A Survey on Visual Mamba

Hanwei Zhang[1,4,5], Ying Zhu[2], Dan Wang[2], Lijun Zhang[1], Tianxiang Chen[3], and Zi Ye[4]

[1] Automotive Software Innovation Center, Chongqing 401331, China
zhang@depend.uni-saarland.de
zhanglj@ios.ac.cn
[2] Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
zhuying23@mails.ucas.ac.cn
wangdan233@mails.ucas.ac.cn
[3] University of Science and Technology of China
txchen@mail.ustc.edu.cn
[4] Institute of Intelligent Software, Guangzhou, China
yezi1022@gmail.com
[5] Saarland University, Germany

Abstract. State space models (SSMs) with selection mechanisms and hardware-aware architectures, namely Mamba, have recently demonstrated significant promise in long-sequence modeling. Since the self-attention mechanism in transformers has quadratic complexity with image size and increasing computational demands, the researchers are now exploring how to adapt Mamba for computer vision tasks. This paper is the first comprehensive survey aiming to provide an in-depth analysis of Mamba models in the field of computer vision. It begins by exploring the foundational concepts contributing to Mamba's success, including the state space model framework, selection mechanisms, and hardware-aware design. Next, we review these vision mamba models by categorizing them into foundational ones and enhancing them with techniques such as convolution, recurrence, and attention to improve their sophistication. We further delve into the widespread applications of Mamba in vision tasks, which include their use as a backbone in various levels of vision processing. This encompasses general visual tasks, Medical visual tasks (e.g., 2D / 3D segmentation, classification, and image registration, etc.), and Remote Sensing visual tasks. We specially introduce general visual tasks from two levels: High/Mid-level vision (e.g., Object detection, Segmentation, Video classification, etc.) and Low-level vision (e.g., Image super-resolution, Image restoration, Visual generation, etc.). We hope this endeavor will spark additional interest within the community to address current challenges and further apply Mamba models in computer vision.

Keywords: Mamba · Computer vision · State space model · Application.

1 Introduction

Deep Neural Networks (DNNs) have demonstrated remarkable performance across various artificial intelligence (AI) tasks, with the fundamental architecture playing a crucial role in determining the model's capabilities. Traditional neural networks typically comprise Multi-Layer Perceptron (MLP) or Fully Connected (FC) layers [1,2]. Convolutional neural networks (CNNs) [3,4] introduce convolutional and pooling layers, which are particularly effective for processing shift-invariant data like images. Recurrent neural networks (RNNs) [5,6] utilize recurrent cells to handle sequential or time series data. To address the issue of CNN, RNN, and GNN models only capturing local relationships, the Transformer [7,8,9], introduced in 2017, excels at learning long-distance feature representations. Transformers primarily rely on attention-based attention mechanisms, *e.g.* self-attention and cross-attention, to extract intrinsic features and improve their representation capability. Pre-trained massive transformer-based models, such as GPT-3 [10], deliver robust performance across various NLP datasets, excelling in natural language understanding and generation tasks. The remarkable performance of transformer-based models has propelled their widespread adoption in vision applications. The core of transformer models is their exceptional skill in capturing long-range dependencies and maximizing the use of large datasets. The feature extraction module is the main component of vision transformer architectures. It processes data using a sequence of self-attention blocks, significantly improving its capacity to analyze images.

However, a primary obstacle for Transformers is the substantial computational demands of the self-attention mechanism, which increases quadratically with image resolution. The Softmax operation within the attention blocks further intensifies the computational demands, presenting significant challenges for implementing these models on edge and low-resource devices. Additionally, real-time computer vision systems utilizing transformer-based models must adhere to stringent low-latency standards to maintain a high-quality user experience. This scenario highlights the continuous evolution of new architectures to enhance performance, although this often comes with the trade-off of higher computational demands. Many new models based on sparse attention mechanisms or innovative neural network paradigms have been proposed to reduce computational costs further while capturing long-range dependencies and maintaining high performance. State space models (SSMs) have emerged as a central focus among these developments. As shown in Fig. 1(a), the number of publications related to SSMs demonstrates an explosive growth trend. Initially devised to simulate dynamic systems in areas such as control theory and computational neuroscience using state variables, SSMs predominantly describe linear invariant (or stable) systems when adapted for deep learning.
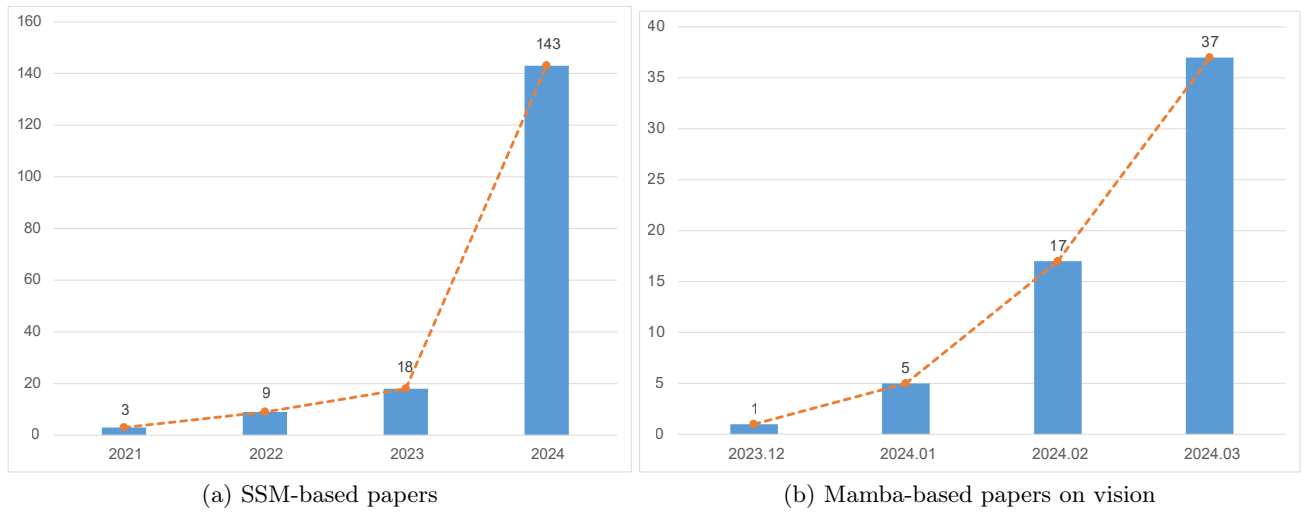

(a) SSM-based papers   (b) Mamba-based papers on vision

Fig. 1. The number of SSMs and Mamba papers released to date(from year 2021 to year 2024.03).

As SSMs have evolved, a new class of selective state space models, termed Mamba [11]. It has advanced the modeling of discrete data, such as text, with state-space models (SSMs) through two key improvements. Firstly, it features an input-dependent mechanism that adjusts SSM parameters dynamically, enhancing information filtering. Secondly, Mamba uses a hardware-aware algorithm that processes data linearly with sequence length, boosting computational speed on modern systems. Inspired by Mamba's achievements in language modeling, several initiatives are now aiming to adapt this success to the field of vision. Several studies have explored its integration with Mixture-of-Experts (MoE) techniques, as evidenced by works like Jamba [12], MoE-Mamba [13], and BlackMamba [14], outperformed the state-of-the-art architecture Transformer-MoE with fewer training steps. As illustrated in Fig. 1(b), since the release of Mamba in December 2023, the number of research papers focusing on Mamba in the vision domain has rapidly increased, reaching a peak in March 2024. This trend suggests that Mamba is emerging as a prominent research area in vision, potentially offering a viable alternative to Transformers. Therefore, A review of current related works is necessary and timely to provide a detailed overview of this new methodology in this evolving field.

Consequently, we present a comprehensive overview of how Mamba models are used in the vision domain. This paper aims to serve as a guide for researchers looking to delve deeper into this area. The critical contributions of our work include:

- This survey paper is the first to provide a comprehensive review of the Mamba technique in the vision domain, explicitly focusing on analyzing the proposed strategies.
- Expanding upon the Naive-based Mamba visual framework, we have investigated how Mamba's capabilities can be enhanced and combined with other architectures to achieve superior performance.

– We offer an in-depth exploration by organizing the literature based on various application tasks. We establish a taxonomy, identify advancements specific to each task, and offer insights on overcoming challenges.

The remainder of the survey is structured as follows: Section 2 examines the general and mathematical concepts underlying Mamba strategies. Section 3 discusses the naive Mamba visual models and how they integrate with other technologies to enhance performance, as proposed in recent years. Section 4 explores the application of Mamba technologies in addressing various computer vision tasks. Finally, Section 5 concludes the survey.

2 Formulation of Mamba

Mamba [11] is initially introduced in the domain of natural language processing. As shown in Fig. 2, the original Mamba Block integrates a Gated MLP into the State Space Model (SSM) architecture of H3 [15], utilizing an SSM sandwiched between two gated connections alongside a standard local convolution. For σ SiLU [16] or Swish activation function [17] is used. The Mamba architecture consists of repeating of Mamba block interleaved with standard normalization and residual connections. An optional normalization layer (LayerNorm chosen by original Mamba) is used in a similar location as RetNet [18].
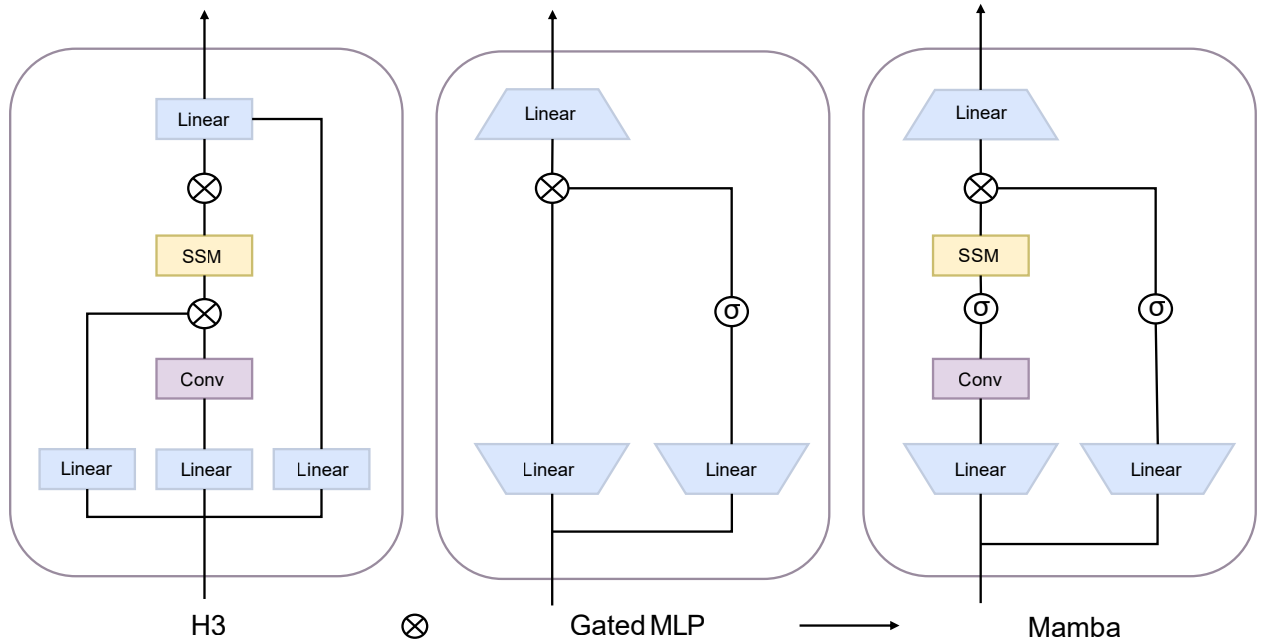

Fig. 2. Mamba Block [11].

2.1 State Space Models (SSMs)

Consider a structured state space model (SSM) that maps one-dimensional sequence $x(t) \in \mathbb{R}^L$ to $y(t) \in \mathbb{R}^L$ through a hidden state $h(t) \in \mathbb{R}^N$. With the evolution parameter $\mathbf{A} \in \mathbb{R}^{N \times N}$ and the projection parameters $\mathbf{B} \in \mathbb{R}^{N \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times N}$, such a model is formulated as linear ordinary differential equations

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t),$$
$$y(t) = \mathbf{C}h(t). \tag{1}$$

Discretization. To adapt for deep learning, State Space Models (SSMs), as continuous-time models, are discretized with a Zero-Order Hold(ZOH) assumption. Thus, the continuous-time parameters $\mathbf{A}, \mathbf{B}$ are transform to their discretized counterparts $\overline{\mathbf{A}}, \overline{\mathbf{B}}$ with a timescale parameter Δ according to

$$
\begin{aligned}
\overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\
\overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}.
\end{aligned}
\tag{2}
$$

Thus, (1) can be rewritten as

$$
\begin{aligned}
h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\
y_t &= \mathbf{C}h_t.
\end{aligned}
\tag{3}
$$

To enhance computational efficiency and scalability, the iterative process in (3) can be synthesized through a global convolution

$$
\begin{aligned}
\overline{\mathbf{K}} &= (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}}\overline{\mathbf{B}}, \cdots, \overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \\
\mathbf{y} &= \mathbf{x} * \overline{\mathbf{K}},
\end{aligned}
\tag{4}
$$

where L is the length of the input sequence $\mathbf{x}$, $\overline{\mathbf{K}} \in \mathbb{R}^L$ serves as the kernel of the SSMs and $*$ represents the convolution operation.

Architectures. SSMs often serve as independent sequence transformations that can be integrated into end-to-end neural network architectures. Here we introduce several fundamental architectures. Linear attention [19] approximates self-attention with a recurrence mechanism as a simplified form of linear SSM. H3 [15], as illustrated in Fig. 2, places an SSM between two gated connections and inserts a standard local convolution before it. Following H3, Hyena [20], replaces the SSM layer with an MLP-parameterized global convolution [21]. RetNet [22] introduces an extra gate and uses simpler SSM. RetNet enables an alternative parallelizable computation path and employs a variant of multi-head attention (MHA) instead of convolutions. Inspired by attention-free Transformer [23], the recent RNN design RWKV [24], can be interpreted as the ratio of two SSMs due to its primary "WKV" mechanism involving Linear Time Invariance (LTI) recurrences.

Selective State Space Models. Traditional SSMs demonstrated linear time complexity but their representativity of sequence context is inherently limited by time-invariant parameterization. To overcome this constraint, Selective State Space Models introduce *selective scan* for interactions among sequential states with

$$
\begin{aligned}
\mathbf{B} &= S_{\mathbf{B}}(\mathbf{x}), \\
\mathbf{C} &= S_{\mathbf{C}}(\mathbf{x}), \\
\Delta &= \tau_{\Delta}(\Delta + S_{\Delta}(\mathbf{x})),
\end{aligned}
\tag{5}
$$

before (2,3), so that parameters $\mathbf{B} \in \mathbb{R}^{B \times L \times N}$, $\mathbf{C}^{B \times L \times N}$ and $\Delta^{B \times L \times D}$ are dependent on the input sequence $\mathbf{x} \in \mathbb{R}^{B \times L \times D}$, where B represents the batch size, and D represents number of channels. Normally, S_B and S_C are linear parameterized projections to dimension N, *i.e.* $Linear_N(\cdot)$, while $S_{\Delta}(\mathbf{x}) = Broadcast_D(Linear_1(\mathbf{x}))$ and $\tau_{\Delta} = softplus$. The choice of S_{Δ} and τ_{Δ} is due to a connection to RNN gating mechanisms explained later.

2.2   Other Key Concepts in Mamba

Selection Mechanism. The connection between RNN gating and the discretization of continuous-time systems is well-established [25]. The classical gating mechanism of RNNs is an instance of the selection mechanism for SSMs. When $N = 1, \mathbf{A} = -1, \mathbf{B} = 1, S_{\Delta} = Linear(\mathbf{x})$ and $\tau_{\Delta} = softplus$, then the selective SSM recurrence takes the form

$$
\begin{aligned}
g_t &= \sigma(Linear(x(t))) \\
h_t &= (1 - g_t)h_{t-1} + g_t x_t.
\end{aligned}
\tag{6}
$$

Scan. The selection mechanism is devised to address the constraints of Linear Time Invariance (LTI) models. However, it reintroduces the computation issue associated with SSMs. To enhance GPU utilization and efficiently materialize the state h within the memory hierarchy, hardware-aware state expansion is enabled by selective scan. By incorporating kernel fusion and recomputation with parallel scan, the fused selective scan layer effectively reduces the amount of memory I/O operations, leading to a significant acceleration compared to conventional implementations.

Discussion. Compared to RNNs and LSTMs, which struggle with vanishing gradients and long-range dependencies, Mamba offers efficient computation and memory utilization. While transformers excel in batch processing and handling long-range dependencies through attention mechanisms, they incur high computational costs, especially during inference. Mamba introduces a selective state space model, incorporating input-dependent matrices to enhance adaptability while maintaining the computational advantages of traditional SSMs. Mamba bridges the gap between traditional SSMs and modern neural network architectures by offering a selective dependency mechanism, optimal GPU memory utilization, and linear scalability with context length, thus providing a promising solution for various sequential data processing tasks.

3   Mamba for Vision

The original Mamba block is designed for one-dimensional sequences, yet vision-related tasks require processing multi-dimensional inputs like images, videos, and 3D representations. Consequently, to adapt Mamba for these tasks, enhancements to the scanning mechanism and architecture of the Mamba block are crucial to effectively handle multi-dimensional inputs.

In this section, we present efforts aimed at enabling Mamba to tackle vision-related tasks while enhancing its efficiency and performance. Initially, we delve into two foundational works: Vision Mamba [26] and VMamaba [27]. These works introduced the ViM block and VSS block, respectively, serving as the foundation for subsequent research endeavors. Subsequently, we explore additional works focused on refining the Mamba architecture as a backbone for vision-related tasks. Lastly, we discuss the work of integrating Mamba with other architectures such as convolution, recurrence, and attention.

3.1   Visual Mamba Block.

Drawing inspiration from the visual transformer architecture, it seems natural to preserve the framework of the transformer model while substituting the attention block with a Mamba block, while keeping the rest of the process intact. At the crux of the matter lies the adaptation of the Mamba block to vision-related tasks. Nearly simultaneously, Vision Mamba and VMamba present their respective solutions: the ViM block and the VSS block.

ViM. ViM block [26], sometimes also mentioned as Bidirectional Mamba block, annotates image sequences with position embeddings and condenses visual representations using bidirectional state space models. It processes input both forward and backward, employing one-dimensional convolution for each direction as shown in (a) Fig. 4. The Softplus function ensures non-negative Δ. Forward and backward $\mathbf{y}$ are computed via the state space model described in equations (2) and (3), and then combined through $SiLU$ gating to produce the output token sequence as (a) Fig. 3.

VSS. The Visual State Space (VSS) block [27] incorporates the pivotal state space model operation. It begins by directing the input through a depth-wise convolution layer, followed by a $SiLU$ activation function, and then through the state space model outlined in equations (2) and (3) employing an approximate $\overline{\mathbf{B}}$. Afterwards, the output of the state space model is subjected to layer normalization before being amalgamated with the output of other information streams as (b) Fig. 3. To address the encountered direction-sensitive issue, they introduce the Cross-Scan Module (CSM) to traverse the spatial domain and convert any non-causal visual image into order patch sequences, as shown in (b) Fig. 4. They refine the approximation of $\overline{\mathbf{B}}$ using the first-order Taylor series $\overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \approx (\Delta\mathbf{A})(\Delta\mathbf{A})^{-1}\Delta\mathbf{B} = \Delta\mathbf{B}$.
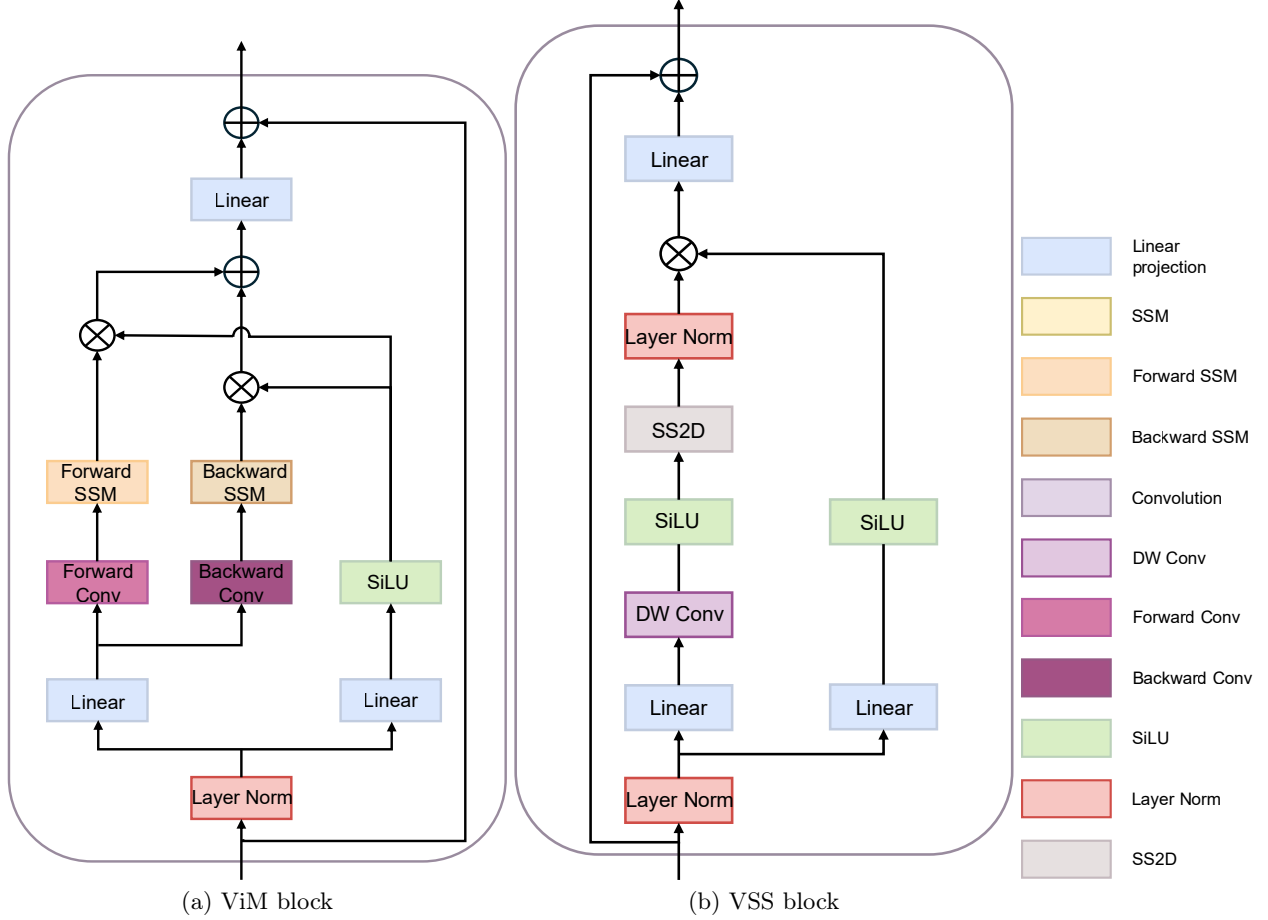
Fig. 3. ViM Block and VSS Block.

3.2 Pure Mamba

ViM-based. Inspired by the vision transformer architecture, Vision Mamba [26] replaces the transformer encoder with a vision mamba encoder based on ViM blocks while retaining the remainder of the process. This involves converting the two-dimensional image into flattened patches, followed by linear projection of these patches into vectors and the addition of position embeddings. A class token represents the entire patch sequence, and subsequent steps involve normalization layers and a MLP layer to derive the final predictions.

LocalMamba [28] is built based on Vim block and it introduces a novel scanning methodology that includes localized scanning within distinct windows to capture detailed local information in conjunction with global context. Plus, LocalMamba searches scanning directions across different network layers to identify and apply the most effective scanning combinations. They propose two variants, *i.e.* with plain and hierarchical structures. They proposed their LocalVim Block including four scanning directions (*cf.* (d) Fig. 4): vim scanning and partitions tokens into distinct windows along with their flipped counterparts facilitating scanning from tail to head, state space module, and spatial and channel attention module (SCAttn).

VSS-based. VMamba [27] has four stages after partitioning the input image into patches as Vision Mamba. VMamba stacks several VSS blocks on the feature map with resolution $\frac{H}{4} \times \frac{W}{4}$ as *Stage 1*. In *Stage 2*, before more VSS blocks involving, the feature map in *Stage 1* goes through a patch merge operation for down-sampling in order to build hierarchical representations, resulting in an output resolution of $\frac{H}{8} \times \frac{W}{8}$. While *Stage 3* and *Stage 4* are the repetition of *Stage 1* and *Stage 2* with resolutions of $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$.

Based on VSS block, PlainMamba block [29] enhances its capability to learn features from two-dimensional images through two main mechanisms: (i) employing a continuous 2D scanning process to improve spatial

continuity, ensuring tokens in the scanning sequence are adjacent, as illustrated in (c) Fig. 4, and (ii) incorporating direction-aware updating to enable the model to discern spatial relations among tokens by encoding directional information. PlainMamba improves the spatial discontinuity when moving to a new row/column in the 2D scanning mechanisms of Vim and VMamba by continuing the scanning with a reversed direction until it reaches the final vision token of the image. Furthermore, PlainMamba eliminates the need for special tokens.

Within lightweight model designs, EfficientVMamba [30] improves the capabilities of VMamba with an atrous-based selective scan approach, *i.e.* Efficient 2D Scanning (ES2D). Instead of scanning all patches from various directions and increasing the total number of patches, ES2D adopts a strategy of scanning forward vertically and horizontally while skipping patches and maintaining the number of patches unchanged, as shown in (e) Fig. 4. Their Efficient Visual State Space (EVSS) block comprises a convolutional branch for local features, uses ES2D as the SSM branch for global features, and all branches end through a squeeze-excitation block. They employ EVSS blocks for both Stage 1 and Stage 2, while opting for Inverted Residual blocks in Stage 3 and Stage 4 to enhance the capture of global representations.

Visual data as multi-dimensional data. As a part of multi-dimensional data, existing models for multi-dimensional data also work for visual-related tasks but often lack the capacity to facilitate inter- and intra-dimension communication or data-independent. The MambaMixer block [31] introduces a dual selection mechanism spanning across tokens and channels. It then links the sequential selective mixers through a weighted averaging mechanism, enabling layers to directly access input and output from various layers. Mamba-ND [32] expands the application of the SSM to higher dimensions by alternating sequence wandering across layers. Utilizing a similar scanning strategy as VMamba in the 2D scenario, it extends this approach to 3D. Additionally, they advocate for the use of multi-head SSM as an analogue to multi-head attention. In response to the inefficiencies and performance challenges encountered by traditional transformers in image and time series processing, a new architecture named Simplified Mamba-based Architecture, *SiMBA* [33] is proposed to incorporate the Mamba block for sequence modeling and Einstein FFT (EinFFT) for channel modeling, with the goal of enhancing the stability and efficiency of the model in handling image and time series tasks. The Mamba block proves effective at processing long sequence data, while EinFFT represents a novel channel modeling technique. Experimental results demonstrate that SiMBA surpasses existing State Space Models and transformers across multiple benchmark tests.

Summary of 2D scanning mechanisms. Scan is a key component for Mamba, when it comes to multi-dimensional inputs, the scanning mechanisms matter. We summarize the existing 2D scanning mechanisms in Fig. 4. In particular, the direction-aware updating employs a set of learnable parameters $\{\Theta_k\}$ to representing the four cardinal directions plus a special begin direction for the initial token, and reformulate (3) as

$$
\begin{aligned}
h'_k(t) &= \overline{\mathbf{A}}_t h_k(t) + (\overline{\mathbf{B}}_t + \overline{\Theta}_{k,t}) x(t), \\
y'(t) &= \sum_{k=1}^{4} \mathbf{C}_t h'_k(t), \\
y(t) &= y'(t) \odot z(t).
\end{aligned}
\tag{7}
$$

Expanding on the fundamental structure of Mamba and (7), we can devise the additional scanning mechanisms depicted in Fig. 4.

As an important element of Mamba, scanning mechanisms not only help in efficiency but also provide information in the scenario of visual-related tasks. We summarize the usage of different scanning mechanisms in existing works as Table. 1. Cross-Scan [27] and BiDirectional Scan [26] stand out as the most widely adopted scanning mechanisms. However, various other scanning mechanisms serve specific purposes. For example, 3D BiDirectional Scan [36] and Spatiotemporal Selective Scan [38] are tailored for video inputs. Local Scan [28] focuses on gathering local information, while ES2D [30] prioritizes efficiency.

(a) BiDirectional Scan [26]

(b) Cross-Scan [27]

(c) Continuous 2D Scanning [29]

(d) Local Scan [28]

(e) Efficient 2D Scanning (ES2D) [30]

(f) Zigzag Scan [34]

(g) Omnidirectional Selective Scan [35]

(h) 3D BiDirectional Scan [36]

(i) Hierarchical Scan [37]

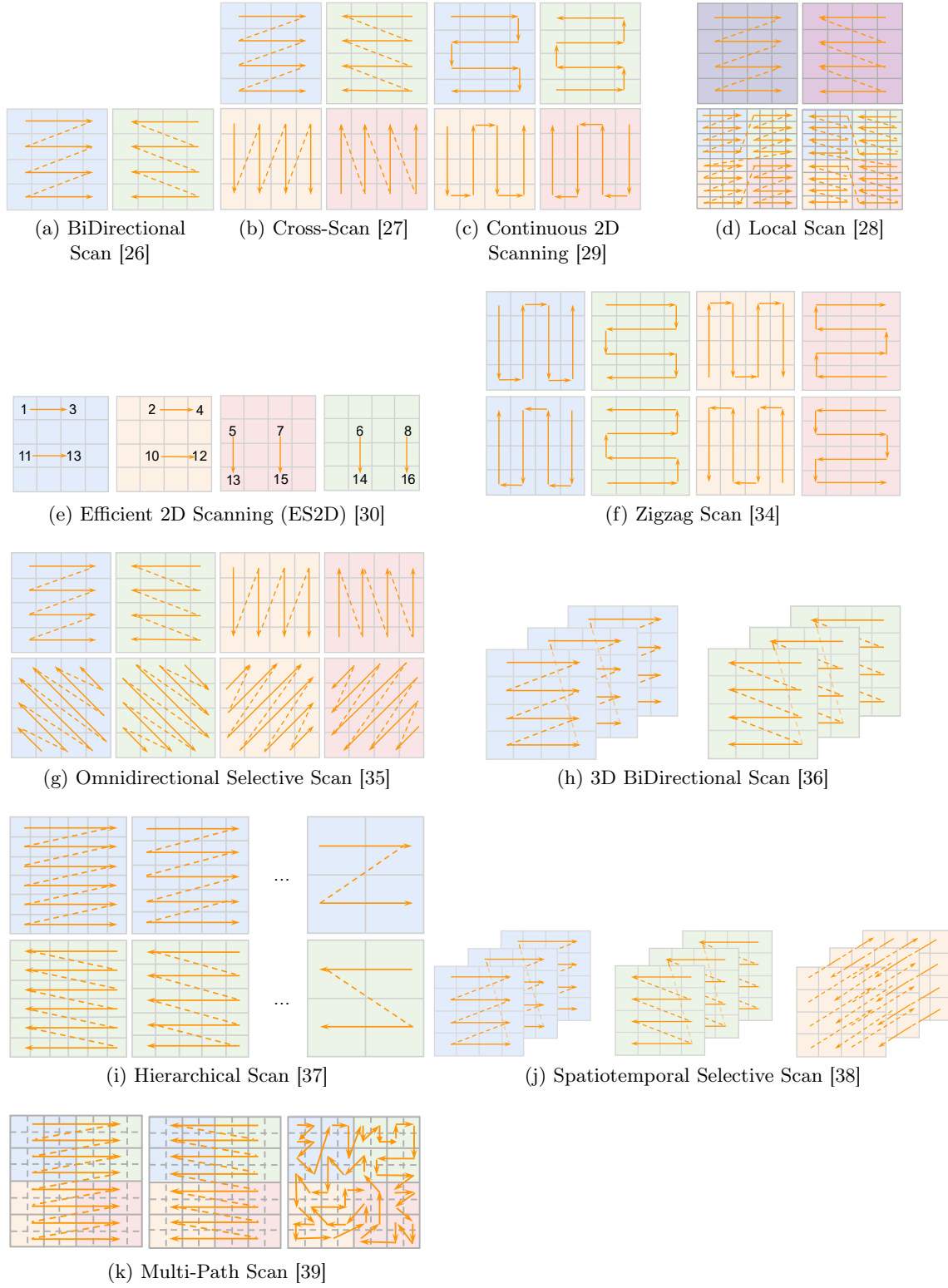(j) Spatiotemporal Selective Scan [38]

(k) Multi-Path Scan [39]

Fig. 4. Comparison between different 2D scanning and the selective scan orders in Vim, VMamba, PlainMamba, LocalMamba, Efficient VMamba, Zigzag, VmambaIR, VideoMamba, Motion Mamba, Vivim and RSMamba.

Table 1. Summary of the Scanning mechanisms used in visual Mamba.

Scanning Mechanisms	Method
BiDirectional Scan [26]	Vision Mamba [26],Motion Mamba [37] HARMamba [40],MMA [41],VL-Mamba[42] Video Mamba Suite [43],Point Mamba [44] LMa-UNet [45] Motion-Guided Dual-Camera Tracker [46]
Cross-Scan [27]	VMamba [27],VL-Mamba[42],VMRNN [47] RES-VMAMBA [48],Sigma [49],ReMamber [50] Mamba-UNet [51],Semi-Mamba-UNet [52] VMambaMorph [53],ChangeMamba [54] H-vmunet [55],MambaMIR [56],MambaIR [57] Serpent [58],Mamba-HUNet [59],TM-UNet [60] Swin-UMamba [61],UltraLight VM-UNet [62] VM-UNet [63],VM-UNET-V2 [64] MedMamba [65],MIM-ISTD [66],RS3Mamba [67]
Continuous 2D Scanning [29]	PlainMamba [29]
Local Scan [28]	LocalMamba [28],FreqMamba [68]
Efficient 2D Scanning (ES2D) [30]	EfficientVMamba [30]
Zigzag Scan [34]	ZigMa [34]
Omnidirectional Selective Scan [35]	VmambaIR [35],RS-Mamba [69]
3D BiDirectional Scan [36]	VideoMamba [36]
Hierarchical Scan [37]	Motion Mamba [37]
Spatiotemporal Selective Scan [38]	Vivim [38]
Multi-Path Scan [39]	RSMamba [39]

3.3 Mamba with Other Architectures

Mamba, being a novel component compared to convolution, recurrence, and attention, offers opportunities for synergistic combinations with other architectures that are still relatively underexplored. In this section, we examine existing exploratory findings on such combinations.

Mamba with Convolution. To combine Mamba with convolution, Mamba introduces the ability to obtain local information, which is essential for tasks related to medical images or segmentation tasks. RES-VMAMBA [48] pioneers incorporating a residual learning framework within the VMamba model to simultaneously leverage global and local state features inherent in the original VMamba architectural design. This architecture commences with a stem module responsible for processing the input image, followed by a series of VSS Blocks organized sequentially across four distinct stages. Diverging from the original VMamba framework, the Res-VMamba architecture adopts the VMamba structure as its backbone and directly integrates raw data into the feature map. They refer to this integration as the global-residual mechanism to distinguish it from the residual structure in the VSS block. This integration aims to facilitate the sharing of global image features alongside the information processed through the VSS blocks. This design seeks to harness the localized details captured by individual VSS blocks and the overarching global features inherent in the unprocessed input, thereby enhancing the model's representational capacity and improving its performance on tasks requiring a comprehensive understanding of visual data.

Mamba with Recurrence. To harness the long-sequence modeling capabilities of Mamba blocks and the spatiotemporal representation prowess of LSTM, the *VMRNN*[47] Cell eliminates all weights and biases in ConvLSTM[70] and employs VSS blocks to learn spatial dependencies vertically. In the VMRNN Cell, long-term and short-term temporal dependencies are captured by updating the information on cell states and

hidden states from a horizontal perspective. Building upon the VMRNN Cell, two variants are proposed: VMRNN-B and VMRNN-D. VMRNN-B primarily relies on stacking VMRNN layers, while VMRNN-D incorporates more VMRNN Cells and introduces Patch Merging and Patch Expanding layers. The Patch Merging layer serves for downsampling, effectively reducing the spatial dimensions of the data, which aids in decreasing computational complexity and capturing more abstract, global features. Conversely, the patch-expanding layer is utilized for upsampling, increasing the spatial dimensions to restore detail and enable precise localization of features in the reconstruction phase. Ultimately, the reconstruction layer takes the hidden state from the VMRNN layer and scales it back to the input size, generating the predicted frame for the next time step. Integrating downsampling and upsampling processes offers significant advantages in our predictive architecture. Downsampling simplifies the input representation, enabling the model to process higher-level features with reduced computational overhead. This is particularly advantageous for more abstractly understanding complex patterns and relationships within the data.

Mamba with Attention. The *SSM-ViT* block [71] is introduced for effective event-based information processing. It comprises three main components: a self-attention block (*Block-SA*), a dilated attention block (*Grid-SA*), and an SSM block. Block-SA focuses on immediate spatial relations and provides a detailed representation of nearby features. Grid-SA offers a global perspective, capturing comprehensive spatial relations and overall input structure. The SSM block ensures temporal consistency and smooth information transition between consecutive time steps. By integrating SSM with self-attention, the *SSM-ViT* block enables faster training and parameter timescale adjustment for temporal aggregation.

The Meet More Areas (*MMA*) block introduced in [41] adopts a MetaFormer-style architecture, comprising two Layer Normalization layers, a token mixer (consisting of a channel attention mechanism and a ViM block in parallel), and an MLP block for deep feature extraction. There are two main reasons for this choice: Firstly, models adopting MetaFormer-style architectures have shown promising results, indicating the potential for achieving favorable outcomes. Secondly, to fully leverage and utilize the global information extracted by the ViM block, the channel attention mechanism is incorporated to activate more pixels, as global details play a role in determining the channel attention weights. Additionally, it's reasonable to suggest that employing a convolution-based module can enhance the visual representation obtained by the ViM block and streamline the training process, similar to the benefits observed with transformers. For restoration, Residual State Space Blocks (RSSBs) [57] block add VSS block in front of the channel attention block, which enables VSS to focus on learning diverse channel representation after which the critical channels are selected by subsequent channel attention, thus avoiding channel redundancy.

4   Visual Mamba in Application Fields

Mamba-based modules elevate the efficiency of processing sequential data, adeptly capturing long-range dependencies and seamlessly integrating into existing systems. In medical visual tasks and remote sensing images, where inputs often entail high-resolution data, Mamba emerges as a pivotal tool in augmenting various visual tasks, particularly those pertinent to medical applications.

In this section, we begin by highlighting the contributions of Mamba-based modules in enhancing general visual-related tasks. Subsequently, we delve into their specific impact on medical visual tasks and remote-sensing images.

4.1   General Visual Mamba

We categorize general vision-related tasks into High/Mid-level vision and Low-level vision. High/Mid-level vision encompasses recognition tasks such as classification, object detection, segmentation, and prediction across various input types, including images, videos, and 3D representation. In contrast, Low-level vision includes restoration, generation *etc.*

High/Mid-level vision The Visual Mamba Backbone [26,27,29,28,32] performance decent on classification, object detection and segmentation. *SSM-ViT* [71] is designed for object detection using event cameras. Unlike standard frame-based cameras, event cameras record per-pixel relative brightness changes in a scene

Table 2. Representative works of general visual mamba

Category	Sub-category	Method	Efficiency	Code
Backbone	Visual Mamba	Vision Mamba [26]	Params Vim-Ti: 7, Vim-S: 26	✓
		VMamba [27]	FLOPs Tiny: 4.5, Small: 9.1, Base: 15.2	✓
		PlainMamba [29]	FLOPs PlainMamba-L1: 3.0 PlainMamba-L2: 8.1 PlainMamba-L3: 14.4	✓
		LocalMamba [28]	FLOPs LocalVMamba-T: 5.7 LocalVMamba-S: 11.4	✓
		Mamba-ND [32]	Params Mamba-2D: 24, Mamba-3D: 36	✓
		SiMBA [33]	-	✓
		RES-VMAMBA [48]	-	✓
	Efficient Mamba	EfficientVMamba [30]	-	✓
		MambaMixer [31]	-	✓
High/Mid-level vision	Object detection	SSM-ViT [71]	Params 17.5	✗
	Segmentation	ReMamber [50]	-	✗
		Sigma [49]	-	✓
	Video classification	ViS4mer [72]	Memory 5273.6	✓
	Video understanding	Video Mamba Suite [43]	-	✓
		VideoMamba [36]	FLOPs VideoMamba-Ti: 7.1 VideoMamba-S: 28, VideoMamba-M: 83.1	✓
		SpikeMba [73]	-	✗
	Multi-Modal understanding	Cobra [74]	-	✓
		ReMamber [50]	-	✗
		VL-Mamba [42]	-	✗
	Video prediction	VMRNN [47]	Params 2.6, FLOPs 0.9	✓
		HARMamba [40]	FLOPs 279.21 (PAMAP2), 237.83 (UCI) 238.36 (UNIMIB HAR), 256.52 (WISDM)	✗
Low-level vision	Image super-resolution	MMA [41]	-	✗
	Image restoration	MambaIR [57]	Params 16.7	✓
		SERPENT [58]	-	✗
		VmambaIR [35]	Params 10.50, FLOPs 20.5	✓
	Image dehazing	UVM-Net [75]	Params 19.25	✓
	Image derain	FreqMamba [68]	Params 14.52	✗
	Image deblurring	ALGNet [76]	FLOPs 17	✗
	Visual generation	MambaTalk [77]	-	✗
		Motion Mamba [37]	-	✓
		DiS [78]	-	✓
		ZigMa [34]	-	✓
	Point cloud	3DMambaComplete [79]	Params 34.06, FLOPs 7.12	✗
		3DMambaIPF [80]	-	✗
		Point Cloud Mamba [81]	Params 34.2, FLOPs 45.0	✗
		POINT MAMBA [44]	Memory 8550	✓
		SSPointMamba [82]	Params 12.3, FLOPs 3.6	✓
	3D reconstruction	GAMBA [83]	-	✗
	Video generation	SSM-based diffusion model [84]	-	✓

[1] For the Efficiency, Inference speed is in ms, Memory is in MB, Params is in M, and FLOPS is in G.

as they occur. Therefore, object detection with event cameras requires processing an asynchronous stream of events in a four-dimensional spatio-temporal space. Earlier studies have employed RNN architectures with convolutional or attention mechanisms to develop models that exhibit superior performance on downstream tasks using event camera data. However, these models often suffer from slow training. In response, the SSM-ViT block is introduced, leveraging SSM for efficient event-based information processing. It explores two strategies to mitigate aliasing effects when deploying the model at higher frequencies.

Leveraging Mamba's significant advancements in efficient training and inference with linear complexity, *ReMamber* [50] is introduced for referring image segmentation (RIS), a challenging task in the realm of multi-modal understanding. Distinguished from conventional segmentation, RIS entails identifying and segmenting specific objects in images based on textual descriptions. The ReMamber architecture comprises several Mamba Twister blocks, each featuring multiple VSS blocks and a Twisting layer. The Mamba Twister block serves as a multi-modal feature fusion block, receiving visual and textual features as input and outputting the fused multi-modal feature representation. Intermediate features are extracted after each Mamba Twister block and subsequently fed into a flexible decoder to generate the final segmentation mask. The VSS layers are tasked with extracting visual features, while the Twisting layer primarily captures effective visual-language interactions. The authors conduct experiments on multiple RIS datasets, achieving state-of-the-art results.Sigma [49] is a novel network for multimodal semantic segmentation tasks. One of them, Siam Mamba encoder, uses cascaded visual state space (VSS) blocks to extract multi-scale global information from different modalities. And an attention-based Mamba fusion mechanism and a channel-aware Mamba decoder are proposed. In the decoding stage, the fused features are further enhanced by channel-aware visual state space (CVSS) blocks, which can effectively capture multi-scale long-range information and realize cross-modal information integration.

Unlike transformers that rely on quadratic complexity attention mechanisms, Mamba, as a pure SSM-based model, excels in handling long sequences with linear complexity and proves particularly adept at processing lengthy videos at high resolutions. *ViS4mer* [72] serves as a model primarily used for recognizing and classifying long videos, particularly for understanding and categorizing lengthy movie clips. ViS4mer consists of two main components: a standard Transformer encoder designed for extracting short-distance spatiotemporal features from videos, and a multi-scale temporal S4 decoder optimized for subsequent long-range temporal reasoning. Leveraging the capability of the fundamental SSM to capture long-range dependencies in sequential data, the multi-scale temporal S4 decoder is implemented based on SSMs to reduce the computation cost of the model.

The *Video Mamba Suite* [43] is not a novel method; rather, it explores and assesses the potential of SSM, represented by Mamba, in video understanding tasks. The ViM block is enhanced into the Decomposed Bidirectionally Mamba (DBM) Block, which separates the input projector while sharing the parameters of SSM in both scanning directions. They classify Mamba into four distinct roles for modeling videos and compare it with existing Transformer-based models to evaluate its effectiveness in various video understanding tasks. Furthermore, the Video Mamba Suite comprises 14 models/modules to evaluate performance across 12 video understanding tasks. The experiments demonstrate that Mamba is applicable in video analysis and can be utilized for more complex, multimodal video understanding challenges. Apart from the Video Mamba Suite, *VideoMamba* [36] is proposed for video understanding tasks, with a specific focus on addressing two major challenges: local redundancy and global dependencies. This study evaluates VideoMamba's capabilities across four key aspects: scalability in the video domain, sensitivity to short-term action recognition, advantages in long-term video understanding, and compatibility with other modalities. To enhance model scalability in the visual domain, VideoMamba employs a self-distillation strategy. This approach significantly enhances VideoMamba's performance as both the model and input sizes increase, without the need for pre-training on large-scale datasets. While the ViM block enhances the model's spatial perception capabilities, VideoMamba extends this capability to 3D video understanding by including spatio-temporal bidirectional scanning. Through the extension of the ViM block, VideoMamba achieves a notable increase in processing speed and a decrease in computational resource consumption without compromising performance.SpikeMba [73] is a novel multimodal video content understanding framework designed to handle the task of temporal video localization in video content.SpikeMba combines Spiking Neural Networks (SNNs) and State Space Models (SSMs) to capture fine-grained relationships between multimodal input features. In particular, the Spike Saliency Detector (SSD) utilizes the thresholding mechanism of SNNs to generate saliency proposal sets that signal highly relevant or salient instances in the video through spikes. The

Multimodal Relevance Mamba Block (MRM) is based on SSM and enhances the modeling of long range dependencies while maintaining linear complexity with respect to the input size.

In recent years, Multimodal Large Language Models (MLLMs) have been used with remarkable success in various domains. However, as a base model for many downstream tasks, current MLLMs consist of well-known Transformer networks with low secondary computational complexity. To improve the efficiency of these models, Cobra[74] has been proposed. This is a new multimodal large language model (MLLM) with linear computational complexity, which integrates the efficient Mamba language model into the visual modality, exploring different modal fusion schemes, and thus identifying ways to produce the most efficient multimodal representation.Cobra consists of three components: a visual encoder, a projector, and a Mamba backbone. Among them, the visual coder is used to extract the visual representation of the image, and the projector is used to transform the dimensions of the visual representation to match the dimensions of the tokens in the Mamba language model.The Mamba backbone, on the other hand, consists of a stack of 64 identical basic blocks that receive a combination of visual and textual embeddings while preserving the connectivity and the RMSNorm, and autoregressively transforms them into the target token sequences. VL-Mamba [42] is also a multimodal large language model consisting of a pre-trained visual coder, a randomly initialized MMC, and a pre-trained Mamba LLM. Among them, the visual coder uses the Vision Transformer (ViT) architecture to generate a sequence of patch features of the original image.For the MMC, a 2D visual selective scanning mechanism is proposed to solve the computer vision task since the state-space model is designed to deal with 1D sequential data with causality and the visual sequences generated by the visual coder are 2D non-causal data . The article explores three multimodal connector variants: the MLP (multilayer perceptron), the VSS-MLP (MLP combined with a VSS module) and the VSS-L2 (two linear layers combined with a VSS module). Input images are first acquired as visual features through a visual coder, then visual sequences are fed into the MMC, and finally the output vectors are fed into the Mamba LLM in combination with a tokenized text query in order to generate the corresponding response.The integration and processing of visual and verbal information is optimized through the synergy of these components.The Referential Image Segmentation (RIS) task requires a model to recognize and segment specific objects in an image based on textual descriptions, and ReMamber[50] is a novel architecture for handling this task.ReMamber combines the Mamba model and introduces multimodal Mamba Twister blocks to explicitly simulate image-text interactions through its unique channeling and spatial warping mechanism Fusing textual and visual features.ReMamber extracts intermediate features after each Mamba Twister block and feeds them into a flexible decoder to generate the final segmentation mask.

To address the unparalleled challenge of predicting temporal and spatial dynamics for spatio-temporal forecasting in videos, the *VMRNN* cell [47] introduces a novel recurrent unit designed to efficiently handle spatio-temporal prediction tasks. Recognizing the challenges in processing extensive global information, the VMRNN cell integrates VSS blocks with LSTM architecture to leverage the long-sequence modeling abilities of VSS blocks and the spatio-temporal representation capabilities of LSTM. This integration enhances the accuracy and efficiency of spatio-temporal predictions. The model conducts image-level analysis by segmenting each frame into patches, which are subsequently flattened and processed through an embedding layer. This process enables the VMRNN layer to extract and predict spatio-temporal features effectively. HAR-Mamba [40] builds on ViT blocks for activity recognition and reaches superior performance while reducing computational and memory overhead on the activity recognition tasks.

Low-level vision In the realm of image super-resolution, Meet More Areas (*MMA*) [41] stands out as a novel model designed for super-resolution tasks. Built on the ViM block, MMA aims to enhance performance by activating a wider range of areas within images. To achieve this, MMA adopts several key strategies, including integrating ViM into MetaFormer-style modules, pre-training ViM on larger datasets, and employing complementary attention mechanisms. MMA comprises three main modules: shallow feature extraction, deep feature extraction, and high-quality reconstruction. Leveraging the ViM module, MMA effectively models global information and further expands the activation region through attention mechanisms.

Existing restoration backbones often face the dilemma between global receptive fields and efficient computation, hindering their application in practice, while Mamba shows great potential for long-range dependency modeling with linear complexity, which offers a way to resolve the above dilemma. *MambaIR* [57] aims to solve the problem by introducing local enhancement and channel attention mechanisms to improve the standard Mamba model. The methodology of the model mainly consists of three stages: shallow feature

extraction, deep feature extraction, and high-quality image reconstruction. Among them, the deep feature extraction stage utilizes multiple residual state space blocks (RSSBs) for feature extraction, which add a VSS block before the channel attention block designed in previous transformer-based restoration networks. *SERPENT* [58] designs a hierarchical architecture that processes input images in a multi-scale manner, including processing steps such as segmentation, embedding, downsampling, and upsampling, and introduces jump connections to facilitate information flow. Among them, the Serpent block is the main processing unit, consisting of multiple VSS blocks stacked on each other. Serpent combines the advantages of Convolutional Networks and Transformers to drastically reduce the computational effort, GPU memory requirement, and model size while maintaining high reconstruction quality. *VmambaIR* [35] proposes the OSS module to comprehensively and efficiently model image features from six directions. The omnidirectional selective scanning mechanism overcomes the unidirectional modeling limitation of SSMs and achieves comprehensive pattern recognition and modeling by modeling the image information flow in all three dimensions.

UVM-Net [75] is a novel single-image defogging network architecture that combines the local feature extraction of convolutional layers and SSM's long-range dependency modeling capability to exhibit efficient performance. The method employs an encoder-decoder network architecture, and the critical component is the ViM block, which leverages the long-range modeling capability of SSM by rolling the feature map over the channel domain. Unlike U-Mamba [85] and Mamba-UNet [51], the ViM block establishes long-range dependencies on another dimension of the feature map (the non-channel domain).

Images lose important frequency information under the influence of raindrops, which affects the performance of visual perception and advanced visual tasks.FreqMamba [68] is a novel image de-raining method that combines Mamba modeling and frequency analysis techniques to solve the image de-raining problem. Specifically, FreqMamba contains three branching structures, including spatial Mamba, frequency band Mamba, and Fourier global modeling. Spatial Mamba processes raw image features to extract details and correlations within the image. Frequency Band Mamba uses the Wavelet Packet Transform (WPT) to decompose the input features into spectral features in different frequency bands and scan them along the frequency dimension. Fourier modeling i.e., processing the input using the Fourier transform captures the global degradation patterns affecting the image. Extensive experiments have shown that FreqMamba outperforms existing state-of-the-art methods both visually and quantitatively.

Image deblurring is a classical problem in low-level computer vision, which aims to recover high-quality clear images from blurred input images.ALGNet [76] is an efficient image deblurring network that utilizes selective state-space models (SSM) to aggregate rich and accurate features.The network consists of multiple ALGBlocks, each of which contains a CLGF module that captures local and global features and a feature aggregation module FA.The CLGF module captures long-range dependent features using SSM and employs a channel-attention mechanism to reduce local pixel forgetting and channel redundancy.The FA module emphasizes the importance of local features in recovery by recalibrating the weights.

The efficiency of Mamba contributes significantly to mitigating the high computational complexity associated with training generation tasks. To address the change in generating long and diverse sequences with low latency, *MambaTalk* [77] implements a two-stage modeling strategy with discrete motion priors to enhance the quality of gestures and employs mamba block to enhance gesture diversity and rhythm through multimodal integration. *Motion Mamba* [37] is introduced to construct a motion generation model based on Mamba, leveraging its efficient hardware-aware design. Motion Mamba consists of two main components: the Hierarchical Temporal Mamba (HTM) block for temporal data handling, and the Bidirectional Spatial Mamba (BSM) block for processing latent poses. The HTM block employs several isolated SSM modules within a symmetric U-Net architecture to maintain motion consistency across frames. Meanwhile, the BSM block enhances the accuracy of motion generation within a temporal frame by processing latent poses bidirectionally. Diffusion State Space Models (*DiS*) [78] substitute the conventional U-Net backbone in diffusion models with SSMs. This framework considers all inputs, including time, conditions, and noisy image patches and tokens. To tackle the oversight of spatial continuity in the scanning scheme of existing Mamba-based vision methods, Zigzag Mamba [34] is introduced as a straightforward, plug-and-play solution inspired by DiT style approaches. Essentially, it retains the scanning scheme of Plain Mamba but expands it from four to eight schemes by incorporating their mirror flipping schemes, as shown in (f) Fig. 4. Subsequently, Zigzag Mamba is integrated with the Stochastic Interpolant framework, forming *ZigMa*, to explore the scalability of the diffusion model on large-resolution visual datasets.*GAMBA* [83] introduces a sequential network based on Mamba, which enables context-dependent reasoning and linear scalability for sequence length. This

architecture accommodates many Gaussians for the 3D Gaussian splatting process. To address the issue of quadratic memory consumption increase with sequence length in traditional attention-based video generative diffusion models, *SSM-based diffusion model* [84] is introduced for generating longer video sequences. Like ViS4mer [72], the SSM-based diffusion model reimagines the attention modules within the conventional temporal layers of Video Diffusion Models (VDMs). It replaces them with a ViM block designed to capture the temporal dynamics of video data, accompanied by a Multi-Layer Perceptron (MLP) to boost model performance. This innovative approach significantly mitigates memory consumption for extended sequences.

The irregularity and sparsity of point cloud data have been a challenge in 3D vision. Although Transformer shows potential in point cloud analysis tasks based on its powerful global information modeling capability, its computational complexity grows significantly with the increase of input length, limiting its application to long sequence models. In this context, SSPoint Mamba [82] is proposed as a simple and effective state space model for point cloud analysis. The model uses embedded point blocks as inputs and enhances the global modeling capability of SSM with a reordering strategy that provides a more rational geometric scanning order. The reordered point tokens (point tokens) are fed into a series of Mamba blocks to causally capture the point cloud structure. The model demonstrates its effectiveness on several point cloud analysis tasks.3DMambaComplete [79] network aims to address the computational complexity challenges posed by the loss of local details and attention mechanisms common in point cloud completion, and it is built on top of the novel Mamba framework. The method first downsamples incomplete point clouds, then enhances feature learning with Mamba Encoder, then predicts and refines hyperpoints, and disperses hyperpoints to different 3D locations by learning specific offsets, and finally performs point deformation to generate a complete point cloud. The model utilizes the concept of structured state-space modeling to improve shape reconstruction by predicting hyperpoints and point deformation to control the deformation at each hyperpoint location.The point cloud filtering model 3DMambaIPF [80], on the other hand, aims to deal with the denoising of large-scale point cloud data. The network integrates Mamba into a filtering module, Mamba-Denoise, to achieve accurate and fast modeling of long sequences of point cloud features.3DMambaIPF consists of several Mamba-Denoise modules, and employs a strategy of iterative point cloud filtering to implement the filtering process for point clouds. The loss function includes reconstruction loss and differentiable rendering loss to minimize the distance between the noisy point cloud and the real point cloud, optimize the visual boundary of the point cloud, and improve the realism of the denoising results.Point Cloud Mamba [81] combines local and global modeling frameworks and proposes a Consistent Traversal Serialization (CTS) approach to transform 3D point cloud data into 1D point sequences while ensuring that adjacent points in the sequence are also spatially adjacent. In addition, point cueing and position encoding based on spatial coordinate mapping are introduced to help Mamba process point sequences more efficiently and inject position information.Point Mamba [44] is a new backbone network for point cloud processing that addresses the causality of state space models on point cloud data by introducing an octree-based ordering strategy. In addition, Point Mamba blocks incorporate a bi-directional selective scanning mechanism to adjust the sequence order dependency of Mamba.

4.2   Medical Visual Mamba

Transformers [8] have profoundly influenced the field of medical imaging with their ability to master complex data representations. They have led to notable advancements across various imaging modalities, including Radiography [86], Endoscopy [87], Computed Tomography (CT) [88], Ultrasound Images [89], and Magnetic Resonance Imaging (MRI) [90]. However, because most medical images are high-resolution and detailed, transformer models typically require considerable computational resources, which scale quadratically with image resolution.

The medical imaging field has recently experienced a surge in the development of Mamba-based methodologies, particularly following the introduction of VMamba. This section provides detailed examples of these design choices, further dividing them into 2D and 3D-based approaches based on the input type, as displayed in Table 3.
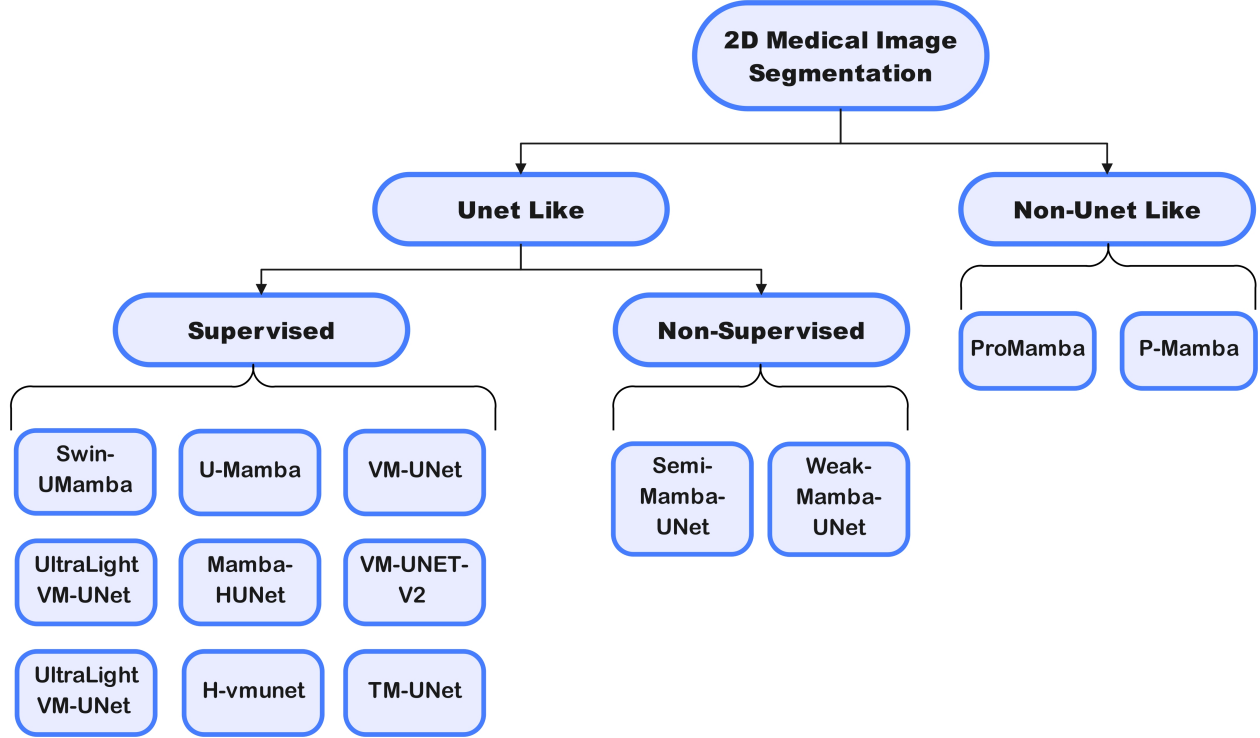
Fig. 5. An overview of Mamba models used for segmentation task in 2D medical images.

2D Medical Image Mamba has shown impressive potential in 2D medical segmentation, as displayed in Fig. 5. Here, we discuss in detail some methods that explore using mamba to model the global structure information of 2D medical segmentation.

Most of the innovative architectures that have been developed are based on U-Net, which has achieved remarkable success in various medical image segmentation tasks. U-Mamba [85] is the first extension of the mamba model to the U-Net architecture for visual segmentation in biomedical imaging, addressing the challenge of long-range dependency modeling, which is captured by a hybrid CNN-SSM block. Wu et al. introduced the High-order Vision Mamba UNet (H-vmunet) [55], an improvement of U-Mamba, which utilizes a High-order 2D-selective-scan at each interaction order to bolster the learning of local features while minimizing the incorporation of redundant information. Shortly after their initial release, the team expanded their work by introducing the UltraLight VM-UNet [62]. This new iteration was developed by in-depth analysis of the critical factors affecting parameter efficiency within the Mamba framework. This resulted in a remarkably lightweight model with a mere 0.049M parameters and a computational efficiency of only 0.060 GFLOPs.In addition, Mamba-UNet [51] combines the encoder-decoder architecture of U-Net with the capabilities of Mamba and maintains spatial information at different network scales through jump connections. A Visual Mamba-based VSS block is used, which utilizes linear embedding layers and deep convolution to extract features while downsampling and upsampling are facilitated by multiple merge operations and extension layers for comprehensive feature learning.

Pyramid ViT (PVT) and Swin-Unet are both pioneering hierarchical designs that apply visual tasks and propose progressive shrinking pyramids and spatial-reduction attention. Drawing inspiration from PVT and Swin-Unet, Ruan et al. introduced VM-UNet [63], a foundational model for purely SSM-based segmentation in medical imaging. This model demonstrates the capabilities of SSM in medical image segmentation and consists of three primary components: an encoder, a decoder, and skip connections. Building on their previous work, the team proposed VM-UNET-V2 [64]. The Visual State Space (VSS) block was introduced to capture a broader range of contextual information. The Semantics and Detail Infusion (SDI) mechanism was

Table 3. Representative works of medical visual mamba

Category	Sub-category	Method	Efficiency	Code
2D	Segmentation	Mamba-UNet [51]	-	✓
		H-vmunet [55]	Memory 0.676 Params 8.97	✓
		Mamba-HUNet [59]	-	✗
		P-Mamba [91]	Inference speed 23.49 Memory 12.22 Params 183.37 FLOPs 71.81×10^9	✗
		ProMamba [92]	Params 102	✗
		TM-UNet [60]	Params 14.86 Total Params 8.41 FLOPs 3.42	✗
		Semi-Mamba-UNet [52]	-	✓
		Swin-UMamba [61]	Params 28 FLOPs 18.9	✓
		UltraLight VM-UNet [62]	Params 0.049 GFLOPs 0.060	✓
		U-Mamba [85]	-	✓
		VM-UNet [63]	Params 34.62 FLOPs 7.56 FPS 20.612	✓
		VM-UNET-V2 [64]	Params 17.91 FLOPS 4.40 FPS 32.58	✓
		Weak-Mamba-UNet [93]	-	✓
	Radiation dose prediction	MD-Dose [94]	Inference speed 18 Params 30.47	✓
	Classification	MedMamba [65]	-	✓
		MambaMIL [95]	-	✓
	Image reconstruction	MambaMIR/MambaMIR-GAN [56]	-	✓
	Exposure correction	FDVM-Net [96]	Inference speed 22.95	✓
3D	Segmentation	LMa-UNet [45]	-	✓
		LightM-UNet [97]	Params 1.87 FLOPs 457.62×10^9	✓
		SegMamba [98]	Inference speed 151	✓
		T-Mamba [99]	-	✓
		Vivim [38]	FPS 35.33	✓
	Classification	CMViM [100]	Params 50	✓
	Motion tracking	Motion-Guided Dual-Camera Tracker [46]	-	✗
	Backbone	nnMamba [101]	Params 15.55 FLOPs 141.14	✗
	Image registration	VMambaMorph [53]	Inference speed 19 Memory 3.93 Params 9.64	✓
		MambaMorph [102]	Inference speed 27 Memory 7.60 Params 7.59	✓

[1] For the Efficiency, Inference speed is in ms, Memory is in Gb, Params is in M, and FLOPS is in G.

also implemented to enhance the fusion of low-level and high-level features. Mamba-HUNet [59], another multi-scale hierarchical upsampling network, incorporates the Mamba technique. It leverages Visual State Space blocks and patch merging layers to extract hierarchical features, ensuring the preservation of spatial information. TM-UNet [60] introduces improvements to the bottleneck layer. This architecture proposes Triplet SSM as the bottleneck layer, representing the first attempt to use a pure SSM approach to integrate spatial and channel features. Existing Mamba-based models are primarily trained from scratch, overlooking potential benefits. Swin-UMamba [61], a new Mamba-based model tailored explicitly for medical image segmentation tasks, leveraged the strengths of ImageNet-based pretraining.

Previous discussions have primarily focused on supervised learning methods, but other supervisory approaches have also been explored. Semi-Mamba-UNet [52] combines a visual Mamba-based U-Shape Encoder-Decoder with a traditional CNN-based UNet in a semi-supervised learning framework. It introduces a self-supervised pixel-level contrastive learning strategy using a pair of projectors to improve feature learning, particularly on unlabeled data. Weak-Mamba-UNet [93] is a novel weakly-supervised learning framework for medical image segmentation, combining CNNs, ViT, and the VMamba. It focuses on scribble-based annotations and uses a collaborative, cross-supervisory mechanism with pseudo labels for iterative network learning and refinement.

Some segmentation approaches diverge from UNet architectures. P-Mamba [91] introduces a novel dual-branch framework for highly efficient left ventricle segmentation in pediatric echocardiograms. This model features an innovative DWT-based encoder branch equipped with Perona-Malik Diffusion (PMD) Blocks. Moreover, P-Mamba adopts vision mamba layers within its vision mamba encoder branch to bolster computational and memory efficiency. PromptMamba [92] represents a groundbreaking integration of the Vision-Mamba and prompt technologies, marking a significant milestone as the first model to leverage the Mamba framework for the specific task of polyp segmentation.

In addition, mamba has expanded its research in 2D medical imaging beyond segmentation, enhancing the precision and speed of image analysis to support diagnosis and treatment planning. Classification is a vital and fundamental task in the field of medical image analysis. Yue et al. introduced Vision Mamba for this purpose, also known as MedMamba [65]. They developed a novel module named Conv-SSM, which combines the local feature extraction of convolutional layers with the long-range dependency capture of SSM, enabling efficient modeling of medical images from different modalities. Furthermore, MambaMIL [95] introduces the Sequence Reordering Mamba (SR-Mamba), a model that recognizes the order and distribution of instances in long sequences to harness the embedded valuable information effectively. Image reconstruction is pivotal in enhancing diagnostic processes, as high-quality and high-fidelity medical images are fundamental to the accuracy and efficiency of clinical decisions. Huang et al. [56] have developed MambaMIR, a model leveraging Mamba technology for the reconstruction of medical images, alongside its advanced counterpart, MambaMIR-GAN, which incorporates Generative Adversarial Networks. Moreover, Zheng et al. introduce FDVision Mamba (FDVM-Net) [96], a frequency-domain-based network that effectively corrects image exposure by reconstructing the frequency domain of endoscopic images, as recorded endoscopic images often suffer from exposure abnormalities. In specialized areas, MD-Dose [94], a cutting-edge diffusion model based on the Mamba architecture, was designed to predict radiation therapy dose distribution for thoracic cancer patients accurately.

3D Medical Image 3D image analysis in medical imaging enables more accurate and comprehensive diagnoses by providing detailed views of complex anatomical structures. Gong et al. present nnMamba [101], an innovative architecture designed for 3D medical imaging applications, which integrates local and global relationship modeling via the MICCSS (Mamba-In-Convolution with Channel-Spatial Siamese input) module. nnMamba was tested on a comprehensive benchmark of six datasets for three crucial tasks: segmentation, classification, and landmark detection, showcasing its capability in long-range relationship modeling at channel and spatial levels.

Precise 3D segmentation outcomes can alleviate physicians' diagnostic workloads in disease management. SegMamba [98], a cutting-edge architecture, is the first method to utilize Mamba specifically for accurate 3D segmentation in medical imaging. It introduced a tri-orientated Mamba (ToM) module for modeling 3D features from three directions and a gated spatial convolution (GSC) module to enhance spatial feature representation before each ToM module. Similarly employing a U-shaped architecture, LightM-UNet [97] harnesses the Residual Vision Mamba Layer exclusively in a solely Mamba approach for extracting deep

semantic features and modeling extensive spatial dependencies within a lightweight framework. Moreover, both LMa-UNet [45] and T-Mamba [99] built upon the foundation of SegMamba, with improvements made to the Mamba block. A notable aspect of LMa-UNet [45] is its use of large windows, which outperforms small kernel-based CNNs and small window-based Transformers in local spatial modeling, while T-Mamba [99] develops a gate selection unit to adaptively combine two features in the spatial domain with one feature in the frequency domain, marking the first instance of incorporating frequency-based features into the vision mamba framework. The issue of long-term temporal dependency in video scenarios has also been addressed by developing a generic Video Vision Mamba-based framework named Vivim [38]. Using the specifically engineered Temporal Mamba Block, this model effectively compresses long-term spatiotemporal data into sequences of different scales.

For image registration task, MambaMorph [102] introduces a groundbreaking multi-modality deformable registration framework that enhances medical image analysis by combining a Mamba-based registration module with an advanced feature extractor for efficient spatial correspondence and feature learning. The VMambaMorph [79] has further enhanced its VMamba-based block by incorporating a 2D cross-scan module, redesigned to process 3D volumetric features efficiently.

In other domains, the Contrastive Masked Vim Autoencoder (CMViM) [100] tackles Alzheimer's disease (AD) classification by incorporating Vision Mamba (ViM) into a masked autoencoder for 3D multi-modal data reconstruction. For endoscopy skill evaluation, a low-cost motion-guided dual-camera tracker [37] provides reliable endoscope tip feedback, and a Mamba-based motion-guided prediction head (MMH) merges visual tracking with historical motion data using a state space model.

Challenge Here we explore some promising future research directions for vision Mamba in medical image analysis. Challenges include the need for pre-training on large datasets, enhancing the interpretability of Mamba-based medical imaging approaches, and improving robustness against adversarial attacks. Additionally, there is a need to design efficient Mamba architectures suitable for real-time medical applications and to address the challenges in deploying Mamba-based models in distributed settings.

Table 4. Representative mamba work in remote sensing image

Category	Method	Highlight	Efficiency	Code
Pan-sharpening	Pan-Mamba [103]	channel swapping Mamba; cross-modal Mamba	Params 0.1827 FLOPs 3.0088	✓
Infrared Small Target Detection	MIM-ISTD [66]	Mamba-in-Mamba architecture	Params 1.16 FLOPs 1.01 Inference speed 30 Memory 1774	✓
Classification	RSMamba [39]	multi-path activation	-	✓
	HSIMamba [104]	process data bidirectionally	Memory 136.53	✓
Image dense prediction	RS-Mamba [69]	omnidirectional selective scan	-	✓
Change detection	ChangeMamba [54]	cross-scan mechanism	-	✓
Semantic segmentation	RS3Mamba [67]	dual-branch network	FLOPs 31.65 Params 43.32 Memory 2332	✓
	Samba [105]	encoder-decoder architecture	Params 51.9	✓

[1] For the Efficiency, Inference speed is in ms, Memory is in MB, Params is in M, and FLOPS is in G.

4.3   Remote Sensing Image

The progress of remote sensing technology has sparked interest in high-resolution earth observation, with the Transformer model offering an optimal solution through its attention mechanism. However, its quadratic complexity poses challenges in modeling efficiency and memory usage. The State Space Model (SSM) addresses these issues by establishing long-distance dependencies with near-linear complexity, and Mamba further enhances efficiency through hardware optimization and time-varying parameters. The representative recent work is shown in Table 4.

Drawing inspiration from TNT, Chen et al. introduced a new Mamba-in-Mamba (MiM-ISTD) [66] architecture to enhance the efficiency of infrared small target detection. In this approach, local patches are considered "visual sentences," while Outer Mamba is utilized to extract global information. Moreover, for remote sensing image classification, RSMamba [39] features a dynamic multi-path activation mechanism to improve Mamba's capability in handling non-causal data. RS-Mamba [69] is adept at handling very-high-resolution (VHR) remote sensing images for dense prediction tasks, utilizing an omnidirectional selective scan module to model images from various angles comprehensively. Classification of hyperspectral images is difficult in remote sensing research due to their complex, high-dimensional data.HSIMamba [104] is designed with a module dedicated to spatial analysis, which includes multiple spectral bands and three-dimensional spatial structures to take advantage of the rich multidimensional nature of the hyperspectral data and to enhance the feature representation capability using linear transformations and activation functions. In addition, HSIMamba employs a bi-directional processing mechanism that improves the model's ability to represent and utilize spectral information through forward and backward spectral dependency capture. Pan-Mamba [103] also provides an innovative network in the pansharpening domain and customizes two crucial components, channel swapping Mamba and cross-modal Mamba, both carefully crafted for efficient cross-modal information exchange and fusion.ChangeMamba [54] explores for the first time the potential of the Mamba architecture for remote sensing change detection (CD) tasks. The MambaBCD, MambaSCD, and MambaBDA network frameworks are designed for Binary Change Detection (BCD), Semantic Change Detection (SCD), and Building Damage Assessment (BDA) tasks, and three spatio-temporal relationship modeling mechanisms are proposed to learn spatio-temporal features thoroughly.ChangeMamba utilizes selective state-space modeling to capture long-range dependent features and maintains linear computational complexity while providing Visual Mamba architecture to learn global spatial context information. Semantic segmentation of remotely sensed images is a fundamental task in geoscientific research.RS3Mamba [67] is a novel two-branch network for semantic segmentation of remotely sensed images. The network incorporates visual state space (VSS) models, especially the Mamba architecture, to improve long-range relational modeling capabilities. In addition, a Co-Completion Module (CCM) is proposed for feature fusion. Experimental results show that RS3Mamba has significant advantages over CNN and transformer-based approaches.Samba [105] is a novel semantic segmentation framework specialized for high-resolution remote sensing images with encoder-decoder architecture. Samba blocks act as encoders to extract multilevel semantic information, and Mamba blocks utilize state space models to capture global semantic information with linear computational complexity.

5   Conclusion

Mamba is gaining prominence in computer vision for its ability to manage long-range dependencies and its significant computational efficiency relative to Transformers. As detailed in recent surveys, various methods have been developed to harness and explore Mambas' capabilities, reflecting ongoing advancements in the field.

We begin by discussing the foundational concepts of SSMs and Mamba architectures, followed by a comprehensive analysis of various competing methodologies across a spectrum of computer vision applications. Our survey encompasses state-of-the-art Mamba models designed explicitly for backbone architectures, high/mid-level vision, low-level vision, medical imaging, and remote sensing. This survey is the first review paper about the recent developments in SSMs and Mamba-based techniques, explicitly focusing on computer vision challenges. Our goal is to generate more interest among the vision community in utilizing the possibilities of Mamba models and finding solutions to their current limitations.

Disclosure of Interests. The authors have no competing interests to declare relevant to this article's content.

References

1. Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.
2. Frank Rosenblatt et al. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*, volume 55. Spartan books Washington, DC, 1962.
3. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
4. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
5. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
6. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
7. Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.
8. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
9. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
10. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
11. Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
12. Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.
13. Maciej Pióro, Kamil Ciebiera, Krystian Król, Jan Ludziejewski, and Sebastian Jaszczur. Moe-mamba: Efficient selective state space models with mixture of experts. *arXiv preprint arXiv:2401.04081*, 2024.
14. Quentin Anthony, Yury Tokpanov, Paolo Glorioso, and Beren Millidge. Blackmamba: Mixture of experts for state-space models. *arXiv preprint arXiv:2402.01771*, 2024.
15. Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.
16. Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
17. Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Swish: a self-gated activation function. *arXiv: Neural and Evolutionary Computing*, 2017.
18. Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.
19. Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.
20. Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pages 28043–28078. PMLR, 2023.
21. David W Romero, Anna Kuzina, Erik J Bekkers, Jakub M Tomczak, and Mark Hoogendoorn. Ckconv: Continuous kernel convolution for sequential data. *arXiv preprint arXiv:2102.02611*, 2021.
22. Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models (2023). *URL http://arxiv.org/abs/2307.08621 v1*.
23. Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. An attention free transformer. *arXiv preprint arXiv:2105.14103*, 2021.
24. Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.
25. Corentin Tallec and Yann Ollivier. Can recurrent neural networks warp time? *arXiv preprint arXiv:1804.11188*, 2018.

26. 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.

27. 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.

28. Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024.

29. Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley. Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695*, 2024.

30. Xiaohuan Pei, Tao Huang, and Chang Xu. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977*, 2024.

31. Ali Behrouz, Michele Santacatterina, and Ramin Zabih. Mambamixer: Efficient selective state space models with dual token and channel selection. *arXiv preprint arXiv:2403.19888*, 2024.

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

33. Badri N Patro and Vijay S Agneeswaran. Simba: Simplified mamba-based architecture for vision and multivariate time series. *arXiv preprint arXiv:2403.15360*, 2024.

34. Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Bjorn Ommer. Zigma: Zigzag mamba diffusion model. *arXiv preprint arXiv:2403.13802*, 2024.

35. Yuan Shi, Bin Xia, Xiaoyu Jin, Xing Wang, Tianyu Zhao, Xin Xia, Xuefeng Xiao, and Wenming Yang. Vmambair: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*, 2024.

36. Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024.

37. Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective ssm. *arXiv preprint arXiv:2403.07487*, 2024.

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

39. Keyan Chen, Bowen Chen, Chenyang Liu, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsmamba: Remote sensing image classification with state space model. *arXiv preprint arXiv:2403.19654*, 2024.

40. Shuangjian Li, Tao Zhu, Furong Duan, Liming Chen, Huansheng Ning, and Yaping Wan. Harmamba: Efficient wearable sensor human activity recognition based on bidirectional selective ssm. *arXiv preprint arXiv:2403.20183*, 2024.

41. Cheng Cheng, Hang Wang, and Hongbin Sun. Activating wider areas in image super-resolution. *arXiv preprint arXiv:2403.08330*, 2024.

42. Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024.

43. Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024.

44. Jiuming Liu, Ruiji Yu, Yian Wang, Yu Zheng, Tianchen Deng, Weicai Ye, and Hesheng Wang. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv preprint arXiv:2403.06467*, 2024.

45. Jinhong Wang, Jintai Chen, Danny Chen, and Jian Wu. Large window-based mamba unet for medical image segmentation: Beyond convolution and self-attention. *arXiv preprint arXiv:2403.07332*, 2024.

46. Yuelin Zhang, Wanquan Yan, Kim Yan, Chun Ping Lam, Yufu Qiu, Pengyu Zheng, Raymond Shing-Yan Tang, and Shing Shin Cheng. Motion-guided dual-camera tracker for low-cost skill evaluation of gastric endoscopy. *arXiv preprint arXiv:2403.05146*, 2024.

47. Yujin Tang, Peijie Dong, Zhenheng Tang, Xiaowen Chu, and Junwei Liang. Vmrnn: Integrating vision mamba and lstm for efficient and accurate spatiotemporal forecasting. *arXiv preprint arXiv:2403.16536*, 2024.

48. Chi-Sheng Chen, Guan-Ying Chen, Dong Zhou, Di Jiang, and Dai-Shi Chen. Res-vmamba: Fine-grained food category visual classification using selective state space models with deep residual learning. *arXiv preprint arXiv:2402.15761*, 2024.

49. Zifu Wan, Yuhao Wang, Silong Yong, Pingping Zhang, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Sigma: Siamese mamba network for multi-modal semantic segmentation. *arXiv preprint arXiv:2404.04256*, 2024.

50. Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. *arXiv preprint arXiv:2403.17839*, 2024.

51. 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*, 2024.

52. Chao Ma and Ziyang Wang. Semi-mamba-unet: Pixel-level contrastive and pixel-level cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. *arXiv e-prints*, pages arXiv–2402, 2024.

53. Ziyang Wang, Jian-Qing Zheng, Chao Ma, and Tao Guo. Vmambamorph: a visual mamba-based framework with cross-scan module for deformable 3d image registration. *arXiv preprint arXiv:2404.05105*, 2024.

54. Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatio-temporal state space model. *arXiv preprint arXiv:2404.03425*, 2024.

55. Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. H-vmunet: High-order vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2403.13642*, 2024.

56. Jiahao Huang, Liutao Yang, Fanwen Wang, Yinzhe Wu, Yang Nan, Angelica I Aviles-Rivero, Carola-Bibiane Schönlieb, Daoqiang Zhang, and Guang Yang. Mambamir: An arbitrary-masked mamba for joint medical image reconstruction and uncertainty estimation. *arXiv preprint arXiv:2402.18451*, 2024.

57. Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024.

58. Mohammad Shahab Sepehri, Zalan Fabian, and Mahdi Soltanolkotabi. Serpent: Scalable and efficient image restoration via multi-scale structured state space models. *arXiv e-prints*, pages arXiv–2403, 2024.

59. Kazi Shahriar Sanjid, Tanzim Hossain, Shakib Shahariar Junayed, Monir Uddin, Dr Mohammad, et al. Integrating mamba sequence model and hierarchical upsampling network for accurate semantic segmentation of multiple sclerosis legion. *arXiv e-prints*, pages arXiv–2403, 2024.

60. Hao Tang, Lianglun Cheng, Guoheng Huang, Zhengguang Tan, Junhao Lu, and Kaihong Wu. Rotate to scan: Unet-like mamba with triplet ssm module for medical image segmentation. *arXiv preprint arXiv:2403.17701*, 2024.

61. Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv preprint arXiv:2402.03302*, 2024.

62. Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:2403.20035*, 2024.

63. Jiacheng Ruan and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

64. Mingya Zhang, Yue Yu, Limei Gu, Tingsheng Lin, and Xianping Tao. Vm-unet-v2 rethinking vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2403.09157*, 2024.

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

66. Tianxiang Chen, Zhentao Tan, Tao Gong, Qi Chu, Yue Wu, Bin Liu, Jieping Ye, and Nenghai Yu. Mim-istd: Mamba-in-mamba for efficient infrared small target detection. *arXiv preprint arXiv:2403.02148*, 2024.

67. Xianping Ma, Xiaokang Zhang, and Man-On Pun. Rs3mamba: Visual state space model for remote sensing images semantic segmentation. *arXiv preprint arXiv:2404.02457*, 2024.

68. Zou Zhen, Yu Hu, and Zhao Feng. Freqmamba: Viewing mamba from a frequency perspective for image deraining. *arXiv preprint arXiv:2404.09476*, 2024.

69. Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. Rs-mamba for large remote sensing image dense prediction. *arXiv preprint arXiv:2404.02668*, 2024.

70. Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

71. Nikola Zubić, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. *arXiv preprint arXiv:2402.15584*, 2024.

72. Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.

73. Wenrui Li, Xiaopeng Hong, and Xiaopeng Fan. Spikemba: Multi-modal spiking saliency mamba for temporal video grounding. *arXiv preprint arXiv:2404.01174*, 2024.

74. Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv preprint arXiv:2403.14520*, 2024.

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

76. Hu Gao and Depeng Dang. Aggregating local and global features via selective state spaces model for efficient image deblurring. *arXiv preprint arXiv:2403.20106*, 2024.

77. Zunnan Xu, Yukang Lin, Haonan Han, Sicheng Yang, Ronghui Li, Yachao Zhang, and Xiu Li. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *arXiv preprint arXiv:2403.09471*, 2024.

78. Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. Scalable diffusion models with state space backbone. *arXiv preprint arXiv:2402.05608*, 2024.

79. Yixuan Li, Weidong Yang, and Ben Fei. 3dmambacomplete: Exploring structured state space model for point cloud completion. *arXiv preprint arXiv:2404.07106*, 2024.
80. Qingyuan Zhou, Weidong Yang, Ben Fei, Jingyi Xu, Rui Zhang, Keyi Liu, Yeqi Luo, and Ying He. 3dmambaipf: A state space model for iterative point cloud filtering via differentiable rendering. *arXiv preprint arXiv:2404.05522*, 2024.
81. Tao Zhang, Xiangtai Li, Haobo Yuan, Shunping Ji, and Shuicheng Yan. Point could mamba: Point cloud learning via state space model. *arXiv preprint arXiv:2403.00762*, 2024.
82. 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.
83. Qiuhong Shen, Xuanyu Yi, Zike Wu, Pan Zhou, Hanwang Zhang, Shuicheng Yan, and Xinchao Wang. Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction. *arXiv preprint arXiv:2403.18795*, 2024.
84. Yuta Oshima, Shohei Taniguchi, Masahiro Suzuki, and Yutaka Matsuo. Ssm meets video diffusion models: Efficient video generation with structured state spaces. *arXiv preprint arXiv:2403.07711*, 2024.
85. Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.
86. Euclid Seeram and Euclid Seeram. *Digital radiography.* Springer, 2019.
87. Rashid N Lui, Sunny H Wong, Sergio A Sánchez-Luna, Gianluca Pellino, Steven Bollipo, Mei-Yin Wong, Philip WY Chiu, and Joseph JY Sung. Overview of guidance for endoscopy during the coronavirus disease 2019 pandemic. *Journal of gastroenterology and hepatology*, 35(5):749–759, 2020.
88. Philip J Withers, Charles Bouman, Simone Carmignato, Veerle Cnudde, David Grimaldi, Charlotte K Hagen, Eric Maire, Marena Manley, Anton Du Plessis, and Stuart R Stock. X-ray computed tomography. *Nature Reviews Methods Primers*, 1(1):18, 2021.
89. Kirsten Christensen-Jeffries, Olivier Couture, Paul A Dayton, Yonina C Eldar, Kullervo Hynynen, Fabian Kiessling, Meaghan O'Reilly, Gianmarco F Pinton, Georg Schmitz, Meng-Xing Tang, et al. Super-resolution ultrasound imaging. *Ultrasound in medicine & biology*, 46(4):865–891, 2020.
90. Arti Tiwari, Shilpa Srivastava, and Millie Pant. Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019. *Pattern recognition letters*, 131:244–260, 2020.
91. 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.
92. Jianhao Xie, Ruofan Liao, Ziang Zhang, Sida Yi, Yuesheng Zhu, and Guibo Luo. Promamba: Prompt-mamba for polyp segmentation. *arXiv preprint arXiv:2403.13660*, 2024.
93. 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*, 2024.
94. Linjie Fu, Xia Li, Xiuding Cai, Yingkai Wang, Xueyao Wang, Yali Shen, and Yu Yao. Md-dose: A diffusion model based on the mamba for radiotherapy dose prediction. *arXiv preprint arXiv:2403.08479*, 2024.
95. Shu Yang, Yihui Wang, and Hao Chen. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. *arXiv preprint arXiv:2403.06800*, 2024.
96. Zhuoran Zheng and Jun Zhang. Fd-vision mamba for endoscopic exposure correction. *arXiv preprint arXiv:2402.06378*, 2024.
97. Weibin Liao, Yinghao Zhu, Xinyuan Wang, Cehngwei Pan, Yasha Wang, and Liantao Ma. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. *arXiv preprint arXiv:2403.05246*, 2024.
98. 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.
99. Jing Hao, Lei He, and Kuo Feng Hung. T-mamba: Frequency-enhanced gated long-range dependency for tooth 3d cbct segmentation. *arXiv preprint arXiv:2404.01065*, 2024.
100. Guangqian Yang, Kangrui Du, Zhihan Yang, Ye Du, Yongping Zheng, and Shujun Wang. Cmvim: Contrastive masked vim autoencoder for 3d multi-modal representation learning for ad classification. *arXiv preprint arXiv:2403.16520*, 2024.
101. Haifan Gong, Luoyao Kang, Yitao Wang, Xiang Wan, and Haofeng Li. nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. *arXiv preprint arXiv:2402.03526*, 2024.
102. 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.
103. 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.
104. Judy X Yang, Jun Zhou, Jing Wang, Hui Tian, and Alan Wee Chung Liew. Hsimamba: Hyperpsectral imaging efficient feature learning with bidirectional state space for classification. *arXiv preprint arXiv:2404.00272*, 2024.
105. Qinfeng Zhu, Yuanzhi Cai, Yuan Fang, Yihan Yang, Cheng Chen, Lei Fan, and Anh Nguyen. Samba: Semantic segmentation of remotely sensed images with state space model. *arXiv preprint arXiv:2404.01705*, 2024.