Mamba Models in Vision: A Brief Survey

Shaofeng Yuan[1] SHAOFENG.YUAN.SMU@GMAIL.COM
[1] *Independent Researcher*

Abstract

Selective state space models (Selective SSMs, or Mamba), a class of state space models, represent a recent emerging topic in computer vision, demonstrating remarkable results in the area of global modeling with linear complexity. In this survey, we provide a comprehensive review of articles on Mamba models applied in perceptual vision domains, including visual recognition, and medical image computing.

Keywords: Selection mechanism, global modeling, linear complexity, visual recognition, medical image computing.

1. Introduction

Recently, State Space Models (SSMs) (Gu, 2023) have attracted considerable interest among artificial intelligence (AI) researchers and practitioners. Building on the foundation of classical SSMs (Kalman, 1960) in control theory, the modern SSMs (e.g., Mamba (Gu and Dao, 2023) or Selective SSMs or S6) not only establish long-distance dependencies, but also exhibit linear complexity with respect to input size. The concept of the SSMs was initially introduced in the S4 (Gu et al., 2022) model (Structured SSMs or S4), presenting a distinctive architecture capable of effectively modeling global information in comparison to conventional ConvNets or Transformer architectures. Building upon S4, the S5 (Smith et al., 2023) model (Simplified S4) emerged, strategically reducing complexity to a linear level. Mamba, in turn, introduced an input-adaptive mechanism to enhance the previous SSMs, resulting in higher inference speed, throughput, and overall metrics compared to Transformers of equivalent scale.

Several latest studies have preliminarily explored the effectiveness of Mamba in the vision domain (Ma et al., 2024; Zhu et al., 2024). In addition, some works extend Mamba to non-vision and non-language field for graph modeling (Wang et al., 2024a; Behrouz and Hashemi, 2024). In this brief survey, we provide a timely literature review of Mamba models applied in computer vision, aiming to provide a fast understanding of the generic sequence modeling framework to our readers.

2. A Categorization of Mamba Models

We categorize Mamba models into a multi-perspective taxonomy considering different criteria of separation. Perhaps the most important criteria to separate the models are defined by (1) the task they are applied to, and (2) the datasets used during training and evaluation. Our categorization of Mamba models according to the criteria enumerated above is present in Table 1.

Table 1: Our multi-perspective categorization of Mamba models applied in computer vision.

Paper	**Task**	**Dataset**
Ma et al. (2024)	Medical image segmentation	Abdomen CT: MICCAI 2022 FLARE, Abdomen MR: MICCAI 2022 AMOS, Endoscopy: MICCAI 2017 EndoVis, Microscopy: NeurIPS 2022 Cell Segmentation
Zhu et al. (2024)	Visual representation learning	ImageNet-1K classification, ADE20K semantic segmentation, COCO object detection, COCO instance segmentation
Liu et al. (2024b)	Visual representation learning	ImageNet-1K classification, ADE20K semantic segmentation, COCO object detection
Xing et al. (2024)	Medical image segmentation	Brain tumor MR: MICCAI 2023 BraTS
Guo et al. (2024)	Medical image registration	SynthRAD2023
Yang et al. (2024)	Medical video segmentation	Polpy: Kvasir, CVC-300, CVC-612, ASU-Mayo, Breast US: SIAT2022(Li et al., 2022)
Ruan and Xiang (2024)	Medical image segmentation	Skin: ISIC17, ISIC18, Abdomen CT: Synapse
Liu et al. (2024a)	Medical image segmentation	Abdomen MR: MICCAI 2022 AMOS, Endoscopy: MICCAI 2017 EndoVis, Microscopy: NeurIPS 2022 Cell Segmentation
Gong et al. (2024)	Medical image segmentation, Medical image classification, Anatomical landmark detection	Brain tumor MR: MICCAI 2023 BraTS GIL track, Alzheimer's Disease: ADNI, Fetal brain: Private dataset
Zheng and Wu (2024)	Single image dehazing, Low light enhancement, Single image deraining	Dehazing: RESIDE, Enhancement: LOL, MIT-Adobe FiveK, Dereining: Rain13K
Wang et al. (2024b)	Medical image segmentation	Cardiac MR: MICCAI 2017 ADDC
Li et al. (2024)	Image classification, Action recognition, Weather forecasting	Image: ImageNet-1K, Video: HMDB-51, Climate: ERA5
Zheng and Zhang (2024)	Medical image enhancement	Polpy: Kvasir
Wang and Ma (2024a)	Medical image segmentation	Cardiac MR: MICCAI 2017 ADDC
Ye and Chen (2024)	Medical image segmentation	Cardiac US: Stanford2023(Reddy et al., 2023)
Liang et al. (2024)	Point cloud analysis	Object classification: ScanObjectNN, Part segmentation: ShapeNetPart
Wang and Ma (2024b)	Medical image segmentation	Cardiac MR: MICCAI 2017 ADDC
He et al. (2024)	Remote sensing image analysis	WorldView-II, WorldView-III

2.1. Medical Image Segmentation

Ma et al. (Ma et al., 2024) introduce **U-Mamba** to address the challenges in modeling long-range dependencies dute to the inherent locality of ConvNets and the computational complexity of ViTs. U-Mamba is designed simultaneously extract multi-scale local features and capture long-distance dependencies, outperforming existing ConvNet- and Transformer-based segmentation networks. [U-Mamba@2401.04722]

Xing et al. (Xing et al., 2024) introduce **SegMamba**, a novel 3D medical image segmentation Mamba model, designed to effectively capture long-range dependencies within whole volume features at various scale by combining the U-shape structure with the Mamba. [SegMamba@2401.13560]

Ruan and Xiang (Ruan and Xiang, 2024) propose a U-shape architecture model, named **Vision Mamba UNet (VM-UNet)**, for medical image segmentation. [VM-UNet@2402.02491]

Liu et al. (Liu et al., 2024a) propose a Mamba-based network, i.e., **Swin-UMamba**, for 2D medical image segmentation. It uses a generic encoder to integrate the power of the pretrained vision model with a well-designed decoder for medical image segmentation tasks. This study reveals the impact of ImageNet-based pretraining for Mamba-based models. [Swin-UMamba@2402.03302]

Gong et al. (Gong et al., 2024) explore the integration of Mamba blocks within ConvNets to enhance long-range dependency modeling. They propose **nnMamba**, a various of structure for both segmentation, classification, and landmark detection. [nnMamba@2402.03526]

Wang et al. (Wang et al., 2024b) propose leveraging Visual Mamba blocks (VSS) within the U-Net architecture to improve long-range dependency modeling in medical image analysis, resulting in **Mamba-UNet**. Mamba-UNet is motivated by UNet and Swin-UNet. [Mamba-UNet@2402.05079]

Wang and Ma (Wang and Ma, 2024a) introduce the **Semi-Mamba-UNet**, a novel framework integrating the Mamba architecture within a pixel-level contrastive, cross-supervised learning for semi-supervised medical image segmentation. [Semi-Mamba-UNet@2402.07245]

Ye and Chen (Ye and Chen, 2024) introduce the **P-Mamba** for efficient pediatric echocardiographic left ventricular segmentation. P-Mamba has two encoder branches. The one is the Vision Mamba encoder, aiming to improve the computing and memory efficiency while modeling global dependencies, the other is the DWT-based Perona-Malik Diffusion (PMD) encoder for noise suppression while preserving the local shape cues of the left ventricle. [P-Mamba-UNet@2402.08506]

Wang and Ma (Wang and Ma, 2024b) introduce the **Weak-Mamba-UNet**, an innovative weakly-supervised learning framework that leverages the capabilities of ConvNet, ViT, and the cutting-edge Visual Mamba (VMamba) architecture for medical image segmentation, especially when dealing with scribble-based annotations. [Weak-Mamba-UNet@2402.10887]

2.2. Medical Video Segmentation

Yang et al. (Yang et al., 2024) present a novel framework, named **Vivim**, that integrates Mamba into the multi-level transformer architecture to transform a video clip into one feature sequence containing spatiotemporal information at each scale. [Vivim@2401.14168]

2.3. Visual Representation Learning

Zhu et al. (Zhu et al., 2024) propose a new generic vision backbone with bidirectional Mamba block, i.e., **Vision Mamba (Vim)**. Vim marks the image sequences with position embeddings and compresses the visual representation with bidirectional state space models. [Vision Mamba@2401.09417]

Liu et al. (Liu et al., 2024b) introduce the **Visual State Space Model (VMamba)** with global receptive fields and dynamic weights for efficient visual representation learning. [VMamba@2401.10166]

2.4. Natural Image Processing

including restoration.

Zheng and Wu (Zheng and Wu, 2024) present a **U-shaped Vision Mamba (UVM-Net)**, which forms a deep network basd on U-Net structure with both local capture capability and efficient long-rang modeling. The Bi-SSM module in UVM-Net scroll the feature maps over the channel domain to fully utilize the long-range modeling capability of SSM. [UVM-Net@2402.04139]

2.5. Medical Image Processing

including restoration and registration.

Zheng and Zhang (Zheng and Zhang, 2024) present a frequency-domain based network, called **FD-Vision Mamba (FDVM-Net)**, which achieves high-quality image exposure correction by reconstructing the frequency domain of endoscopic images. [FDVM-Net@2402.06378]

Guo et al. (Guo et al., 2024) introduce **MambaMorph**, an innovative multi-modality deformable registration network, specifically designed for Magnetic Resonance (MR) and Computed Tomography (CT) image alignment. [MambaMorph@2401.13934]

2.6. Multi-Dimensional Data Analysis

Li et al. (Li et al., 2024) present **Mamba-ND**, a generalized design extending the Mamba architecture to arbitrary multi-dimensional data. [Mamba-ND@2402.05892]

2.7. Point Cloud Analysis

Liang et al. (Liang et al., 2024) introduce the **Point State Space Model (PointMamba)**, which has global receptive fields with linear complexity. [PointMamba@2402.10739]

2.8. Remote Sensing Image Analysis

He et al. (He et al., 2024) introduce **Pan-Mamba**, a pan-sharpening network that leverages Mamba as the core module. Mamba is utilized for global information modeling, extracting global information from both high-resolution texture-rich panchromatic (PAN) and low-resolution multi-spectral (LRMS) images. [Pan-Mamba@2402.12192]

References

Ali Behrouz and Farnoosh Hashemi. Graph mamba: Towards learning on graphs with state space models. *arXiv preprint arXiv:2402.08678*, 2024.

Haifan Gong, Luoyan Kang, Yitao Wang, Xiang Wan, and Haofeng Li. nnMamba: 3D biomedical image segmentation, classification and lankmark detection with state space model. *arXiv preprint arXiv: 2402.03526*, 2024.

Albert Gu. *Modeling Sequences with Structured State Spaces*. PhD thesis, Stanford University, 2023.

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

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Tao Guo, Yinuo Wang, and Cai Meng. Mambamorph: a mamba-based backbone with contrastive feature learning for deformable mr-ct registration. *arXiv preprint arXiv:2401.13934*, 2024.

Xuanhua He, Ke Cao, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Pan-Mamba: Effective pan-sharpening with state space model. *arXiv preprint arXiv: 2402.12192*, 2024.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

Jialu Li, Qingqing Zheng, Mingshuang Li, Ping Liu, Qiong Wang, Litao Sun, and Lei Zhu. Rethinking breast lesion segmentation in ultrasound: A new video dataset and a baseline network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 391–400. Springer, 2022.

Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-ND: Selective state space modeling for multi-dimensional data. *arXiv preprint arXiv:2402.05892*, 2024.

Dingkang Liang, Xin Zhou, Xinyu Wang, Xingkui Zhu, Wei Xu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv: 2402.10739*, 2024.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, and Shanshan Wang. Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining. *arXiv preprint arXiv: 2402.03302*, 2024a.

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*, 2024b.

Jun Ma, Feifei Li, and Bo Wang. U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv: 2401.04722*, 2024.

Charitha D Reddy, Leo Lopez, David Ouyang, James Y Zou, and Bryan He. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. *Journal of the American Society of Echocardiography*, 36(5):482–489, 2023.

Jiacheng Ruan and Suncheng Xiang. VM-UNet: Vision Mamba UNet for medical image segmentation. *arXiv preprint arXiv: 2402.02491*, 2024.

Jimmy TH Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023.

Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.00789*, 2024a.

Ziyang Wang and Chao Ma. Semi-Mamba-UNet: Pixel-level contrastive cross-supervised visual Mamba-based UNet for semi-supervised medical image segmentation. *arXiv preprint arXiv: 2402.07245*, 2024a.

Ziyang Wang and Chao Ma. Weak-Mamba-UNet: Visual Mamba makes cnn and vit work better for scribble-based medical image segmentation. *arXiv preprint arXiv: 2402.10887*, 2024b.

Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-UNet: UNet-like pure visual Mamba for medical image segmentation. *arXiv preprint arXiv: 2402.05079*, 2024b.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range sequential modeling Mamba for 3D medical image segmentation. *arXiv preprint arXiv: 2401.13560*, 2024.

Yijun Yang, Zhaohu Xing, and Lei Zhu. Vivim: a video vision Mamba for medical video object segmentation. *arXiv preprint arXiv: 2401.14168*, 2024.

Zi Ye and Tianxiang Chen. P-Mamba: Marrying Perona Malik diffusion with Mamba for efficient pediatric echocardiographic left ventricular segmentation. *arXiv preprint arXiv: 2402.08506*, 2024.

Zhuoran Zheng and Chen Wu. U-shaped vision mamba for single image dehazing. *arXiv preprint arXiv: 2402.04139*, 2024.

Zhuoran Zheng and Jun Zhang. Fd-vision mamba for endoscopic exposure correction. *arXiv preprint arXiv: 2402.06378*, 2024.

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.