [10]		
		OA	Pre	AUC	OA	Pre	AUC
Linear (baseline)	4.89	87.60	89.28	98.15	85.74	85.92	91.83
3×3 Conv	4.94	86.14	88.25	97.99	84.21	84.02	89.88
ConvGLU [33]	5.35	87.98	89.88	97.52	86.34	86.22	90.11
DFN [34]	5.27	87.25	88.69	97.68	85.02	84.97	90.08
LAEF (ours)	4.32	88.41	89.68	98.29	87.30	87.67	93.49

Training Details. We implemented our MambaMIC with PyTorch 2.0.0 and trained it on an NVIDIA RTX 3090, processing 200 epochs with a batch size of 16. We employed the Adam optimizer with an initial learning rate of 0.0001, a weight decay of 1e-4, and Cross-Entropy Loss to optimize the model parameters. Additionally, we incorporated a cosine annealing learning rate decay strategy and an early stopping strategy with a 10-epoch warm-up period during training.

B. Comparison with SOTA Models

To validate the effectiveness of MambaMIC, we compared it with state-of-the-art methods, including CNN-based approaches (ConvNext [22], RepViT [23], MobileOne [24]), Transformer-based approaches (PVT [25], MpViT [26], Twins [27], ViT [28]), hybrid CNN-Transformer methods (FastViT [29], SwiftFormer [19], Uniformer [30], SMT [20]), and Mamba-based methods (MedMamba [31], VMamba [21]). As shown in Table I and Fig. 1, MambaMIC achieves the best results in terms of model parameters, FLOPs, OA, and AUC across five datasets. Notably, our model performs exceptionally well on large datasets (e.g., TissueMNIST [8]) and achieves optimal performance on small datasets (e.g., RPE Data [7]). This highlights MambaMIC's capability to efficiently handle microscopic image recognition tasks without requiring extensive data or computational resources.

C. Ablation Study

In this section, we conduct ablation experiments on the RPE Data [7] and MHIST [10] to investigate the impact of individual components on the final performance. For a fair comparison, all ablation studies are performed under identical settings and training configurations.

Ablation experiments with different components. To validate the effectiveness of the proposed components, we conducted a detailed ablation study in Table II. The LAEF significantly improves accuracy while further reducing model parameters and computational complexity. This is achieved through the channel clearing mechanism, which alleviates information redundancy and enhances the local perceptual ability of the VSSM, mitigating the local pixel forgetting issue. Additionally, the FMIAM achieves deep information fusion and key expert information relocation with fewer parameters, leading to a notable performance improvement.

Ablation study of the parallel VSSM mechanism. In Table III, we further analyze the parallel VSSM mechanism, which reduces the parameter count by half while maintaining high performance. This is because the parallel mechanism not only effectively alleviates the computational burden caused by an excessive number of hidden states, but also promotes the interaction of channel context information.

Ablation study of LAEF. To further validate the effectiveness of LAEF, we replaced it with a Linear layer (baseline), 3 $\times$3 convolution, Convolutional Gated Linear Unit (ConvGLU) [33], and Depth-wise Convolution Equipped Feed-forward Network (DFN) [34]. The quantitative results, shown in Table IV, demonstrate that our LAEF achieves the best performance in both parameters and accuracy.

Ablation analysis of partial channel ratio. In LAEF, we enhance local perception by retaining only a subset of channels through channel dropping and pruning mechanisms. The choice of partial channel rate r is therefore crucial, and we further analyze its selection. The experimental results, shown in Fig. 4, reveal that when r is set too large or all channels are selected, no significant performance gain is observed. On the contrary, the redundancy of information increases the computational burden and introduces noise, which interferes with the accurate localization of key features. When r is set too small, valuable information is lost. Consequently, we explore the optimal ratio, and when r is set to 1/4 (the default setting for MambaMIC), it achieves the best trade-off between accuracy and speed.

IV. Conclusion

In this paper, we explore the power of Mamba in MIC and reconsider its limitations. Specifically, we design a Local-Global dual-branch architecture, the MambaMIC Block. The local branch uses convolutions to enhance perception, while the global branch employs VSSM to capture global dependencies, incorporating LAEF to reduce channel redundancy and pixel forgetting. Additionally, FMIAM recalibrates features from both branches for multi-class fusion and key feature re-localization. Extensive experiments show that MambaMIC outperforms state-of-the-art methods, providing a new strong baseline for MIC.

References

[1] Xiaoyou Ying and Thomas M Monticello, "Modern imaging technologies in toxicologic pathology: An overview," *Toxicologic pathology*, 2006.

[2] Fatima Merchant and Kenneth Castleman, *Microscope image processing*, Academic press, 2022.

[3] Zhichao Liu, Luhong Jin, Jincheng Chen, Qiuyu Fang, et al., "A survey on applications of deep learning in microscopy image analysis," *Computers in Biology and Medicine*, 2021.

[4] Satish Kumar, Tasleem Arif, Abdullah S Alotaibi, Majid B Malik, and Jatinder Manhas, "Advances towards automatic detection and classification of parasites microscopic images using deep convolutional neural network: methods, models and research directions," *Archives of Computational Methods in Engineering*, 2023.

[5] Long D Nguyen, Dongyun Lin, Zhiping Lin, and Jiuwen Cao, "Deep cnns for microscopic image classification by exploiting transfer learning and feature concatenation," *ISCAS*, 2018.

[6] A Vaswani et al., "Attention is all you need," in *NeurIPS*, 2017.

[7] Loris Nanni, Michelangelo Paci, et al., "Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium," *PLoS One*, 2016.

[8] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, et al., "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, 2023.

[9] Hang Yu, Fayad O Mohammed, Muzamil Abdel Hamid, Feng Yang, Yasmin M Kassim, Abdelrahim O Mohamed, Richard J Maude, et al., "Patient-level performance evaluation of a smartphone-based malaria diagnostic application," *Malaria Journal*, 2023.

[10] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, et al., "A petri dish for histopathology image analysis," in *AIM*, 2021.

[11] Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, et al., "A real-world dataset and benchmark for foundation model adaptation in medical image classification," *Scientific Data*, 2023.

[12] Neha Sengar, Radim Burget, and Malay Kishore Dutta, "A vision transformer based approach for analysis of plasmodium vivax life cycle for malaria prediction using thin blood smear microscopic images," *Computer Methods and Programs in Biomedicine*, 2022.

[13] Mohamad Abou Ali, Fadi Dornaika, and Ignacio Arganda-Carreras, "White blood cell classification: Convolutional neural network (cnn) and vision transformer (vit) under medical microscope," *Algorithms*, 2023.

[14] Mehedi Masud, Hesham Alhumyani, et al., "Leveraging deep learning techniques for malaria parasite detection using mobile application," *Wireless Communications and Mobile Computing*, 2020.

[15] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, et al., "Maxvit: Multi-axis vision transformer," *ECCV*, 2022.

[16] Tao Huang, Lang Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu, "Lightvit: Towards light-weight convolution-free vision transformers," *arXiv preprint arXiv:2207.05557*, 2022.

[17] Ankit Gupta, Albert Gu, and Jonathan Berant, "Diagonal state spaces are as effective as structured state spaces," *NeurIPS*, 2022.

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

[19] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan, "Swiftformer: Efficient additive attention for transformer-based real-time mobile vision applications," in *ICCV*, 2023.

[20] Weifeng Lin, Ziheng Wu, et al., "Scale-aware modulation meet transformer," *arxiv preprint arxiv:2307.08579*, 2023.

[21] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[22] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *CVPR*, 2022.

[23] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding, "Repvit: Revisiting mobile cnn from vit perspective," in *CVPR*, 2024.

[24] Pavan Kumar Anasosalu Vasu et al., "An improved one millisecond mobile backbone," *arXiv preprint arXiv:2206.04040*, 2022.

[25] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao, "Pvtv2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, 2022.

[26] Youngwan Lee, Jonghee Kim, et al., "Mpvit: Multi-path vision transformer for dense prediction," in *CVPR*, 2022.

[27] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in *NeurIPS*, 2021.

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan, "Fastvit: A fast hybrid vision transformer using structural reparameterization," in *ICCV*, 2023.

[30] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, et al., "Uniformer: Unifying convolution and self-attention for visual recognition," *arxiv preprint arxiv:2201.09450*, 2022.

[31] Yubiao Yue and Zhenzhang Li, "Medmamba: Vision mamba for medical image classification," *arXiv preprint arXiv:2403.03849*, 2024.

[32] Qilong Wang, Banggu Wu, et al., "Eca-net: Efficient channel attention for deep convolutional neural networks," in *CVPR*, 2020.

[33] Dai Shi, "Transnext: Robust foveal visual perception for vision transformers," in *CVPR*, 2024.

[34] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool, "Localvit: Bringing locality to vision transformers," *arXiv preprint*, 2021.