Bridging Local Inductive Bias and Long-Range Dependencies with Pixel-Mamba for End-to-end Whole Slide Image Analysis

Supplementary Material

In this supplementary material, we provide more information that can not be included in the manuscript's main text due to the page limitation. In Section 1, the detailed network configurations of Pixel-Mamba are introduced. In Section 2, the implemented details of pre-training Pixel-Mamba on the ImageNet-1K dataset and fine-tuning Pixel-Mamba on the downstream tasks of pathological images are provided. In Section 3 and 4, we provide the ablation study of hyper-parameter α on BLCA dataset and more experiments on more datasets and baseline Methods.

1. Network Configurations of Pixel-Mamba

The detailed network configurations of Pixel-Mamba-6M and Pixel-Mamba-21M are provided in Table 1 and Table 2, respectively.

Take Pixel-Mamba-6M for example, it consists of 24 Pixel-Mamba layers, each composed of three key components: Mamba Block, Region Fusion (RF), and Token Expansion (TE). Throughout the forward process, the Mamba Block models global context among tokens to capture long-range dependencies. The Region Fusion module identifies similar regions at each level and merges them to reduce redundancy, improving memory efficiency. Finally, the Token Expansion module enlarges the token receptive field, progressively increasing it from 1 × 1 to 32 × 32, which is crucial for learning hierarchical representations. With the expansion of the token receptive field, the feature channels of tokens are increasing from 3 to 384. Pixel-Mamba includes horizontal TE and vertical TE. In each one, the token expansion operation then increases the receptive field by concatenating or averaging adjacent tokens along the horizontal or vertical dimensions.

Pixel-Mamba-21M also consists of 24 Pixel-Mamba layers. The difference is that its token feature channel increases to 768 throughout the forward process.

Table 1. The detailed network configurations of Pixel-Mamba-6M. In the column of Pixel-Mamba Layers, Mamba represents the Mamba Block, RF represents the Region Fusion, and TE represents Token Expansion. The Token column indicates the receptive field of tokens in this layer.

Pixel-Mamba-6M			
Layer id	Token	Channels	Pixel-Mamba Layers
1	1×1	3	Mamba + RF + Horizontal TE-Cat
2	1×2	6	Mamba + RF + Vertical TE-Cat
3	2×2	12	Mamba + RF + Horizontal TE-Cat
4	2×4	24	Mamba + RF + Vertical TE-Cat
5	4×4	48	Mamba + RF
6	4×4	48	Mamba + RF + Horizontal TE-Cat
7	4×8	96	Mamba + RF
8	4×8	96	Mamba + RF + Vertical TE-Cat
9	8×8	192	Mamba + RF
10	8×8	192	Mamba + RF
11	8×8	192	Mamba + RF + Horizontal TE-Avg
12	8×16	192	Mamba + RF
13	8×16	192	Mamba + RF
14	8×16	192	Mamba + RF + Vertical TE-Avg
15	16×16	192	Mamba + RF
16	16×16	192	Mamba + RF
17	16×16	192	Mamba + RF
18	16×16	192	Mamba + RF
19	16×16	192	Mamba + RF
20	16×16	192	Mamba + RF + Horizontal TE-Avg
21	16×32	192	Mamba + RF
22	16×32	192	Mamba + RF + Vertical TE-Cat
23	32×32	384	Mamba + RF
24	32×32	384	Mamba + RF

2. Implemental Details

Pre-training: Pixel-Mamba is pre-trained on the ImageNet [1] to ensure model convergence, utilizing over 1.28 million natural images for a classification task encompassing 1,000 categories. Following [2], Pixel-Mamba is trained on images of size 224×224. The optimizer is AdamW with a momentum of 0.9, batch size of 1024, and weight decay of 0.05. Pixel-Mamba is trained for 300 epochs using a cosine learning rate schedule, starting with an initial learning rate of 0.001. Data augmentation techniques include random cropping, random horizontal flipping, label-smoothing regularization, mixup, and random erasing. 8 NVIDIA A100 GPUs are used for pre-training, costing approximately 2 to 3 days.

Fine-tuning on Downstream Tasks: Pixel-Mamba serves as the backbone network for all downstream tasks involving pathology images, including tumor staging and survival analysis. During fine-tuning, a task-specific head is added to the backbone network, and the entire network

Table 2. The detailed network configurations of Pixel-Mamba-21M. In the column of Pixel-Mamba Layers, Mamba represents the Mamba Block, RF represents the Region Fusion, and TE represents Token Expansion. The Token column indicates the receptive field of tokens in this layer.

Pixel-Mamba-21M			
Layer id	Token	Channels	Pixel-Mamba Layers
1	1×1	3	Mamba + RF + Horizontal TE-Cat
2	1×2	6	Mamba + RF + Vertical TE-Cat
3	2×2	12	Mamba + RF + Horizontal TE-Cat
4	2×4	24	Mamba + RF + Vertical TE-Cat
5	4×4	48	Mamba + RF
6	4×4	48	Mamba + RF + Horizontal TE-Cat
7	4×8	96	Mamba + RF
8	4×8	96	Mamba + RF + Vertical TE-Cat
9	8×8	192	Mamba + RF
10	8×8	192	Mamba + RF
11	8×8	192	Mamba + RF + Horizontal TE-Cat
12	8×16	384	Mamba + RF
13	8×16	384	Mamba + RF
14	8×16	384	Mamba + RF + Vertical TE-Avg
15	16×16	384	Mamba + RF
16	16×16	384	Mamba + RF
17	16×16	384	Mamba + RF
18	16×16	384	Mamba + RF
19	16×16	384	Mamba + RF
20	16×16	384	Mamba + RF + Horizontal TE-Avg
21	16×32	384	Mamba + RF
22	16×32	384	Mamba + RF + Vertical TE-Cat
23	32×32	768	Mamba + RF
24	32×32	768	Mamba + RF

is fine-tuned on WSIs at a $2.5\times$ magnification over 100 epochs. The AdamW optimizer is employed alongside a cosine learning rate schedule, starting with an initial learning rate of 0.0004. 8 NVIDIA A100 GPUs are used. In each iteration, one GPU is assigned to process one WSI. Gradient accumulation is performed every 8 iterations, resulting in an effective batch size of 64.

3. The ablation of Region Fusion

In the Region Fusion module of Pixel-Mamba, the number of merged regions, k, is defined as $\lceil \alpha * n/L \rceil$, where $0 < \alpha < 1$ is a hyper-parameter controlling the retention rate of regions for the final output, and L is the total number of layers in the network.

We conduct the ablation study of hyper-parameter α and the results are reported in Table 3. Pixel-Mamba-Surv achieves the best C-index of 0.6507 with the $\alpha = 0.8$ on the BLCA dataset. Thus, we suggest $\alpha = 0.8$, and all results of experiments in the main text of the manuscript are obtained with $\alpha = 0.8$.

Table 3. The ablation study of α in Region Fusion.

α	0.4	0.6	0.8	1.0
C-index	0.6176	0.6260	0.6507	0.6104

Table 4. Metastasis Detection Results on CAMELYON16

Method	AUC	ACC
R50+ABMIL	0.8708 ± 0.0429	0.8597 ± 0.0297
R50+MambaMIL	0.8812 ± 0.0162	0.7693 ± 0.0456
CONCH+ABMIL	0.9853 ± 0.0101	0.9750 ± 0.0111
CONCH+MambaMIL	0.9850 ± 0.0151	0.9650 ± 0.0122
Pixel-Mamba	0.8923 ± 0.0418	0.8793 ± 0.0437

4. Experiments on More Datasets and Baseline Methods

To provide a comprehensive evaluation of Pixel-Mamba, we test it on the CAMELYON16 metastasis detection dataset. For the baseline methods, we adopt their default settings. For Pixel-Mamba, we randomly sample 100 regions of size 2048×2048 at 20× magnification per image and aggregate the results via a voting ensemble (see Table 4). Although classification tasks typically favor large-scale pre-trained VLMs like CONCH, our model performs well—thanks to its larger receptive field and end-to-end training—significantly outperforming two-stage approaches that use a ResNet-50 encoder. This experiment demonstrates that for tasks requiring finer details at higher magnifications, Pixel-Mamba can effectively process WSIs by simply slicing them into a few large regions and applying a voting-based ensemble, rather than relying on neural networks to aggregate features from numerous patches as in two-stage methods.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[2] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 1