SCIENCE CHINA
Information Sciences

• REVIEW •

November 2026, Vol. 69, Iss. 11, 211302:1–211302:23
https://doi.org/10.1007/s11432-025-4780-1

State space models meet remote sensing: a survey

Qinzhe YANG[1], Chenyang LIU[1,2,3], Jia XU[4], Zhenwei SHI[2,3] & Zhengxia ZOU[2,3*]

[1]*Shen Yuan Honors College, Beihang University, Beijing 100191, China*
[2]*Department of Aerospace Intelligent Science and Technology, School of Astronautics, Beihang University, Beijing 100191, China*
[3]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China*
[4]*Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing 100094, China*

Received 24 July 2025/Revised 9 November 2025/Accepted 29 January 2026/Published online 20 October 2026

Abstract State space models (SSMs), designed for long-range modeling, offer linear computational complexity and strong capabilities in capturing long-range dependencies. In the field of remote sensing, SSMs have gained popularity due to their effectiveness in addressing unique challenges such as dense visual predictions, multi-modal remote sensing data, and temporal remote sensing data, which have also yielded significant advancements in customized architectures. This study presents a comprehensive review of SSM-based approaches in remote sensing, covering most of the relevant studies since SSMs were first introduced to the field. We offer a multi-dimensional analysis examining SSM applications in remote sensing tasks and discussing advancements in architecture design. This study not only synthesizes the rapid progress in SSM-based research but also identifies key challenges and future opportunities. By providing a detailed perspective, this study aims to serve as a foundational resource for remote sensing researchers, offering actionable insights to foster further advancements in this evolving domain. We will keep tracing related studies at https://github.com/QinzheYang/Awesome-RS-State-Space-Model.

Keywords remote sensing, state space model, Mamba, survey, foundation model

Citation Yang Q Z, Liu C Y, Xu J, et al. State space models meet remote sensing: a survey. Sci China Inf Sci, 2026, 69(11): 211302, https://doi.org/10.1007/s11432-025-4780-1

1 Introduction

State space models (SSMs) are a novel framework introduced to the field of artificial intelligence last year. They achieve linear complexity, making them particularly suitable for tasks involving lengthy sequences [1–6]. They have demonstrated advantages in long-sequence modeling tasks over existing models [7–9]. Since images can also be unfolded into sequential data in a certain order, this efficiency has led to significant success in the vision domain. SSMs have shown remarkable performance in tasks such as image classification, object detection, and semantic segmentation [10–18].

In the field of remote sensing, the beginning of 2024 has seen a notable surge of interest in SSMs, fueled by the unique challenges and demands inherent to this domain [19–24]. Remote sensing tasks are characterized by dense prediction requirements, multi-modal data types, and the need to process temporal remote sensing data [1,20,25–36]. These challenges have necessitated the development of specialized architectures capable of efficiently modeling long sequences while addressing the specific demands of remote sensing applications.

SSM, with its efficient long-sequence modeling capabilities, has been widely studied in various tasks to address the unique challenges mentioned above [37–39]. For instance, in dense prediction tasks such as semantic segmentation, SSM-based backbones can be computationally attractive because their sequence modeling scales approximately linearly with token length, which may ease modeling large spatial extents [40, 41]. Moreover, remote sensing often involves abundant multi-modal data, such as hyperspectral images (HSI), multi-spectral images (MSI), and panchromatic images (Pan). By integrating the spatial and spectral scanning mechanisms, SSMs can establish correlations among these diverse modalities [42, 43]. Additionally, when processing multi-temporal remote sensing information, SSMs can sensitively perceive feature differences across different time periods, leading to numerous advancements in the field [44–46]. Since 2024, the uniqueness of remote sensing has catalyzed a series of model architectural innovations, including novel scanning strategies, hybrid architectures, and lightweight designs, all aimed at improving the efficiency and effectiveness of SSMs in remote sensing applications. Therefore, a comprehensive review of SSMs in remote sensing is urgently needed.

* Corresponding author (email: zhengxiazou@buaa.edu.cn)

© The Author(s) 2026. This article is published with open access at link.springer.com.

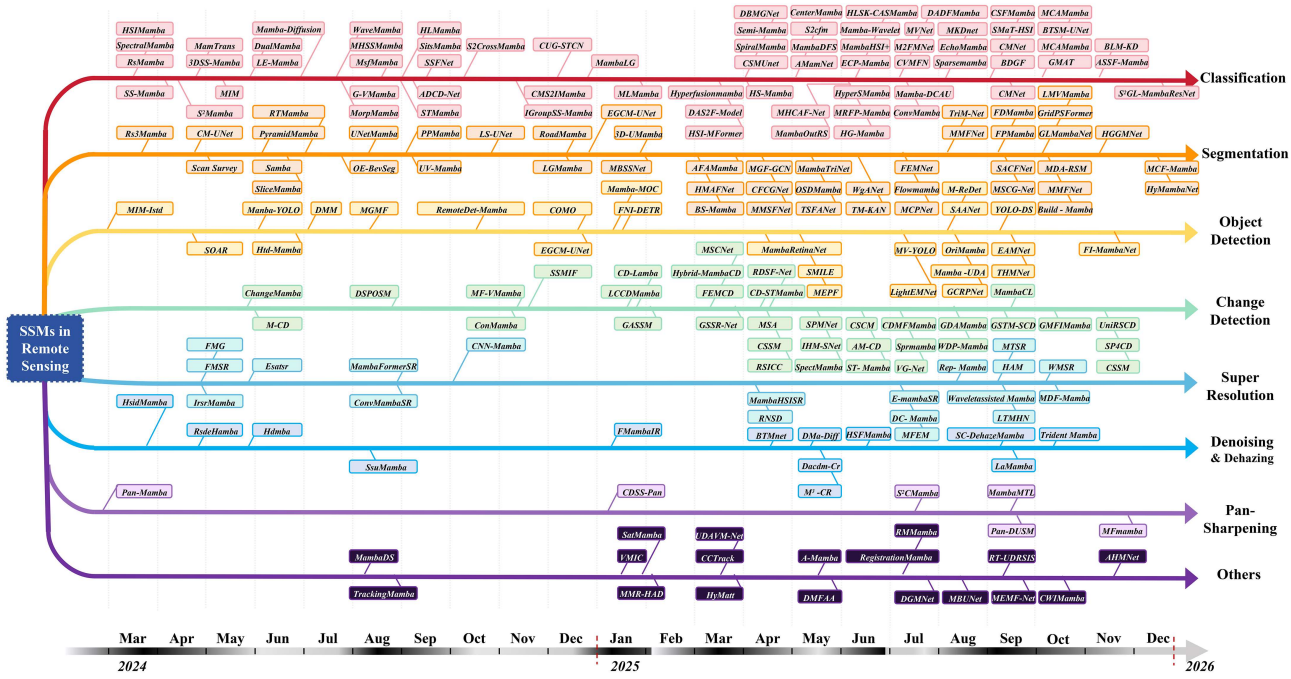

Figure 1 (Color online) The recent development of the SSMs in remote sensing since March 2024. It has showcased new research of SSMs on different remote sensing tasks.

Although there are currently some reviews on SSMs, most of them focus on fields such as computer vision [20, 28, 29, 31, 32]. However, remote sensing images differ from general images and inherently involve characteristics such as multi-modal heterogeneity, frequency domain noise, and long-range dependencies. This necessitates that the design of SSMs takes into account issues such as multi-modal fusion, noise removal, and global-local dependency relationships. Many researchers have conducted excellent studies on RS-SSM to address these issues. Therefore, it is essential to conduct a review specifically on SSMs for remote sensing. Recently, Bao et al. have presented a review on SSMs in remote sensing [47], but their emphasis primarily lies in the comparisons and structural framework of the models across various tasks, such as the specific differences among SSM methods within each task category. In contrast, our study not only outlines the developmental trajectories of methods across different tasks, covering work from March 2024 to December 2025, as shown in Figure 1, but also summarizes eight architectural designs of SSMs. Within each category, we further detail more granular classifications and analyze typical methods along with their strengths.

In this study, we present a comprehensive review and analysis of SSMs in remote sensing, systematically tracing their evolution and aiming to provide researchers with a clear understanding of the advancements over the past year[1)]. We also outline potential future research directions, offering insights into how SSMs can continue to evolve to meet the demands of remote sensing applications. By highlighting the strengths and limitations of existing approaches, this review serves as a valuable resource for researchers to develop SSM-based methods for various remote sensing tasks.

This study is organized along several key dimensions.

- Task-specific adaptations: We analyze how SSMs have been tailored to address the diverse range of tasks in remote sensing, such as classification and object detection, highlighting representative methods. This comprehensive analysis provides valuable guidance for researchers to understand the progress in various tasks, and also helps to identify the strengths of existing methods.
- Architectural innovations: We review eight architectural designs of SSMs. By systematically summarizing and analyzing these innovations, our work offers valuable insights into how these innovations specifically tackle the challenges posed by remote sensing data, including dense predictions, high dimensionality, and temporal variations

1) We conducted a structured literature search for over 300 RS-SSMs between March 2024 and December 2025. We queried common scholarly databases and digital libraries (e.g., Google Scholar, arXiv, IEEE Xplore, and major publishers' portals) using combinations of keywords including "state space model", "Mamba", together with "remote sensing". We included papers that (i) target remote sensing data or tasks, and (ii) use an SSM/Mamba-derived block as a primary modeling component rather than a minor auxiliary module. We removed duplicates across preprint and published versions, and excluded studies that only mention SSMs without presenting an SSM-based method. For each included work, we extracted task type, data modality, scanning strategy, architectural design choices, and the key reported metrics.

in remote sensing information.

• Challenges and opportunities: We summarize the key challenges and opportunities in future research, such as the SSM-based remote sensing foundation model, which have not been addressed in previous reviews and are closely linked to the unique characteristics of remote sensing. Our insights can help researchers quickly familiarize themselves with the latest advances and future directions in this field.

The rest of this study is organized as follows. In Section 2, we outline and analyze the development process from the state space model to Vision Mamba. Section 3 reviews SSMs in remote sensing tasks. Section 4 focuses on the key technological advancements of SSM in remote sensing. Section 5 discusses the current challenges and offers an in-depth analysis of potential research directions for opportunities. Finally, Section 6 concludes the study.

2 From state space model to Vision Mamba

In this section, we will introduce the evolution of SSM, transitioning from models designed for sequential modeling to those specifically tailored for visual modeling.

2.1 Preliminary of SSMs

The state space model is a mathematical model of a physical system that uses state variables to describe how inputs influence the system's behavior over time, typically through first-order differential or difference equations. Originating in the 1950s [19], this approach was applied in fields such as control engineering and system identification. SSM comprises two key equations: (1) the state equation, which defines the relationship between inputs and previous states through matrix operations, (2) the output equation, which determines how the state matrix is transformed into observable outputs:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ \dot{y}(t) &= Cx(t) + Du(t), \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{R}^n$ represents the n state variables, $u(t) \in \mathbb{R}^m$ represents the m state inputs, $y(t) \in \mathbb{R}^p$ represents the p outputs. $A \in \mathbb{R}^{n\times n}$ is the state matrix (controlling the latent state x), $B \in \mathbb{R}^{n\times m}$ is the control matrix, $C \in \mathbb{R}^{p\times n}$ is the output matrix, and $D \in \mathbb{R}^{p\times m}$ is the command matrix. This structure makes SSM highly versatile, leading to widespread applications in various fields over the past 70 years, including control theory, economic forecasting, and robotics [48–50].

2.2 SSMs for sequential modeling

With the rapid growth of sequence length and data volume, improving the efficiency of attention-based models has become an important research direction. SSMs provide a complementary alternative to self-attention by enabling linear-time sequence mixing, making it a compelling choice for sequential modeling. In Transformers, the self-attention mechanism results in computational complexity that scales quadratically with the number of tokens, making it challenging to adapt to long-sequence modeling or prediction tasks. In contrast, SSMs benefit from state and output equations that can be interwoven in a manner that mathematically unfolds into a structure akin to one-dimensional convolution operations, resulting in linear complexity. Therefore, in August 2022, Gu et al. introduced SSMs to sequence processing tasks and developed the S4 model [51] to address the inherent computational complexity of the self-attention mechanism. The S4 model introduces a memory mechanism that incorporates time steps and the HiPPO framework [52], allowing it to compress all currently observed input signals into a coefficient vector. The discretized SSM can be described in the following form:

$$\begin{aligned} h_k &= \bar{A}h_{k-1} + \bar{B}x_k, \\ y_k &= Ch_k, \end{aligned} \tag{2}$$

where $h_k \in \mathbb{R}^{n\times l}$ represents the state variable at step k. $h_{k-1} \in \mathbb{R}^{n\times l}$ denotes the state variable at step $k-1$. $\bar{A} \in \mathbb{R}^{n\times n}$, $\bar{B} \in \mathbb{R}^{n\times d}$, and $C \in \mathbb{R}^{n\times n}$ are the parameter matrices. $x_k \in \mathbb{R}^{d\times l}$ and $y_k \in \mathbb{R}^{n\times l}$ represent the input and output at step k, respectively. l, d, and n denote the sequence length, input vector dimension, and state size, respectively.

At the same time, when SSM is stacked in multiple layers, it can be represented in a form similar to one-dimensional convolution, thereby achieving fast inference.

$$\begin{aligned} y_k &= C\bar{A}^k\bar{B}x_0 + C\bar{A}^{k-1}\bar{B}x_1 + \cdots + C\bar{A}\bar{B}x_{k-1} + C\bar{B}x_k \\ &= \left[C\bar{A}^k\bar{B}, C\bar{A}^{k-1}\bar{B}, \cdots, C\bar{A}\bar{B}, C\bar{B}\right] \times \left[x_0, x_1, \cdots, x_{k-1}, x_k\right]. \end{aligned} \tag{3}$$

Early SSMs did not perform as well as self-attention mechanisms in key modalities such as language. Therefore, in December 2023, Gu et al. introduced two key improvements by proposing the structured SSM (S6), also known as Mamba [1]. At the model level, SSM's parameters are designed to be functions of the input, allowing the model to selectively propagate or forget information throughout the sequence. At the hardware level, it employs a hardware-aware algorithm aimed at minimizing I/O access, thereby achieving further acceleration. In language modeling, S6 achieves an inference speed five times faster than Transformers, while matching the accuracy of Transformers that are twice its size.

2.3 SSMs for vision (Vision Mamba)

In the realm of vision tasks, where images are highly position-sensitive and require comprehensive global context, SSMs have garnered widespread attention for their ability to model long-sequence dependencies with linear complexity. As a result, in January 2024, Zhu et al. sought to introduce SSMs into computer vision applications and developed Vision Mamba [41]. The structure of Vision Mamba is similar to that of Vision Transformer, while the self-attention blocks have been replaced with Mamba blocks. Liu et al. proposed VMamba [53] during the same period, which introduced 2D-selective scanning. This implementation features various scanning techniques through the 2D selective scan (SS2D) block, integrating them with internal visual state-space (VSS) blocks. The design of the model structure and scanning strategy for Vision Mamba and VMamba provides insights for subsequent research.

3 SSMs in remote sensing tasks

In this section, we review the application of SSMs across various remote sensing tasks, including classification, detection, change detection, semantic segmentation, denoising, super-resolution, dehazing, and pan-sharpening. Figure 2 illustrates the typical SSMs methods and the total and average citation counts for papers in each task category. Table 1 [27,42–46,54–93] shows the categories and highlights, and whether the source code is open to the public. It demonstrates the popularity of SSM research in the field of remote sensing.

3.1 SSMs for remote sensing image classification

The research on SSMs in remote sensing image classification broadly falls into two main categories.

The first category focuses on designing different scanning strategies to balance computational efficiency with the ability to capture global context [54, 56, 57, 60, 65, 68]. For example, RSMamba [54] is a typical example of the first approach, which tackled the inherent limitations of SSMs in spatial sensitivity and unidirectional sequence modeling by innovatively implementing multi-directional scanning strategies (forward, reverse, and random). This approach significantly enhanced spatial-context awareness in dynamic remote sensing scenes while establishing a critical paradigm for subsequent SSM-based architectures. Furthermore, RSM-SS [44] and other methods [56, 57, 60, 65, 68] have investigated a range of scanning strategies to better capture context.

The second category focuses on designing micro and macrostructure variations that combine SSMs with methods such as CNNs, Transformers, and graph neural networks (GNNs), allowing the model to leverage the unique strengths of each framework [55, 58, 61, 62, 64, 66, 67, 69]. For example, HSIMamba [55] is a typical representative of this category; it synergistically integrates the sequential processing of spectral bands by SSMs with the spatial attention mechanisms of CNNs. This hybrid design efficiently models both spectral and spatial information in hyperspectral images. Furthermore, this approach has been adopted by methods such as SpectralMamba and 3DSS-Mamba, which apply SSMs to 3D data while retaining their linear computational scalability [56, 60].

Methods primarily based on scanning strategies can alleviate the bias of unidirectional scanning through multi-directional scanning, making better use of image information. These methods are more suitable for sequential data such as HSI. However, they are relatively less sensitive to local texture details. In contrast, methods focused on structural mixing can effectively compensate for modeling local details and key information, but they also increase the computational burden of the model.

Table 1 Introduction of typical SSMs studies in the field of remote sensing. This table shows the categories and highlights, and whether the source code is open to the public.

Task	Method	Highlight	Code
Classification	RSMamba [54]	Designing multi-path activation scanning to model non-causal data	Link
	HSIMamba [55]	Processing remote sensing data bidirectionally	Link
	SpectralMamba [56]	Extracting HSI feature by segmented scanning on frequency bands	Link
	S^2Mamba [57]	Capturing spatial contextual relations by patch cross scanning	Link
	SS-Mamba [58]	Extracting feature by dual-branch spatial and spectral Mamba	Link
	MIM [59]	Designing Mamba-in-Mamba architecture for multi-scale feature	Link
	3DSS-Mamba [60]	Designing spectral-spatial selective scanning	Link
	MamTrans [61]	Integrating SSM with attention to improve accuracy	Link
	DualMamba [62]	Designing cross-attention spectral-spatial Mamba	–
	LE-Mamba [63]	Fusing SSM with local spectral features	–
	GraphMamba [64]	Spatial encoding with weighted multi-hop aggregation (WMA)	Link
	MHSSMamba [65]	Gating mechanisms for spectral-spatial feature enhancement	Link
	WaveMamba [66]	Combining SSM with wavelet feature linear transformation	Link
	MorpMamba [67]	Designing erosion and dilation based feature processing	–
	MSFMamba [68]	Designing multi-scale feature fusion Mamba for multi-modal data	Link
	Enhanced-Mamba [69]	Designing multi-scale feature fusion Mamba for multi-modal data	–
	HLMamba [63]	Designing a Mamba fusion module to fuse multi-modal features	–
Object detection	MIM-ISTD [70]	Designing a Mamba-in-Mamba architecture for small target	Link
	RemoteDet-Mamba [71]	Designing a cross-modal fusion SSM to capture the local features	–
	Mamba-MOC [72]	Designing a cross-scale SSM to capture global and local information	–
	HTD-Mamba [73]	Designing a pyramid SSM to capture and fuse spectral-wise features	Link
	DMM [43]	Difference-guided multi-spectral Mamba feature enhancement	Link
	MGMF [74]	Designing mask-guided fusion SSM block to enhance target features	–
Segmentation	RS3Mamba [75]	Designing SSM with integrated spatial and spectral features	Link
	CM-UNet [76]	Designing U-shaped Mamba to enhance global-local information fusion	Link
	PyramidMamba [77]	Designing multi-scale pyramid fusion to reduce semantic redundancy	Link
	OE-BevSeg [78]	Designing double surround scanning to perceive the surroundings	–
	RTMamba [79]	Designing inverted triangle pyramid pooling feature aggregation	–
	Samba [80]	Combining Mamba block with multi-head self-attention	Link
	UNetMamba [81]	Designing an UNet-SSM decoder for high-resolution image decoding	Link
	UV-Mamba [82]	Introducing deformable convolution to alleviate SSM memory loss	Link
	PPMamba [83]	Combines a local auxiliary mechanism with an omnidirectional SSM	–
Change detection	ChangeMamba [45]	Designing a spatio-temporal modeling scanning strategy	Link
	RSM [44]	Designing a omnidirectional selective scanning module	Link
	RSCama [84]	Designing an SSM with integrated temporal features	Link
	CDMamba [27]	Designing an adaptive global local guided fusion block	Link
	M-CD [46]	Designing a multi-scale score checking module	Link
Super resolution	FreMamba [85]	Designing SSM block with internal embedded Fourier transform	Link
	IRSRMamba [86]	Designing SSM block with internal embedded Wavelet transform	Link
	Esatsr [87]	Enhancing the global-local feature of SSM	–
	MambaFormerSR [88]	Combining attention mechanism with Mamba	–
Denoising	HSIDMamba [89]	Designing 8-directions scanning strategy	–
	SSUMamba [90]	Designing spatial-spectral continuous scanning strategy	Link
Recovery	RSDeHamba [91]	Designing a U-Net structure and Mamba former variant	–
	Hdmba [92]	Designing a novel window SSM to capture local and global correlations	Link
Pan-sharpening	Pan-Mamba [42]	Designing channel exchange and cross-modal scanning strategy	Link
Meteorological downscaling	MambaDS [93]	Using SSM to enhance the utilization of multi-variate correlations	–

Tables 2 [55–58,60–62,64–66,94–100] and 3 [55,57,58,60,61,64,95–101] present comparisons of hyperspectral image classification models on the Houston2013 dataset [94] and the Indian dataset [101], respectively. In all tables in this study, the data are sourced from the original papers and their code repositories, which can be accessed through the GitHub repository we maintain. The shaded cells indicate the top three performing models, with darker colors

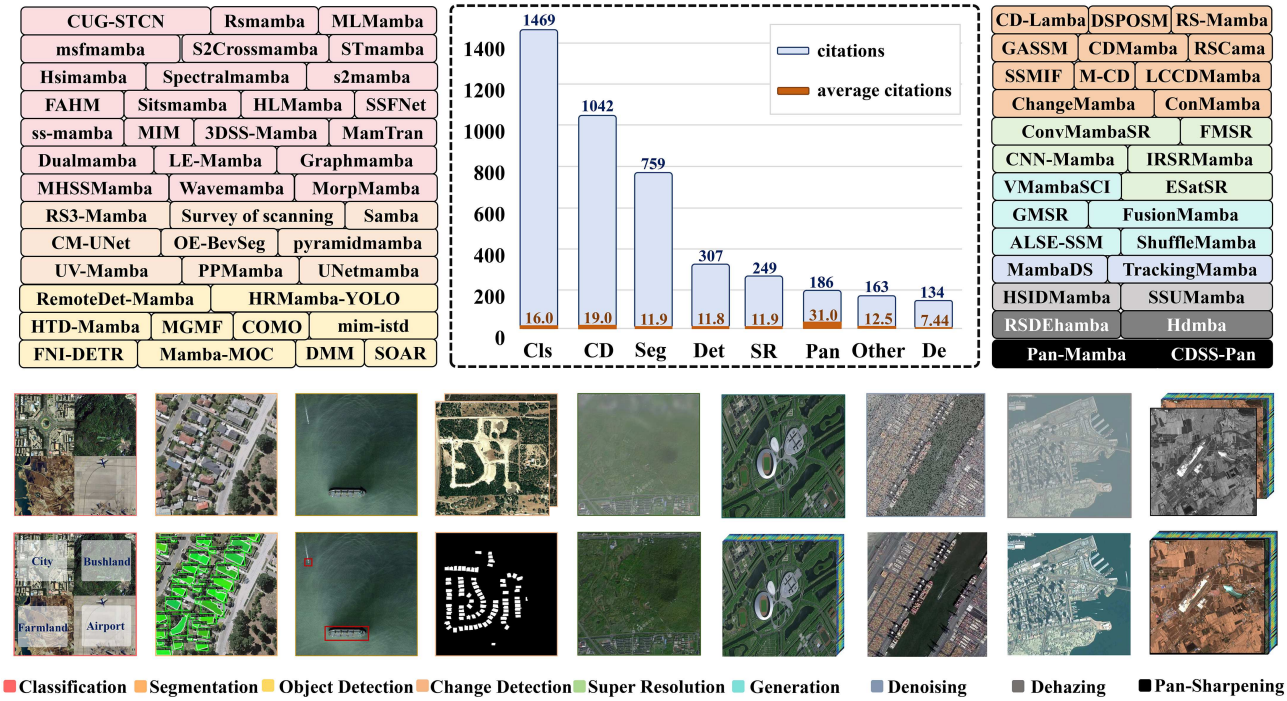

Figure 2 (Color online) Statistics on representative remote sensing tasks addressed by SSMs, including image classification (Cls), segmentation (Seg), object detection (Det), change detection (CD), super resolution (SR), denoising and dehazing (De), and pan-sharpening (Pan). We also report the total and average citation counts for papers in each task category, as of December 18, 2025.

Table 2 Comparison results of classical methods and previous SSMs methods on remote sensing classification dataset Houston2013 [94]. The comparison metrics include: accuracy (%) for each category (C1–C15), overall accuracy (OA) (%), average accuracy (AA) (%), and kappa coefficient (κ).

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	OA	AA	κ
1d CNN [100]	87.3	98.21	100.0	92.99	97.35	95.10	77.33	51.38	27.95	90.83	79.32	76.56	69.47	99.19	98.10	80.04	82.74	78.35
2d CNN [95]	85.1	99.91	77.23	97.73	99.53	92.31	92.16	79.39	86.31	43.73	87.00	66.28	90.18	90.69	77.80	83.72	84.35	82.31
RNN [96]	82.3	94.27	99.60	97.54	93.28	95.10	83.77	56.03	72.14	84.17	82.83	70.61	69.12	98.79	95.98	83.23	85.04	81.83
mini GCN [97]	98.4	99.11	99.60	96.68	97.73	95.10	76.77	68.09	53.92	77.41	84.91	77.23	50.88	98.38	98.52	81.71	83.09	80.18
ViT [98]	82.6	92.82	99.80	99.24	97.73	95.10	76.77	55.65	67.42	68.05	82.35	58.50	60.00	98.79	98.73	80.41	82.50	78.76
Spectralformer [99]	81.9	100.0	95.25	96.12	99.53	94.41	83.12	76.73	79.32	78.86	88.71	87.32	72.63	100.0	99.79	86.14	87.48	84.97
HSIMamba [55]	99.8	99.34	100.0	99.34	100.0	100.0	94.96	95.06	93.86	91.60	97.06	97.69	99.65	100.0	100.0	97.29	97.89	97.06
SpectralMamba [56]	95.9	98.87	98.61	99.05	99.91	94.41	87.03	64.86	80.64	98.17	80.36	81.56	79.65	100.0	98.52	89.52	90.50	88.64
S^2Mamba [57]	83.1	100.0	99.60	98.20	100.0	95.80	89.37	88.60	92.45	92.57	91.56	90.97	89.12	100.0	100.0	93.36	94.09	92.79
SSMamba [58]	92.9	95.99	100.0	99.17	98.29	96.39	91.60	83.33	92.05	98.05	92.48	91.24	92.87	100.0	100.0	94.30	94.96	93.84
3DSS-Mamba [60]	99.0	99.56	100.0	97.68	99.91	98.29	97.55	96.16	96.27	99.55	98.20	98.38	96.68	99.48	100.0	98.37	98.44	98.24
DualMamba [62]	98.8	100.0	100.0	98.75	100.0	100.0	98.55	99.60	98.22	100.0	99.73	99.63	100.0	100.0	100.0	99.57	99.67	99.51
GraphMamba [64]	96.6	98.86	100.0	100.0	100.0	100.0	94.40	83.52	83.36	99.70	99.92	93.57	93.47	100.0	100.0	95.62	96.23	95.27
MHSS-Mamba [65]	97.6	99.84	99.13	99.03	99.83	95.67	93.53	96.94	92.17	99.83	95.95	94.97	89.74	97.66	99.09	96.92	96.73	96.67
WaveMamba [66]	97.9	98.72	100.0	98.39	99.67	93.82	87.85	97.42	91.69	98.20	96.27	98.70	82.90	99.53	100.0	96.39	96.07	96.1
MamTrans [61]	82.7	100.0	98.61	99.81	100.0	95.80	87.59	95.44	83.95	91.99	97.53	90.01	84.91	100.0	100.0	93.41	93.89	92.84

representing higher rankings.

Table 2 showcases the results on the Houston2013 dataset [94], which consists of 15 classes. Among these, C1–C4 represent different vegetations, while C5 and C6 represent natural environments (soil and water). C7 and C8 represent extensive building coverage, which requires contextual scene understanding and a larger receptive field for differentiation. C9–C11 correspond to road systems, both of which are geometrically narrow structures. C12–C15 represent targets with distinct optical features but are relatively small in size. Based on the class-wise performance scores, some notable characteristics of RS-SSMs can be identified. (i) For C1–C4, which possess similar spectral features, 3DSS-Mamba [60] and HSIMamba [55] employ spectral-spatial joint scanning models that effectively reduce misclassifications. (ii) C12–C15 require the model to have strong capabilities for small target learning. Methods

Table 3 Comparison results of classical methods and previous SSM methods on the remote sensing classification dataset Indian [101]. The comparison metrics include: accuracy (%) for each category (C1–C16), overall accuracy (OA) (%), average accuracy (AA) (%), and kappa coefficient (κ).

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	OA	AA	κ
1d CNN [100]	47.83	42.35	60.87	89.49	97.04	59.69	64.81	48.68	44.33	96.30	74.28	15.45	91.11	33.33	100.0	80.00	70.43	79.60	66.42
2d CNN [95]	65.90	76.66	72.39	93.96	97.27	77.23	58.85	57.03	72.87	93.44	100.0	88.18	100.0	84.62	100.0	100.0	75.89	86.64	72.81
RNN [96]	69.00	58.93	77.17	82.33	97.07	69.06	53.56	63.06	65.07	95.06	88.67	50.00	97.78	66.67	81.82	100.0	70.66	76.37	66.73
mini GCN [97]	72.54	55.99	92.93	92.62	98.63	64.71	68.78	63.38	69.33	98.77	87.78	50.00	100.0	48.72	72.73	80.00	75.11	78.03	71.64
ViT [98]	53.25	66.20	88.60	89.71	89.98	72.22	66.00	57.09	57.09	97.53	87.62	63.94	95.56	79.49	90.91	80.00	71.86	78.97	68.04
Spectralformer [99]	70.52	81.39	91.30	95.53	99.32	81.81	75.48	73.76	73.76	98.77	93.17	78.48	100.0	79.49	100.0	100.0	81.76	87.81	79.19
HSIMamba [55]	100.0	80.36	79.89	95.97	90.39	99.99	100.0	99.79	100.0	83.95	51.77	85.45	100.0	79.49	90.91	100.0	89.92	89.82	88.57
S^2Mamba [57]	94.44	100.0	100.0	98.43	100.0	100.0	98.47	98.10	95.04	100.0	97.67	100.0	100.0	100.0	100.0	100.0	97.92	98.88	97.61
SSMamba [58]	100.0	80.30	88.54	99.49	93.69	98.34	100.0	100.0	100.0	89.61	89.51	93.32	99.46	98.63	96.67	99.73	91.59	95.46	90.42
3DSS-Mamba [60]	78.05	90.27	93.04	93.43	97.70	97.26	96.00	99.30	38.89	96.80	98.91	91.39	95.68	99.65	94.52	84.52	95.82	90.83	95.23
GraphMamba [64]	100.0	94.25	94.41	98.39	94.63	99.71	100.0	100.0	100.0	95.33	95.30	93.31	100.0	100.0	99.91	100.0	96.43	97.83	95.91
MamTrans [61]	94.00	100.0	99.45	98.43	99.86	100.0	91.83	97.06	95.92	100.0	98.24	100.0	100.0	100.0	100.0	100.0	97.07	98.42	96.64

such as DualMamba [62], which leverage local feature fusion, achieve notable performance improvements in these scenarios.

Table 3 illustrates the results on the Indian dataset [101], which includes 16 classes. This dataset features crops at different growth stages and has a limited total number of samples. Traditional local convolution-based models or unidirectional spectral modeling approaches tend to struggle with inter-class confusion in such cases. However, SSM's multi-directional scanning effectively mitigates this issue. Additionally, the dataset is highly imbalanced in terms of class distribution; for instance, C1, C7, and C9 each contain no more than 10 images. Nevertheless, SSM exhibits minimal performance discrepancies across different classes, highlighting its ability to learn from a small number of samples. This characteristic warrants further exploration in future studies.

3.2 SSMs for remote sensing image segmentation

Semantic segmentation is a dense image processing task that requires models to simultaneously understand both global semantics and local details. There are two main approaches to incorporating SSMs in semantic segmentation.

The first category focuses on leveraging SSMs and CNN to separately extract global and local information, which are then fused together [75, 82, 83]. For example, RS3Mamba [75] is a typical example of the first approach, which was developed shortly after RSMamba [54] and Pan-Mamba [42]. RS3Mamba employs a residual network as its main branch while utilizing an auxiliary encoder composed of Mamba to enhance global dependencies. RS3Mamba demonstrates superior performance across various datasets, producing less noise and achieving smoother boundaries in segmentation results.

The second category [76, 77, 80] focuses on integrating SSMs with network architectures such as U-Net [102]. For example, Liu et al. developed CM-UNet [76], which also follows a typical U-shaped architecture but replaces the traditional skip connections with SSM blocks. The SSM blocks retain a gated structure, and integrate multi-scale convolutional kernels (3×3, 5×5, and 7×7) to process combined features. The spatial features are then aggregated through mean and max pooling.

The first category balances global semantics and local boundaries, resulting in smoother boundaries. However, the additional branches increase the complexity of feature alignment and fusion, leading to increased inference latency. The second category employs a mature multi-scale framework, but most current implementations only replace parts of SSM locally. There is a need for further research to fully utilize SSM for modeling multi-scale information.

Table 4 [75,79–81,103–107] presents comparisons of these models on the LoveDA dataset [103]. UNetMamba and Samba achieved better results, and their commonality lies in the direct construction of multi-scale features. SSM itself does not incorporate multi-scale information, and this construction method compensates for the limitations of SSM.

3.3 SSMs for remote sensing object detection

Current research on SSMs in object detection broadly falls into two main categories, reflecting different strategies to handle the unique challenges of remote sensing data.

Table 4 Comparison results (%) of classical methods and previous SSMs methods on remote sensing semantic segmentation dataset LoveDA [103].

Method	Background	Building	Road	Water	Barren	Forest	Agriculture	MIoU
PidNet [104]	44.15	54.72	56.54	76.81	17.49	47.13	57.54	50.63
Manet [105]	43.29	55.35	56.2	74.18	14.28	42.85	58.62	49.26
TransUNet [106]	43.00	56.10	53.70	78.00	9.30	44.07	57.85	48.31
RS3Mamba [75]	41.60	58.23	54.03	77.34	17.97	43.81	61.37	50.62
UNetMamba [81]	47.08	59.16	56.74	81.37	18.15	46.61	64.31	53.35
Samba [80]	51.71	57.85	49.86	59.85	21.93	41.00	47.56	47.11
RTmamba-s [79]	42.40	51.06	50.45	76.09	18.25	43.81	58.51	48.65
RTmamba-b [79]	44.72	56.43	52.49	79.90	20.96	44.46	55.06	50.57
SctNet-s [107]	38.17	46.06	51.06	75.70	18.50	42.19	54.44	47.02
SctNet-b [107]	43.37	52.17	50.51	75.71	18.15	43.21	52.97	47.73

The first approach focuses on leveraging SSMs to enable long-sequence modeling, making full use of both local and distant contextual information around each pixel [70, 73]. For example, HTD-Mamba [73] is a typical example of the first approach proposed by Shen et al. For each pixel's surrounding area, HTD-Mamba [73] utilizes a pyramid state space model, performs multi-scale downsampling of the input SSMs token sequence. Following the extraction of features through the SSM, it re-fuses these features via upsampling. This approach effectively leverages both local and remote spatial information surrounding each pixel, highlighting the competitiveness and potential of SSMs in object detection tasks.

The second approach emphasizes combining SSMs with multi-modal information, designing specialized scanning methods to effectively fuse data from different sensors, such as hyperspectral and multi-spectral images [43, 74]. For example, DMM [43] and MGMF [74] are typical examples of the second approach, which further explores the capabilities of SSMs in object detection across multiple modalities. They designed various model variants of SSMs and investigated the effectiveness of multi-modal feature fusion.

The first approach is more favorable for small and dense targets, achieving a balance between the receptive field range and efficiency. The second approach leverages SSM's selective gating to suppress cross-modal redundant information. However, this may lead to fluctuating performance in cases of modality loss or uneven quality.

3.4 SSMs for remote sensing change detection

In SSM-based change detection (CD), a common approach is to arrange multi-temporal data in different sequences and use SSMs to capture the surface changes between multi-temporal remote sensing images.

The ChangeMamba [45] proposed by Chen et al. is the typical SSM model in remote sensing CD tasks, addressing tasks including binary change detection, building damage assessment, and semantic change detection. Three modeling approaches were utilized: (1) spatio-temporal sequential modeling, which unfolds the tokens of two temporal phases of data and then arranges them in temporal order; (2) spatio-temporal cross modeling, where the tokens of the two temporal phases are processed in a cross-ordered fashion; and (3) spatial-temporal parallel modeling, which concatenates the tokens of the two temporal phases along the channel dimension. This exploration has established notable methodologies for subsequent change detection efforts. Unlike ChangeMamba [45], which utilizes SSMs for feature extraction, subsequent studies such as M-CD [46] and RSCama [84] investigate multi-temporal feature fusion through SSMs by defining scanning orders for sequential data to capture spatiotemporal relationships. These methods provide strong evidence for understanding the complex relationships between images at different times. However, it remains to be explored which modeling approach is optimal for specific change detection tasks and diverse types of land cover distribution characteristics.

Table 5 [27,44–46,108–113] presents comparisons of these models on datasets WHU [108], Levir-CD [109], and Levir-CD+ [109]. CDMamba and M-CD both focus on modifying the internal structure of SSM by introducing additional branches and utilizing SSM for the fusion of multi-temporal features, thereby maximizing SSM's capability to capture key differential information.

3.5 SSMs for remote sensing image super resolution

Research on applying SSM to image super-resolution tends to utilize Fourier transforms and other frequency-domain transformation methods, integrating SSM blocks to achieve SSM scanning at the frequency domain level [85, 86].

FMSR [85] is a typical representative and marks the first attempt at applying SSMs in remote sensing image super-resolution. FMSR employs a frequency selection module based on two-dimensional fast Fourier transforms

Table 5 Comparison results (%) of classical methods and previous SSM methods on remote sensing change detection dataset WHU [108], Levir-CD [109], and Levir-CD+ [109].

Method	WHU [108]			LEVIR [109]			LEVIR+ [109]		
	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
FC-EF [110]	91.36	84.10	99.32	88.90	80.03	98.89	76.41	61.82	98.08
IFNet [111]	89.81	81.50	99.16	89.69	81.31	98.96	79.23	65.61	98.34
SwinsUnet [112]	89.93	81.71	99.22	87.77	78.21	98.77	78.31	64.35	98.22
ChangeFormer [113]	90.30	82.32	99.26	88.83	79.90	98.88	77.54	63.31	98.16
RSM [44]	92.79	86.55	99.44	89.77	81.44	98.97	80.91	67.95	98.42
ChangeMamba [45]	92.55	86.13	99.42	90.16	82.09	99.01	80.77	67.74	98.41
CDMamba [27]	93.76	88.26	99.51	90.75	83.07	99.06	83.01	70.95	98.65
M-CD [46]	95.3	91.1	99.6	92.1	85.0	99.2	–	–	–

to process features, while also incorporating a VSSM branch to extract remote features. To effectively handle local features, a hybrid gate module (HGM) is introduced. The subsequent IRSRMamba [86] follows a similar framework, utilizing wavelet transforms.

These approaches utilize frequency domain methods such as the Fourier transform and the wavelet transform, integrated with SSM blocks, and can capture the global structural information of images. SSM can selectively model this information, combined with CNNs to process local details, achieving excellent reconstruction results. However, these approaches face the challenge of effectively fusing features extracted from the frequency domain with those from the spatial domain. Suboptimal fusion strategies may lead to information conflict or insufficient interaction, ultimately hampering the overall reconstruction performance.

3.6 SSMs for remote sensing image denoising and dehazing

Currently, the exploration of SSMs in denoising and dehazing is still in the exploratory stage, primarily focusing on denoising and dehazing.

HSIDMamba [89] addresses HSI denoising through state space model scanning strategies. HSIDMamba [89] incorporates eight different scanning starting points and directions, allowing the extracted features to be processed by the state space model in multiple orientations. Additionally, HSIDMamba [89] incorporates a spectral attention-gating structure to filter out redundant features and enhance denoising performance through average pooling and CNN activation.

Remote sensing image dehazing focuses on removing uneven and physically irregular haze factors to achieve high-quality image restoration. Zhou et al. explored the feasibility of SSMs for dehazing tasks [91]. RSDeHamba [91] has a relatively simple, U-shaped architecture, akin to FusionMamba [114]. The network alternates between vision dehamba blocks and up-sampling techniques during the reconstruction process. The vision dehamba block is a variant of MambaFormerSR [88] and incorporates multi-directional state space model, convolution operations, and channel attention mechanisms.

These studies provide new perspectives for tasks traditionally dominated by CNNs, as the long-range modeling capability of SSM helps address non-uniform noise or haze. Moreover, compared to other tasks, related research is relatively limited, and the model structures and their effectiveness still require further validation.

3.7 SSMs for remote sensing image pan-sharpening

Pan-sharpening involves integrating low-resolution MSI with high-resolution panchromatic images to produce a high-resolution counterpart of the MSI. Pan-Mamba [42] represents the first application of Mamba technology in remote sensing pan-sharpening, introducing channel exchange scanning and cross-modal scanning to enhance the efficient exchange and fusion of cross-modal information in image sharpening. Specifically, the first half of the features from both types of images is concatenated with the second half from the other type, and an SSMs gating mechanism is incorporated to facilitate feature fusion while suppressing redundant features. This feature fusion approach is also reflected in subsequent studies, such as FMG [85], FMSR [85], and IRSRMamba [86].

The above methods in pan-sharpening achieve efficient cross-modal fusion through clever feature concatenation and the SSM gating mechanism. However, these methods may be highly dependent on the specific relationship between panchromatic and multi-spectral images, and adjustments may be needed when generalizing to other modal combinations.

4 Advancements of SSM architecture design in remote sensing

This section will review advancements in SSMs' architecture design for remote sensing. We will organize these advanced methods along two lines: (1) the model structure from the outside in, which includes high-level framework configurations, hybrid architecture designment, and the fundamental component refinement of SSM; (2) the progressive processing of data, which includes scanning strategies for feature extraction, feature fusion, as well as how to remove redundant features and how to handle frequency domain features.

4.1 Scanning strategies for multi-modal remote sensing data

The scanning strategies involve designing different scanning paths for the SSM. The purpose of the scanning strategy is to study specific model observation paths to achieve more efficient extraction of spatial and spectral information. The scanning mode adjusts the order of the input sequence and restores it to its original order after processing [115–118]. Although some methods introduce multi-paths, the increase in complexity still remains insignificant compared to the Transformer [28, 119]. Furthermore, when processing remote sensing data such as hyperspectral imagery, the scanning strategies can integrate spatial and spectral information effectively.

The reason for designing the scanning strategy for SSM is that remote sensing images lack inherent directionality. Relying on a single path sequence may introduce directional bias and lead to inadequate information integration in areas with non-uniform textures. In contrast, employing multi-path or adaptive sequences transforms the two-dimensional structure into a more balanced one-dimensional sequence, thereby reducing causal bias. This improvement enhances global consistency and the separability of small targets in complex remote sensing scenarios, as well as facilitates the efficient processing of high-dimensional data such as HSI. However, excessive scanning paths can easily introduce redundant information. When faced with diverse tasks and data modalities, choosing or combining the optimal scanning strategy poses a significant challenge.

We provide a detailed analysis of the advantages and disadvantages of the various methods, as seen in Table 6 [44,45,53,54,60,69,78,90,120–127]. These advantages and disadvantages are derived from our analysis and summary. Additionally, Figure 3 visually represents all 23 categories of scanning mechanisms employed in Mamba-based architectures for remote sensing tasks. Among these methods, (a)–(g) are regarded as conventional practices commonly employed in computer vision [121, 128, 129], with a primary emphasis on the continuity and directionality of scanning. In contrast, the methods from (h)–(n) place greater emphasis on scanning arrangement strategies, such as forward and reverse asymmetry (h) and (i), scale variation (n), randomness (j) and (k), and orientation (i)–(m). For multi-modal remote sensing data, the six methods labeled (o) explore asymmetric spectral and spatial forward and reverse sweeps. Methods (p)–(r) address specific spatial or spectral dimensions, while the six methods categorized as (s) alternately mix spatial and spectral information from the perspectives of arrangement and combination. Methods (t)–(w) focus on the evolving relationship between space and time.

4.2 High-level framework configurations

High-level frameworks refer to strategies that incorporate SSM modules at a structural level. The main focus is to showcase the model's construction approach. In the field of remote sensing, these variations can be classified into two categories: Backbone-Centric methods and U-Net-based methods.

Backbone-Centric methods focus on feature extraction pipelines [27, 46, 54–58, 70, 79, 89, 127, 130–139]. These approaches embed SSMs as core computational units in backbone networks. Key innovations include: (1) geometric-aware token scanning strategies for remote sensing data alignment [54, 55, 58, 89]; (2) dynamic feature weighting mechanisms for noise suppression [56, 57, 127]; (3) hybrid designs integrate multi-scale modules (e.g., pyramid frameworks) to compensate for SSM's local feature limitations [44, 75, 77, 84].

U-Net-based methods integrate SSM modules into U-Net architectures through two primary strategies [45,76,80, 85,86,90,91,114,124,140,141]: (1) the first strategy retains the complete U-Net [102,142] framework but replaces its central computational layers with SSM blocks, as seen in designs using residual blocks with integrated SSMs [92]; (2) the second strategy deploys U-Net components for specialized purposes, such as parallel U-Net branches for spatial/spectral feature isolation [114], and combining SSMs with convolutional layers in skip connections to refine multi-scale fusion [143].

Overall, Backbone-Centric methods treat SSM as an efficient feature extractor, with a design that prioritizes the construction of hierarchical feature representations. Although structures such as FPN can be added to obtain multi-scale features, their feature fusion mechanisms are typically not as native and straightforward as those of U-Net-like methods. U-Net-based methods can better integrate multi-scale information, but designing complex structures, such as parallel branches, can add additional complexity and training difficulties for SSMs.

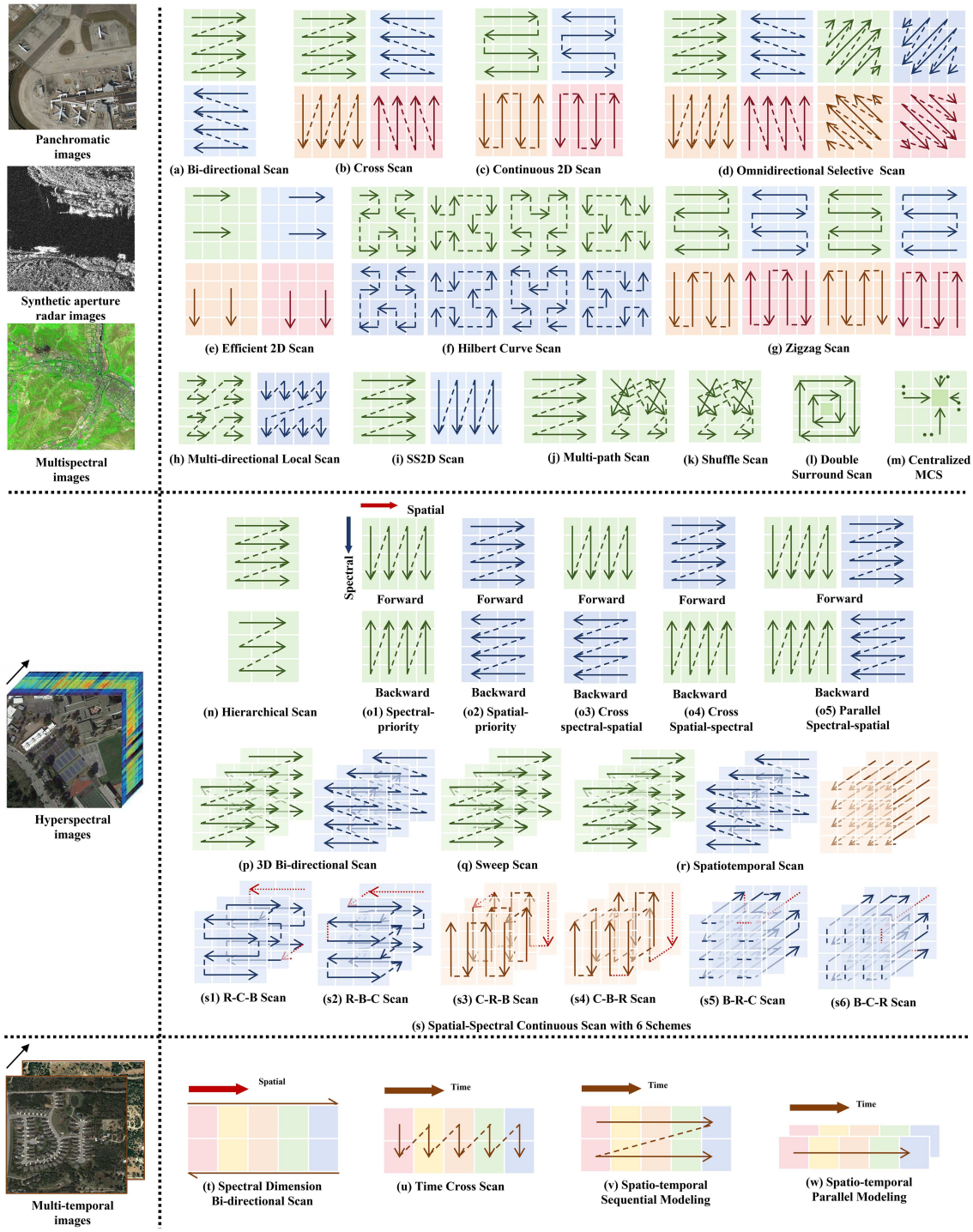

Figure 3 (Color online) Comparison of different SSMs' scanning strategies in remote sensing. Methods (a)–(m) are applied to the processing of 2D remote sensing data, while methods (n)–(w) are applied to the processing of others, including hyperspectral imagery, time series data, and other forms of information.

4.3 Hybrid architecture designment

The hybrid architecture designment aims to combine SSMs with classical structures such as CNN and Transformer, either in parallel, serial, or nested ways, in order to compensate shortcoming of pure SSMs. In this subsection, we review hybrid architectures that combine SSMs with other powerful techniques: CNN for enhanced local understanding, Transformer for enhanced global understanding, and GNN for graph-structured data.

In SSM-CNN hybrids, convolutional inductive biases enhance local feature extraction, which serves as a critical

Table 6 Comparison of advantages and disadvantages of typical scanning strategies in remote sensing.

Scanning method	Strengths	Weaknesses
(a) Bidirectional Scan [120]	Compared to the simple forward and reverse directions, redundancy in information is reduced because of the asymmetry of the paths.	Artifacts or changes in resolution may lead to increased computational complexity.
(b) Cross Scan [53]	Enhancing the capability to capture multi-directional spatial information.	The increase of computational complexity may necessitate additional steps to integrate information from cross-scans.
(c) Continuous 2D Scan [121]	Enhancing spatial continuity, ensuring no jumps in the field of view.	The increase of computational complexity may not effectively capture the discrete spatial structure when the image is large.
(d) Omnidirectional Selective Scan [44]	Aligned with the distinct physical characteristics of remote sensing images, regardless of the horizontal direction.	The increase of computational complexity may introduce redundant information.
(e) Efficient 2D Scan [122]	Both local and global contexts are effectively captured while reducing computational costs.	Some patches are omitted, resulting in a loss of spatial continuity.
(g) Zigzag Scan [123]	Providing comprehensive 2D spatial coverage while capturing both horizontal and vertical dependencies.	Increasing computational complexity.
(h) Multi-directional Local Scan [69]	Capturing fine-grained details and ensuring efficient handling of local regions.	Long-range dependencies may be overlooked, necessitating careful resizing of the window.
(i) SS2D Scan [124]	Maintaining two-dimensional spatial dependencies by scanning in multiple directions to improve feature discriminability.	Increasing computational complexity.
(j) Multi-path Scan [54]	The information from these different paths mitigates the effects of causal sequences.	Increasing computational complexity.
(k) Shuffle Scan [125]	Reduce the inductive bias caused by the causal scanning method during multi-modal fusion.	Weak to learn spatial continuity.
(l) Double Surround Scan [78]	Prioritizing the modeling of relevant hierarchical environmental components enables the model to capture the spatial relationships of BEV features, enhancing its context-aware capability to better distinguish between foreground and background.	Weak to capture dependence on long-distant pixels
(n) Hierarchical Scan [126]	Local and global features are captured at different scales, providing effective features for complex images.	Increasing model complexity and computational requirements necessitate careful design to balance local and global information.
(o) Spatial-Spectral Scan [60]	Instead of employing a multi-directional scan, a new direction is used to replace the original reverse scan, reducing computational complexity.	Weak to capture spatial and spectral information.
(p) 3D Bi-directional Scan [120]	The spatial relationships between 3D data are captured to effectively process volumetric image data.	Capturing 3D information requires increased memory capacity.
(q) Sweep Scan [90]	The bidirectional spatial dependencies of the HSI are effectively captured.	The long-sequence spectral dependencies are overlooked.
(s) Spatial-Spectral Continuous Scan [127]	Preserving the spatial-spectral correlation within the local context and enhancing local texture exploration.	The ability to capture long-distant spatial correlation is diminished.
(u) Time Cross Scan [45]	Temporal cross-scanning enhances the understanding through time dimension.	Weak to capture spatial information.

supplement to SSMs' weaker spatial awareness [43,76,85,91,92,127,144–149]. Techniques like depthwise convolution (DW) for multi-scale perception [55, 68] and pointwise convolution (PW) for cross-modal interaction [150] are systematically employed. This can be seen in HSIMamba's [55] hierarchical feature refinement and SpectralMamba's [56] dynamic gated spectral-spatial fusion. Meanwhile, channel attention mechanisms and dual-branch designs have become ubiquitous in advanced frameworks. This is exemplified by DualMamba's CNN-SSM co-training [62] and SSFNet's DW-PW cross-modal fusion [150]. Another method, such as PyramidMamba [77], uses SSM to process long-distance or multi-scale features, enabling effective extraction and fusion of global contextual information. After that, CNN is applied to further process these features, capturing fine-grained spatial characteristics.

In SSM-Transformer hybrids, a common approach is to explore mixed architectures, which combine the global modeling capabilities of the self-attention mechanism with the selective information acquisition of SSM: MamTrans adds a Transformer branch outside the order-sensitive SSM branch, both extract features in parallel [61]; while MHSSMamba [65] and RS3Mamba [75] synergize spectral-spatial attention with SSMs' long-sequence dependencies for enhanced segmentation granularity. Architectures like MambaFormerSR further innovate through nested SSM-attention structures and sparse cross-attention variants [80, 87–89, 151–153].

In SSM-GNN hybrids, Yang et al. proposed GraphMamba, which replaces the residual normalization layer and multi-head attention mechanism layer of ViT with SSM and graph network, respectively [64]. A filter matrix that changes with the nodes of the graph is designed to better represent the global information and abstract the

relationship between different nodes, which is conducive to the extraction of spatial [64, 154–157].

The purpose of the hybrid architecture of SSM is to balance the long-range context of remote sensing images with the texture and boundaries of fine-grained objects. In this framework, SSM is responsible for long-range modeling, while CNNs, deformable convolutions, or sparse attention mechanisms focus on local refinement. This approach enables the model to simultaneously capture both global and local information effectively. This cross-paradigm integration systematically addresses SSMs' limitations while preserving their efficiency advantages, establishing hybrid architectures as the dominant paradigm in contemporary remote sensing SSMs research. Furthermore, effectively integrating models of different paradigms while balancing their contributions poses a significant design challenge. Additionally, the introduction of complex modules like Transformers may partially negate the original efficiency advantages of SSM, requiring careful design to maintain high efficiency. Moreover, the internal working mechanisms of hybrid architectures and the interactions between different components still need more in-depth theoretical research.

4.4 Fundamental component refinement

Fundamental component refinement represents an enhancement that begins with the smallest structural unit of the network. It involves adjusting the structure of SSMs through four methods: deletion, modification, addition, and complementation.

Deletion aims to reduce the computational load by removing unnecessary components, such as linear projection layers. As shown in Figure 4, models such as SS-Mamba (Figure 4(b2)) [58], RS3Mamba (Figure 4(b3)) [75], and RSDeHamba (Figure 4(b4)) [91] have implemented similar strategies. Additionally, since Mamba architectures are often integrated with other structures, to streamline the architecture, internal normalization layers can be removed if external normalization is already applied.

Modification focuses on altering internal layers. For instance, GraphMamba (Figure 4(c2)) [64] replaces the skip connection with an adaptive residual branch, which enhances the model's capability to extract features. Besides, changing the scan order is categorized as a variant of the scan strategy mentioned earlier (Figure 4(c1)).

Addition involves incorporating an extra branch into the existing framework, introducing a new trainable computational module to better manage the original information [71, 72, 158, 159]. For example, Cdmamba (Figure 4(d1)) [27] introduces an additional two-dimensional convolutional and activation branch, while FMSR (Figure 4(d2)) [85] enhances the learning of spectral features by integrating Fourier transform branches. Additionally, GMSR (Figure 4(d3)) incorporates both spectral and spatial branches [141].

Complementation involves providing supplementary inputs [125]. When the newly added modal information is removed, we can observe that this type of SSM can be approximated as a vanilla SSM. For instance, Pan-Mamba (Figure 4(e1)) [42] and FusionMamba (Figure 4(e2)) [114] leverage the differences among them or the characteristics of various modalities and visual field information to enhance and complement each other's data.

Fundamental component refinement is quite challenging because improper designment may disrupt the flow of critical information in the model, impacting performance. Guiding the refinement of SSMs from a theory-driven perspective could offer a more reliable and promising path for future development.

4.5 Feature fusion of multi-modal remote sensing data

Multi-modal fusion encompasses pixel-level, feature-level, and decision-level integration. In the context of SSM, the current emphasis is on feature-level fusion. This is because SSMs place greater importance on global spatial information, and such feature fusion can supplement SSMs with more spectral/multi-modal information. Pixel-level and decision-level fusion require further research. The feature-level fusion approaches can be categorized into complementary use of data from different time periods, spatial features, and spectral features, as well as the integration of information from different modalities [42, 68, 84, 125]. Table 7 [42–46,58,61,62,65,67–69,84,114,125,127,160,161] summarizes the information categories, and the main idea used by various models [161–167].

The purpose of multi-modal fusion in SSM is to address the challenges of cross-modal alignment and redundant information filtering in remote sensing images, where noise is easily amplified post-fusion [168–170]. By employing cross-modal guided selective gating and dedicated scanning paths, the approach suppresses channels or regions with low mutual information while enhancing complementary frequency bands. However, this method relies on alignment quality and hyperparameter tuning, resulting in a training cost that is higher than that of single-modal approaches. In addition, under the survey framework adopted in this study, the multi-modal fusion mentioned in Section 3 specifically refers to methods that arise in particular tasks, without focusing on data modalities, but rather on task categories. In contrast, the multi-modal fusion discussed in this section inclined more towards focusing on fusion methods rather than the tasks themselves.

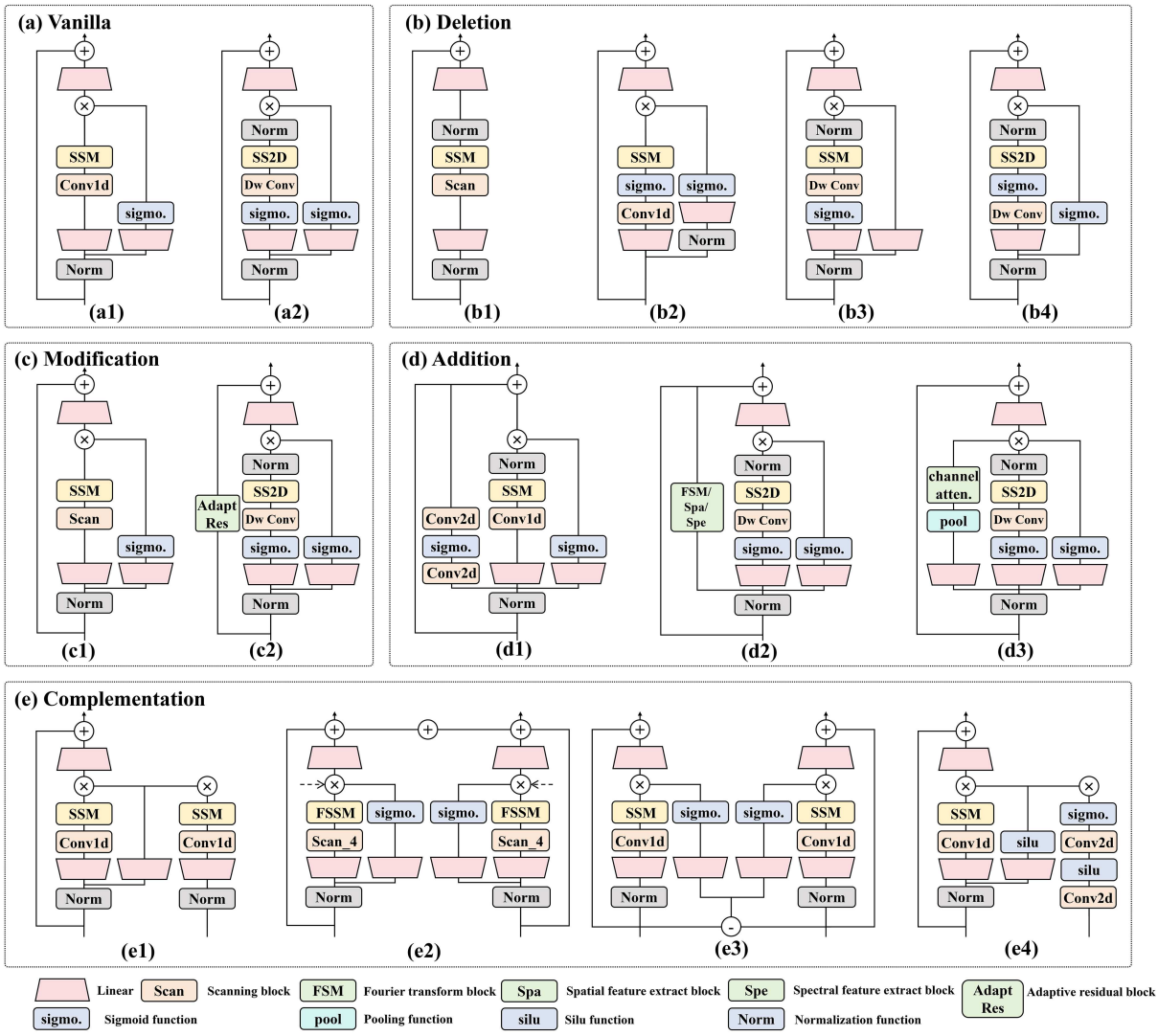

Figure 4 (Color online) Comparison of various methods for internal deformation of SSM. (a) vanilla: (a1) vanilla Mamba; (a2) its second generation, VMamba; the four specific methods of adjusting the structure of SSM: (b) deletion, (c) modification, (d) addition, and (e) complementation. (b1) DualMamba; (b2) SS-Mamba; (b3) RS3Mamba; (b4) RSDeHamba; (c1) RSMamba; (c2) GraphMamba; (d1) Cdmamba; (d2) FMSR; (d3) GMSR; (e1) Pan-Mamba; (e2) FusionMamba; (e3) DMM; (e4) MGMF.

4.6 Gated redundancy removal

In multi-directional scanning of remote sensing data, the combination of multi-modal data and various hybrid structures inevitably results in the generation of redundant information [171, 172]. As a result, many studies in SSMs employ a gating mechanism to filter out redundant information, a process we named gated redundancy removal.

Gated redundancy removal can be either conducted within the main branch or the auxiliary branches of the networks. The extracted features are simultaneously fed into two branches, where the auxiliary branch utilizes dual pooling, filtering, or mixing with other modalities for further processing. Meanwhile, the main branch decides which parts of the extracted features to retain or discard, based on signals from the auxiliary branches. At the junction where the main path and the auxiliary branches converge, a gated element-wise multi-plication is typically implemented. In addition, a threshold can be introduced to further enhance the information screening process, filtering out unnecessary information. Based on the characteristics of the auxiliary branches, we categorize these into two specific approaches, including self-guidance and cross-guidance, as shown in Figure 5. The purpose of gated redundancy removal is to reduce significant redundancy and ineffective calculations brought about by large-scale or multi-modal inputs. This can decrease computational load and memory usage, maintain performance in key areas, and enhance the efficiency of remote sensing interpretation. However, the selection of features and the setting of

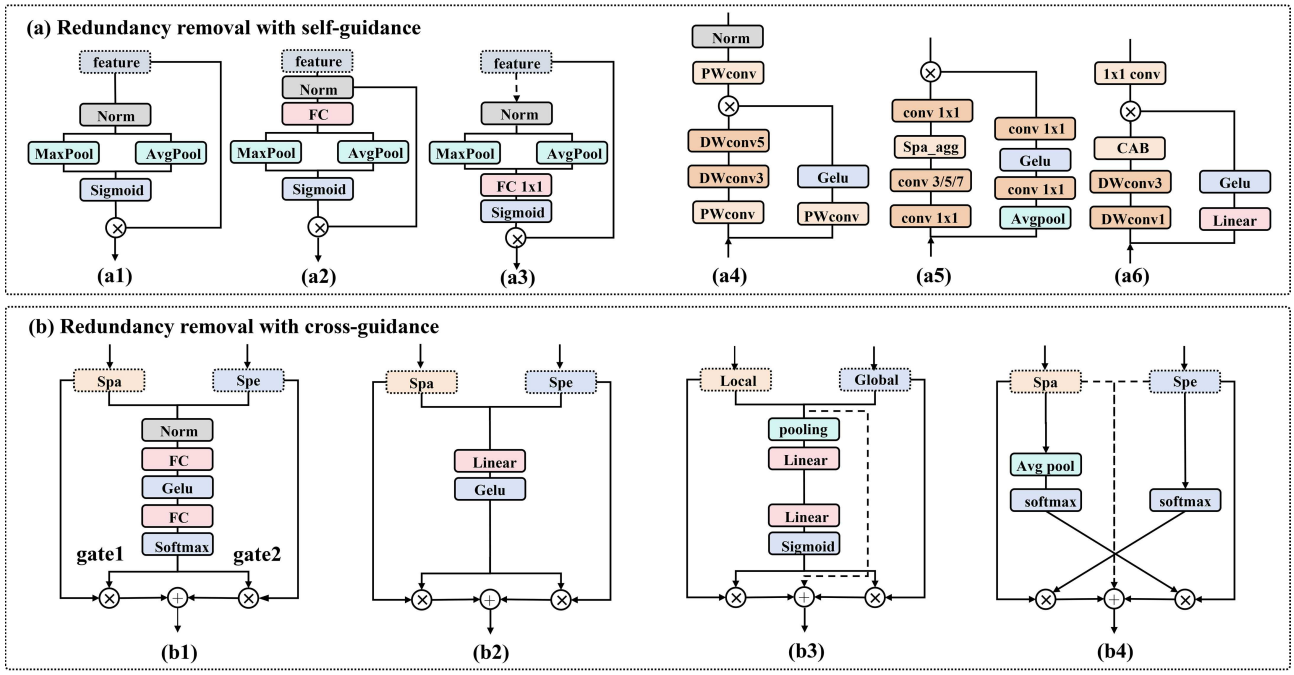

Figure 5 (Color online) Comparison of the redundancy removal methods, including self-guidance and cross-guidance, where spa and spe represent spatial and spectral features, respectively. (a1) GMSR's spatial gradient attention; (a2) M-CD; (a3) spectral gradient attention; (a4) LE-Mamba's MSCGU modules; (a5) CM-UNet; (a6) FMSR; (b1) S^2Mamba; (b2) SS-Mamba; (b3) DualMamba; (b4) Spectral-Spatial Mamba.

Table 7 Comparison of different methods for feature fusion of multi-modal remote sensing data.

Method	Category	Main idea
SSRFN [160]	HSI/MSI	Constructing a Mamba cross-attention network to fuse features
Msfmamba [68]	HSI/LiDAR	Designing a cross-guided Mamba for fusion
Shufflemamba [125]	LRMS/PAN	Designing a cross-guided Mamba for fusion
Pan-Mamba [42]	IRMS/PAN	Designing a cross-guided Mamba for fusion
DMM [43]	MSI/RGB	Designing a cross-guided Mamba for fusion
TFFNet [161]	HSI/SAR	Designing a squeeze-and-excitation transform fusion module
RSM [44]	Bi-temporal	Performing element-wise addition on the features and decode
ChangeMamba [45]	Bi-temporal	Concatenating the multi-temporal features and decode using SSM
RSCama [84]	Bi-temporal	Sorting tokens in a specific order and then processing through SSM
M-CD [46]	Bi-temporal	Concatenating the multi-temporal features and decode using SSM
LE-Mamba [69]	Spectral/spatial	Fusing the features through element-wise multiplication
FusionMamba [114]	Spectral/spatial	Designing a cross-guided Mamba for fusion
SS-Mamba [58]	Spectral/spatial	Modulating spatial and spectral tokens using center region information
GMSR [127]	Spectral/spatial	Concatenating the spatial and spectral features
MamTrans [61]	Spectral/spatial	Concatenating the spatial and spectral features
DualMamba [62]	Spectral/spatial	Designing a cross-guided Mamba for fusion
MHSSMamba [65]	Spectral/spatial	Constructing a cross-attention network to fuse features
Morpmamba [67]	Spectral/spatial	Constructing a cross-attention network to fuse features

thresholds have a significant impact on the performance, necessitating careful tuning.

Self-guidance focuses on redundancy removal within a single stream by comparing different transforms of the same feature. Such methods commonly utilize a variety of dual feature processing techniques, including average pooling versus max pooling, as well as CNN features from different scales. For example, GMSR's [127] spatial gradient attention (Figure 5(a1)), spectral gradient attention (Figure 5(a3)), and M-CD (Figure 5(a2)) [46] utilize average pooling and max-pooling operations along the channel dimension to effectively address channel information. LE-Mamba's MSCGU modules (Figure 5(a4)) [69], CM-UNet (Figure 5(a5)) [76], and FMSR (Figure 5(a6)) [85] utilize CNNs with varying kernel sizes within the main branch to aggregate information from different receptive fields, enabling the capture of a wide range of features.

Cross-guidance utilizes information from another stream to guide redundancy removal, such as from a different modality or scale. Leveraging the differences between these modalities and scales facilitates a comprehensive control feature process. Specifically, S^2Mamba (Figure 5(b1)) [57] filters out low-contribution features by artificially defining a threshold at the gate. This gating mechanism effectively truncates redundant features that may hinder HSI classification while promoting the integration of spatial and spectral attributes. SS-Mamba (Figure 5(b2)) [58] utilizes the average value of spectral information and the intermediate value of spatial information to filter out the low-contribution features of the opponent. DualMamba (Figure 5(b3)) [62] and Spectral-Spatial Mamba (Figure 5(b4)) [56] employ pooling methods to achieve a broader range of feature capture.

4.7 Processing in frequency domain

Remote sensing images inherently contain noise and distortions introduced by complex acquisition conditions, such as atmospheric scattering and sensor noise. These interferences manifest in the frequency domain as anomalous energy distribution patterns, such as random noise in the high-frequency range or smooth distortions in the low-frequency range [173]. SSMs can leverage their long-sequence modeling capabilities to learn the relationships between multiple frequency ranges. Therefore, SSM frameworks increasingly integrate domain-specific transformations to address these challenges [85, 86, 88]. FMSR [85] embeds Fourier transforms in the SSM block, to decompose global-local spectral features through frequency-domain analysis, while MambaFormerSR enhances this paradigm with frequency band attention for adaptive feature refinement [88]. Extending beyond spectral processing, IRSRMamba employs wavelet transforms to capture multi-directional texture patterns, effectively generalizing frequency-based approaches to spatial-textural domains [86]. WMSR [174] uses a discrete wavelet transform to separate features into low-frequency and high-frequency components in the frequency domain, and processes the low-frequency feature with SSM. These methods enable SSMs to simultaneously focus on spatial and frequency domain features.

5 Challenges, opportunities and limitations

In this section, we will discuss the challenges, opportunities, and limitations, including the SSM-based remote sensing foundation model, efficient SSMs for edge devices and large-scale image processing, the SSM-driven remote sensing vision-language model, and limitations of applying SSMs in remote sensing.

5.1 SSM-based remote sensing foundation model

The foundation model represents a new trend in the development of remote sensing, defined as models with billions of parameters, pre-trained on large-scale datasets, and capable of adapting to various downstream tasks. However, current methods employ Transformer architectures rather than SSMs, and they face significant limitations. The performance of the Transformer diminishes in remote sensing applications with numerous spatially dispersed small targets [175]. While increasing the input resolution can mitigate this issue, it leads to high costs due to quadratic growth in GPU consumption and computational complexity [175–190]. This drastically reduces the efficiency of remote sensing foundation models in tasks such as wide area data processing and dense prediction. In contrast, the structure of SSMs is designed for long-distance modeling and features approximately linear computational complexity, making it inherently suitable for these tasks. At the same time, SSMs possess selective modeling capabilities, making them inherently well-suited for the detection of sparse targets in remote sensing. Therefore, designing an SSM-based remote sensing foundation model is crucial [12, 61, 191, 192].

The challenges and opportunities of SSM-based remote sensing foundation models can be summarized in three key points. Firstly, scaling to global-scale datasets is essential for remote sensing. This can help remote sensing models generalize across the globe rather than overfitting to a few regions. It ensures that RS-SSMs can support mission-critical applications such as deforestation monitoring, maritime domain awareness, and disaster response on a planetary scale. Secondly, current models often suffer from the issue of long sequence forgetting, particularly in scenarios involving sparse small targets or densely arranged repetitive targets in remote sensing images. Thus, exploring mechanisms for context scanning or fusion methods designed specifically for ultra-long distances is crucial for advancing foundational models. One feasible approach is to leverage SSM's selective information capture and selective activation of parameters to construct a remote sensing backbone model with long-range modeling capabilities, making it particularly suited for tasks such as ship detection on global-scale datasets. Thirdly, as the number of parameters in foundation models and the amount of data increase rapidly, pretraining with fine-tuning is considered a viable approach. However, existing mainstream efficient fine-tuning methods, such as adapters and LoRA, do not support training for SSM models, indicating a significant need for further research [193–197]. Current

work has demonstrated that SSM has the capability for large-scale data learning [175]. It is important to explore efficient pre-training and fine-tuning techniques specifically for SSM. For example, assigning an increment to the parameter matrix of SSM during fine-tuning and allowing only the increment to be learned could be beneficial.

5.2 Efficient SSMs for edge devices and large-scale image processing

In many remote sensing scenarios, particularly those involving large-scale image processing and edge devices, efficient models have always garnered significant attention. The SSMs possess natural advantages. The existing challenges and opportunities can be summarized as follows.

In SSMs' acceleration techniques, there have been several studies on acceleration methods for 2D image processing in the field of computer vision [186]. However, remote sensing data are more complex and diverse. SSMs were designed with hardware acceleration at the beginning [1], and there has already been progress in further accelerating SSMs for hardware [198]. The research on acceleration techniques for multi-modal remote sensing data (such as HSI, MSI, text data, and SAR) holds great promise, especially when it fully leverages hardware efficiency [199, 200]. One potential area of exploration is the development of transmission algorithms that enable the adaptation of 2D and 3D data to the I/O layer of SSM. In this context, transmission algorithms refer to the process where data are transferred between the CPU and GPU during certain matrix operations. By designing specific operators, the computation order after uploading to the GPU can be altered to reduce the transfer time between different hardware.

At the same time, model lightweight techniques, including distillation, pruning, and quantization, have found various applications in remote sensing scenarios [201–203]. Since methods such as pruning and distillation need to be designed based on the structural characteristics of different models, methods tailored specifically for the structure of SSMs and particular types of remote sensing data remain to be further explored [204–210]. In the future, in lightweight methods such as distillation, introducing simulated remote sensing noise as interference could compel the SSM, functioning as the student model, to concurrently learn its adaptability to remote sensing images. Additionally, pruning and quantization can be utilized to enhance the inference speed of the SSM, making it suitable for time-sensitive practical deployment scenarios in remote sensing. Moreover, by designing new operators that enhance the parallelism of the internal state equations and output equations of SSM in conjunction with specific hardware, it could be better suited for deployment in resource-constrained remote sensing scenarios.

5.3 SSM-driven remote sensing vision-language model

Currently, in the field of visual-language models specifically for remote sensing, contrastive models and autoregressive models are the two dominant approaches. SSMs present both advantages and challenges in relation to these two models.

Contrastive models like CLIP align the embeddings of images and text through contrastive learning. In the context of remote sensing, the complexity of the scenes and their broad coverage require lengthy textual descriptions. However, not every word in the long text description is relevant to the corresponding image (such as adjectives or background information). CLIP's text encoder, which is based on Transformers, treats all tokens equally and lacks the ability to dynamically filter them. In contrast, SSMs possess the capability for selective modeling of information, and their linear complexity makes them well-suited for handling long text descriptions in remote sensing applications. Research on SSMs in this area is relatively sparse, and how to achieve effective modality alignment warrants further investigation.

Autoregressive models are currently the most common architecture in multi-modal generative modeling [211–214]. These models primarily rely on Transformer backbones, leading to a high computational complexity. This issue is particularly pressing in the current era of large models, where today's vision-language models (VLMs) possess a significant number of parameters and computational demands. Furthermore, the complexity of remote sensing scenarios, which necessitates the generation of long detailed texts, further exacerbates the computational burden. In contrast, SSMs, as a type of efficient sequence modeling framework, are inherently well-suited for autoregressive architectures for generating long detailed texts. However, there is limited existing research on the design of SSM architectures for autoregressive modeling, specifically in the context of remote sensing. Recently, DeepSeek-OCR [215] has explored the feasibility of processing text as images, achieving a tenfold model compression. Similarly, SSM also possesses efficient information summarization capabilities. The exploration of SSM as a purpose encoder for large-scale remote sensing images and textual data represents a viable approach.

5.4 Limitations of applying SSMs in remote sensing

Despite the rapid development of RS-SSMs, there are still limitations and open issues in applying SSMs in remote sensing, including scale variation, fine-grained texture modeling, and long-sequence stability.

SSMs are challenged first by the extreme scale variation and strong geometric diversity in large remote sensing images, which makes purely sequential modeling difficult to consistently excel. Many SSM-based designs still rely on additional locality and multi-scale priors to achieve strong dense prediction performance [216]. However, it remains unclear when a pure SSM backbone is preferable to a hybrid design, and how to combine SSMs with local operators or multi-scale mechanisms without sacrificing efficiency, remains an open problem.

SSMs also face open questions regarding long-sequence stability and memory behaviors, especially when remote sensing images are converted into extremely long sequences. For very large images, sparse selection mechanisms in SSMs may introduce information loss or instability as sequence length grows, and robustness across different sensors, resolutions, and acquisition conditions remains insufficiently understood. Addressing these limitations requires future SSM research to more systematically evaluate stability under ultra-high-resolution remote sensing images.

SSMs further encounter limitations in capturing fine-grained textures and high-frequency details that are crucial for remote sensing interpretation, such as building edges, road boundaries, and subtle material patterns in HSI. Scanning strategies and long-range mixing in SSMs alone may be less sensitive to these local texture cues, motivating the integration of CNN-like local operators, frequency-domain processing, or explicit multi-scale feature fusion within SSM frameworks. However, a systematic understanding is still missing regarding which augmentation best complements SSMs under different tasks, modalities, and spatial resolutions.

6 Conclusion

Since the introduction of SSMs in remote sensing in 2024, a series of advanced research studies have emerged. This study extensively reviews the SSMs applications in remote sensing tasks, advancements in architecture design, existing challenges, and future opportunities. With the rapid development of technology today, we believe this study can help researchers quickly grasp the development context of SSMs in remote sensing and the key ideas of model improvement, and better integrate into the future direction in this field.

Acknowledgements The work was supported by National Natural Science Foundation of China (Grant Nos. 62125102, 62471014, U24B20177, 624B2017) and Fundamental Research Funds for the Central Universities.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1 Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. In: Proceedings of the 1st Conference on Language Modeling, 2024
2 Zhang H, Gong Y, Shen Y, et al. Poolingformer: long document modeling with pooling attention. In: Proceedings of International Conference on Machine Learning, 2021. 12437–12446
3 Dao T, Fu D, Ermon S, et al. Flashattention: fast and memory-efficient exact attention with IO-awareness. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 16344–16359
4 Wu C, Wu F, Qi T, et al. Fastformer: additive attention can be all you need. 2021. ArXiv:2108.09084
5 Beltagy I, Peters M E, Cohan A. Longformer: the long-document transformer. 2020. ArXiv:2004.05150
6 Ma W, Chen C, Ma M, et al. An adaptive dual-supervised cross-deep dependency network for pixel-wise classification. IEEE Trans Geosci Remote Sens, 2025, 63: 1–13
7 Chen K, Liu C, Chen H, et al. RSPrompter: learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Trans Geosci Remote Sens, 2024, 62: 1–17
8 Zhu W, Guo L, Zhang T, et al. Fastformer: transformer-based fast reasoning framework. In: Proceedings of the 14th International Conference on Graphics and Image Processing (ICGIP 2022), 2023. 12705: 378–387
9 Cheng B, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 1290–1299
10 Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: a survey. Proc IEEE, 2023, 111: 257–276
11 Cheng G, Han J. A survey on object detection in optical remote sensing images. ISPRS J Photogrammetry Remote Sens, 2016, 117: 11–28
12 Awais M, Naseer M, Khan S, et al. Foundation models defining a new era in vision: a survey and outlook. IEEE Trans Pattern Anal Mach Intell, 2025, 47: 2245–2264
13 Redmon J. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016
14 Redmon J, Farhadi A. Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 7263–7271

15 Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: Proceedings of the 14th European Conference on Computer Vision–ECCV 2016, 2016. 21–37
16 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, 2012. 25
17 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778
18 Sun X Q, Weng X X, Pang C, et al. Mitigating representation bias for class-incremental semantic segmentation of remote sensing images. Sci China Inf Sci, 2025, 68: 182301
19 Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. 2021. ArXiv:2111.00396
20 Bansal S, Madisetty S, Rehman M Z U, et al. A comprehensive survey of Mamba architectures for medical image analysis: classification, segmentation, restoration and beyond. 2024. ArXiv:2410.02362
21 Qu S, Tao X, Qu Z, et al. Almrr: anomaly localization Mamba on industrial textured surface with feature reconstruction and refinement. In: Proceedings of Chinese Conference on Pattern Recognition and Computer Vision (PRCV), 2024. 378–391
22 Tian W, Zeng H, Zhao Y P, et al. Empowering snapshot compressive imaging: Spatial-spectral state space model with across-scanning and local enhancement. 2024. ArXiv:2408.00629
23 Li H, Hu Q, Yao Y, et al. Cfmw: cross-modality fusion Mamba for multispectral object detection under adverse weather conditions. 2024. ArXiv:2404.16302
24 Díaz A H, Davidson R, Eckersley S, et al. E-mamba: using state-space-models for direct event processing in space situational awareness. In: Proceedings of IAA Conference on AI in and for Space (SPAICE), 2024. 509–514
25 Teng Y, Wu Y, Shi H, et al. Dim: diffusion Mamba for efficient high-resolution image synthesis. 2024. ArXiv:2405.14224
26 Li K, Li X, Wang Y, et al. Videomamba: state space model for efficient video understanding. In: Proceedings of European Conference on Computer Vision, 2024. 237–255
27 Zhang H, Chen K, Liu C, et al. CDMamba: incorporating local clues into Mamba for remote sensing image binary change detection. IEEE Trans Geosci Remote Sens, 2025, 63: 1–16
28 Zhu Q, Fang Y, Cai Y, et al. Rethinking scanning strategies with vision Mamba in semantic segmentation of remote sensing imagery: an experimental study. IEEE J Sel Top Appl Earth Observations Remote Sens, 2024, 17: 18223–18234
29 Rahman M M, Tutul A A, Nath A, et al. Mamba in vision: a comprehensive survey of techniques and applications. 2024. ArXiv:2410.03105
30 Sun X, Tian Y, Lu W X, et al. From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy. Sci China Inf Sci, 2023, 66: 140301
31 Lenz B, Lieber O, Arazi A, et al. Jamba: hybrid transformer-mamba language models. In: Proceedings of the 13th International Conference on Learning Representations, 2025. 67959-67984
32 Zhao H, Zhang M, Zhao W, et al. Cobra: extending Mamba to multi-modal large language model for efficient inference. AAAI, 2025, 39: 10421–10429
33 Zeng F, Liu R Y, Sun X Y, et al. Multi-static ISAC based on network-assisted full-duplex cell-free networks: performance analysis and duplex mode optimization. Sci China Inf Sci, 2025, 68: 150303
34 Liu N, Li W, Wang Y J, et al. A survey on hyperspectral image restoration: from the view of low-rank tensor approximation. Sci China Inf Sci, 2023, 66: 140302
35 Gu Y F, Liu T Z, Gao G M, et al. Multimodal hyperspectral remote sensing: an overview and perspective. Sci China Inf Sci, 2021, 64: 121301
36 Ma X, Zhang X, Pun M O, et al. A unified framework with multimodal fine-tuning for remote sensing semantic segmentation. IEEE Trans Geosci Remote Sens, 2025, 63: 1–15
37 Qu H, Ning L, An R, et al. A survey of Mamba. 2024. ArXiv:2408.01129
38 Xu R, Yang S, Wang Y, et al. A survey on vision Mamba: models, applications and challenges. 2024. ArXiv:2404.18861
39 Zhang H, Zhu Y, Wang D, et al. A survey on visual Mamba. Appl Sci, 2024, 14: 5683
40 Dosovitskiy A. An image is worth 16×16 words: transformers for image recognition at scale. 2020. ArXiv:2010.11929
41 Zhu L, Liao B, Zhang Q, et al. Vision Mamba: efficient visual representation learning with bidirectional state space model. 2024. ArXiv:2401.09417
42 He X, Cao K, Zhang J, et al. Pan-Mamba: effective pan-sharpening with state space model. Inf Fusion, 2025, 115: 102779
43 Zhou M, Li T, Qiao C, et al. DMM: disparity-guided multispectral Mamba for oriented object detection in remote sensing. IEEE Trans Geosci Remote Sens, 2025, 63: 1–13
44 Zhao S, Chen H, Zhang X, et al. RS-Mamba for large remote sensing image dense prediction. IEEE Trans Geosci Remote Sens, 2024, 62: 1–14
45 Chen H, Song J, Han C, et al. ChangeMamba: remote sensing change detection with spatiotemporal state space model. IEEE Trans Geosci Remote Sens, 2024, 62: 1–20
46 Paranjape J N, Melo C D, Patel V M. A Mamba-based siamese network for remote sensing change detection. In: Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2025. 1186–1196
47 Bao M, Lyu S, Xu Z, et al. Vision Mamba in remote sensing: a comprehensive survey of techniques, applications and outlook. 2025. ArXiv:2505.00630
48 Fairman F W. Linear Control Theory: the State Space Approach. Hoboken: John Wiley & Sons, 1998
49 Pao H T. Forecast of electricity consumption and economic growth in Taiwan by state space modeling. Energy, 2009, 34: 1779–1791
50 Johansson R, Robertsson A, Nilsson K, et al. State-space system identification of robot manipulator dynamics. Mechatronics, 2000, 10: 403–418
51 Gu A, Goel K, Gupta A, et al. On the parameterization and initialization of diagonal state space models. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 35971–35983
52 Gu A, Dao T, Ermon S, et al. Hippo: recurrent memory with optimal polynomial projections. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33: 1474–1487
53 Liu Y, Tian Y, Zhao Y, et al. VMamba: visual state space model. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 103031–103063
54 Chen K, Chen B, Liu C, et al. RSMamba: remote sensing image classification with state space model. IEEE Geosci Remote Sens Lett, 2024, 21: 1–5
55 Yang J X, Zhou J, Wang J, et al. Hsimamba: hyperpsectral imaging efficient feature learning with bidirectional state space for classification. 2024. ArXiv:2404.00272
56 Yao J, Hong D, Li C, et al. Spectralmamba: efficient Mamba for hyperspectral image classification. 2024. ArXiv:2404.08489
57 Wang G, Zhang X, Peng Z, et al. S^2Mamba: a spatial-spectral state space model for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2025, 63: 1–13
58 Huang L, Chen Y, He X. Spectral-spatial Mamba for hyperspectral image classification. 2024. ArXiv:2404.18401
59 Zhou W, Kamata S, Wang H, et al. Mamba-in-Mamba: centralized Mamba-cross-scan in tokenized Mamba model for hyperspectral image classification. Neurocomputing, 2025, 613: 128751
60 He Y, Tu B, Liu B, et al. 3DSS-Mamba: 3D-spectral-spatial Mamba for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2024, 62: 5534216
61 Shi X, Zhang Y, Liu K, et al. State space models meet transformers for hyperspectral image classification. Signal Processing, 2025, 226:

109669
62 Sheng J, Zhou J, Wang J, et al. DualMamba: a lightweight spectral-spatial Mamba-convolution network for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2025, 63: 1–15
63 Liao D, Wang Q, Lai T, et al. Joint classification of hyperspectral and LiDAR data based on Mamba. IEEE Trans Geosci Remote Sens, 2024, 62: 1–15
64 Yang A, Li M, Ding Y, et al. GraphMamba: an efficient graph structure learning vision Mamba for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2024, 62: 1–14
65 Ahmad M, Butt M H F, Usama M, et al. Multi-head spatial-spectral Mamba for hyperspectral image classification. 2024. ArXiv:2408.01224
66 Ahmad M, Usama M, Mazzara M, et al. WaveMamba: spatial-spectral wavelet Mamba for hyperspectral image classification. IEEE Geosci Remote Sens Lett, 2025, 22: 1–5
67 Ahmad M, Butt M H F, Khan A M, et al. Spatial-spectral morphological Mamba for hyperspectral image classification. Neurocomputing, 2025, 636: 129995
68 Gao F, Jin X, Zhou X, et al. MSFMamba: multiscale feature fusion state space model for multisource remote sensing image classification. IEEE Trans Geosci Remote Sens, 2025, 63: 1–16
69 Wang C, Huang J, Lv M, et al. A local enhanced Mamba network for hyperspectral image classification. Int J Appl Earth Observation GeoInf, 2024, 133: 104092
70 Chen T, Ye Z, Tan Z, et al. MiM-ISTD: Mamba-in-Mamba for efficient infrared small-target detection. IEEE Trans Geosci Remote Sens, 2024, 62: 1–13
71 Ren K, Wu X, Xu L, et al. Remotedet-mamba: a hybrid Mamba-CNN network for multi-modal object detection in remote sensing images. 2024. ArXiv:2410.13532
72 Liu P, Lei S, Li H C. Mamba-moc: a multicategory remote object counting via state space model. 2025. ArXiv:2501.06697
73 Shen D, Zhu X, Tian J, et al. HTD-Mamba: efficient hyperspectral target detection with pyramid state space model. IEEE Trans Geosci Remote Sens, 2025, 63: 1–15
74 Wang S, Wang C, Shi C, et al. Mask-guided Mamba fusion for drone-based visible-infrared vehicle detection. IEEE Trans Geosci Remote Sens, 2024, 62: 1–12
75 Ma X, Zhang X, Pun M O. RS3Mamba: visual state space model for remote sensing image semantic segmentation. IEEE Geosci Remote Sens Lett, 2024, 21: 1–5
76 Liu M, Dan J, Lu Z, et al. CM-Unet: hybrid CNN-Mamba UNet for remote sensing image semantic segmentation. 2024. ArXiv:2405.10530
77 Wang L, Li D, Dong S, et al. Pyramidmamba: rethinking pyramid feature fusion with selective space state model for semantic segmentation of remote sensing imagery. 2024. ArXiv:2406.10828
78 Sun J, Dai Y, Vong C M, et al. OE-BevSeg: an object informed and environment aware multimodal framework for bird's-eye-view vehicle semantic segmentation. IEEE Trans Intell Transp Syst, 2025, 26: 12210–12223
79 Ding H, Xia B, Liu W, et al. A novel Mamba architecture with a semantic transformer for efficient real-time remote sensing semantic segmentation. Remote Sens, 2024, 16: 2620
80 Zhu Q, Cai Y, Fang Y, et al. Samba: semantic segmentation of remotely sensed images with state space model. Heliyon, 2024, 10: e38495
81 Zhu E, Chen Z, Wang D, et al. UNetMamba: an efficient UNet-like Mamba for semantic segmentation of high-resolution remote sensing images. IEEE Geosci Remote Sens Lett, 2025, 22: 1–5
82 Li L, Chen B, Zou X, et al. Uv-mamba: a dcn-enhanced state space model for urban village boundary identification in high-resolution remote sensing images. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025. 1–5
83 Hu Y, Ma X, Sui J, et al. Ppmamba: a pyramid pooling local auxiliary SSM-based model for remote sensing image semantic segmentation. 2024. ArXiv:2409.06309
84 Liu C, Chen K, Chen B, et al. RSCaMa: remote sensing image change captioning with state space model. IEEE Geosci Remote Sens Lett, 2024, 21: 1–5
85 Xiao Y, Yuan Q, Jiang K, et al. Frequency-assisted Mamba for remote sensing image super-resolution. IEEE Trans Multimedia, 2025, 27: 1783–1796
86 Huang Y, Miyazaki T, Liu X, et al. Irsrmamba: infrared image super-resolution via Mamba-based wavelet transform feature modulation model. 2024. ArXiv:2405.09873
87 Wang Y, Yuan W, Xie F, et al. ESatSR: enhancing super-resolution for satellite remote sensing images with state space model and spatial context. Remote Sens, 2024, 16: 1956
88 Zhi R, Fan X, Shi J. MambaFormerSR: a lightweight model for remote-sensing image super-resolution. IEEE Geosci Remote Sens Lett, 2024, 21: 1–5
89 Liu Y, Xiao J, Guo Y, et al. Hsidmamba: exploring bidirectional state-space models for hyperspectral denoising. 2024. ArXiv:2404.09697
90 Fu G, Xiong F, Lu J, et al. SSUMamba: spatial-spectral selective state space model for hyperspectral image denoising. IEEE Trans Geosci Remote Sens, 2024, 62: 1–14
91 Zhou H, Wu X, Chen H, et al. RSDeHamba: lightweight vision Mamba for remote sensing satellite image dehazing. 2024. ArXiv:2405.10030
92 Fu H, Sun G, Li Y, et al. Hdmba: hyperspectral remote sensing imagery dehazing with state space model. 2024. ArXiv:2406.05700
93 Liu Z, Chen H, Bai L, et al. MambaDS: near-surface meteorological field downscaling with topography constrained selective state-space modeling. IEEE Trans Geosci Remote Sens, 2024, 62: 1–15
94 Debes C, Merentitis A, Heremans R, et al. Hyperspectral and LiDAR data fusion: outcome of the 2013 GRSS data fusion contest. IEEE J Sel Top Appl Earth Observations Remote Sens, 2014, 7: 2405–2418
95 Chen Y, Jiang H, Li C, et al. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Trans Geosci Remote Sens, 2016, 54: 6232–6251
96 Hang R, Liu Q, Hong D, et al. Cascaded recurrent neural networks for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2019, 57: 5384–5394
97 Hong D, Gao L, Yao J, et al. Graph convolutional networks for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2020, 59: 5966–5978
98 Vaswani A. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017
99 Hong D, Han Z, Yao J, et al. Spectralformer: rethinking hyperspectral image classification with transformers. IEEE Trans Geosci Remote Sensing, 2021, 60: 1–15
100 Rasti B, Hong D, Hang R, et al. Feature extraction for hyperspectral imagery: the evolution from shallow to deep: overview and toolbox. IEEE Geosci Remote Sens Mag, 2020, 8: 60–88
101 Gunther S, Ruthotto L, Schroder J B, et al. Layer-parallel training of deep residual neural networks. SIAM J Mathematics Data Sci, 2020, 2: 1–23
102 Weng W, Zhu X. INet: convolutional networks for biomedical image segmentation. IEEE Access, 2021, 9: 16591–16603
103 Wang J, Zheng Z, Ma A, et al. Loveda: a remote sensing land-cover dataset for domain adaptive semantic segmentation. 2021. ArXiv:2110.08733
104 Xu J, Xiong Z, Bhattacharyya S P. Pidnet: a real-time semantic segmentation network inspired by pid controllers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 19529–19539

105 Li R, Zheng S, Zhang C, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Trans Geosci Remote Sens, 2022, 60: 1–13
106 Chen J, Lu Y, Yu Q, et al. Transunet: transformers make strong encoders for medical image segmentation. 2021. ArXiv:2102.04306
107 Xu Z, Wu D, Yu C, et al. SCTNet: single-branch CNN with transformer semantic information for real-time segmentation. AAAI, 2024, 38: 6378–6386
108 Ji S, Wei S, Lu M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans Geosci Remote Sens, 2018, 57: 574–586
109 Chen H, Shi Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sens, 2020, 12: 1662
110 Daudt R C, Saux B L, Boulch A. Fully convolutional siamese networks for change detection. In: Proceedings of the 25th IEEE international conference on image processing (ICIP), 2018. 4063–4067
111 Zhang C, Yue P, Tapete D, et al. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS J Photogrammetry Remote Sens, 2020, 166: 183–200
112 Zhang C, Wang L, Cheng S, et al. SwinSUNet: pure transformer network for remote sensing image change detection. IEEE Trans Geosci Remote Sens, 2022, 60: 1–13
113 Bandara W G C, Patel V M. A transformer-based siamese network for change detection. In: Proceedings of IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, 2022. 207–210
114 Peng S, Zhu X, Deng H, et al. FusionMamba: efficient image fusion with state space model. 2024. ArXiv:2404.07932
115 Huang T, Pei X, You S, et al. Localmamba: visual state space model with windowed selective scan. In: Proceedings of European Conference on Computer Vision, 2025. 12–22
116 Chen H, Luo H, Wang C. AfaMamba: adaptive feature aggregation with visual state space model for remote sensing images semantic segmentation. IEEE J Sel Top Appl Earth Observations Remote Sens, 2025, 18: 8965–8983
117 Shao M, Tan X, Shang K, et al. A hybrid model of state-space model and attention for hyperspectral image denoising. IEEE J Sel Top Appl Earth Observations Remote Sens, 2025, 18: 9904–9918
118 Ma J, Li B, Li H, et al. Remote sensing change detection by pyramid sequential processing with Mamba. IEEE J Sel Top Appl Earth Observations Remote Sens, 2025, 18: 19481–19495
119 Ren Y, Li X, Guo M, et al. Mambacsr: dual-interleaved scanning for compressed image super-resolution with ssms. 2024. ArXiv:2408.11758
120 Gao R, Xiao Z, Xiong Z. Mamba-based light field super-resolution with efficient subspace scanning. In: Proceedings of the Asian Conference on Computer Vision, 2024. 531–547
121 He H, Bai Y, Zhang J, et al. Mambaad: exploring state space models for multi-class unsupervised anomaly detection. 2024. ArXiv:2404.06564
122 Liu C, Cao Y, Jiang D, et al. Efficient vision Mamba with dynamic feature fusion for remote sensing image semantic segmentation. In: Proceedings of IEEE International Conference on Signal, Information and Data Processing (ICSIDP), 2024. 1–6
123 Hu V T, Baumann S A, Gui M, et al. Zigma: a dit-style zigzag Mamba diffusion model. In: Proceedings of European Conference on Computer Vision, 2024. 149–166
124 Cao Y, Liu C, Wu Z, et al. Remote sensing image segmentation using vision Mamba and multi-scale multi-frequency feature fusion. Remote Sens, 2025, 17: 1390
125 Cao K, He X, Hu T, et al. Shuffle Mamba: state space models with random shuffle for multi-modal image fusion. 2024. ArXiv:2409.01728
126 Qiao J, Liao J, Li W, et al. Hi-mamba: hierarchical Mamba for efficient image super-resolution. 2024. ArXiv:2410.10140
127 Wang X, Huang Z, Zhang S, et al. Gmsr: gradient-guided Mamba for spectral reconstruction from RGB images. 2024. ArXiv:2405.07777
128 Jiang X, Han C, Mesgarani N. Dual-path Mamba: short and long-term bidirectional selective structured state space models for speech separation. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025. 1–5
129 Li K, Chen G, Yang R, et al. Spmamba: state-space model is all you need in speech separation. 2024. ArXiv:2404.02063
130 Verma T, Singh J, Bhartari Y, et al. Soar: advancements in small body object detection for aerial imagery using state space models and programmable gradients. 2024. ArXiv:2405.01699
131 Du R, Tang X, Ma J, et al. MLMamba: a Mamba-based efficient network for multi-label remote sensing scene classification. IEEE Trans Circuits Syst Video Technol, 2025, 35: 6245–6258
132 Pan Z, Li C, Plaza A, et al. Hyperspectral image classification with Mamba. IEEE Trans Geosci Remote Sens, 2025, 63: 1–14
133 Liu Q, Yue J, Fang Y, et al. HyperMamba: a spectral-spatial adaptive Mamba for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2024, 62: 1–14
134 Huang X, Zhang Y, Luo F, et al. Dynamic token augmentation Mamba for cross-scene classification of hyperspectral image. IEEE Trans Geosci Remote Sens, 2024, 62: 1–13
135 Yan L, Zhang X, Wang K, et al. Contour-enhanced visual state-space model for remote sensing image classification. IEEE Trans Geosci Remote Sens, 2025, 63: 1–14
136 Wang Y, Liang F, Wang S, et al. Towards an efficient remote sensing image compression network with visual state space model. Remote Sens, 2025, 17: 425
137 Fu X, Zhang T, Cheng J, et al. MMR-HAD: multiscale Mamba reconstruction network for hyperspectral anomaly detection. IEEE Trans Geosci Remote Sens, 2025, 63: 1–14
138 Shen Y, Xiao L, Chen J, et al. Learning cross-task features with Mamba for remote sensing image multitask prediction. IEEE Trans Geosci Remote Sens, 2025, 63: 1–16
139 Huang J, Yuan X, Lam C T, et al. LCCDMamba: visual state space model for land cover change detection of VHR remote sensing images. IEEE J Sel Top Appl Earth Observations Remote Sens, 2025, 18: 5765–5781
140 Zheng J, Fu Y, Chen X, et al. EGCM-UNet: edge guided hybrid CNN-Mamba UNet for farmland remote sensing image semantic segmentation. Geocarto Int, 2025, 40: 2440407
141 Cui X, Zhang L. Mamba-DCAU: state space dual attention center-sampling U-Net for hyperspectral image classification. Int J Remote Sens, 2025, 46: 5523–5547
142 Du S, Xing J, Wang S, et al. STMNet: scene classification-assisted and texture feature-enhanced multiscale network for large-scale urban informal settlement extraction from remote sensing images. IEEE J Sel Top Appl Earth Observations Remote Sens, 2024, 17: 13169–13187
143 Dong J, Yin H, Li H, et al. Dual hyperspectral Mamba for efficient spectral compressive imaging. 2024. ArXiv:2406.00449
144 Liu Y, Wang X, Shi L, et al. Ls-unet for semantic segmentation of fine-resolution remotely sensed images. In: Proceedings of China Automation Congress (CAC), 2024. 5373–5378
145 Li D, Zhao J, Chang C, et al. LGMamba: large-scale ALS point cloud semantic segmentation with local and global state-space model. IEEE Geosci Remote Sens Lett, 2025, 22: 1–5
146 Feng Y, Zhuo L, Zhang H, et al. Hybrid-MambaCD: hybrid Mamba-CNN network for remote sensing image change detection with region-channel attention mechanism and iterative global-local feature fusion. IEEE Trans Geosci Remote Sens, 2025, 63: 1–12
147 Ye F, Tan S, Huang W, et al. MambaTriNet: a Mamba-based tribackbone multimodal remote sensing image semantic segmentation model. IEEE Geosci Remote Sens Lett, 2025, 22: 1–5
148 Li L, Yi J, Fan H, et al. A lightweight semantic segmentation network based on self-attention mechanism and state space model for efficient urban scene segmentation. IEEE Trans Geosci Remote Sens, 2025, 63: 1–15
149 Wang X, Sun Q, Qi H, et al. Multiscale CNN-Mmaba hybrid network for building change detection in remote sensing images. In:

Proceedings of the 3rd International Conference on Remote Sensing, Mapping, and Geographic Information Systems (RSMG 2025), 2025. 13791: 247–252
150 Wang Q, Fan X, Huang J, et al. Spectral-spatial feature extraction network with SSM-CNN for hyperspectral-multispectral image collaborative classification. IEEE J Sel Top Appl Earth Observations Remote Sens, 2024, 17: 17555–17566
151 Sun H, Liu J, Yang J, et al. HMAFNet: hybrid Mamba-attention fusion network for remote sensing image semantic segmentation. IEEE Geosci Remote Sens Lett, 2025, 22: 1–5
152 Sun H, Yang X, Lu R, et al. Mscnet: Mamba-based self-correction remote sensing change detection network. Remote Sensing, 2026, doi: 10.2139/ssrn.5184974
153 He Y, Tu B, Liu B, et al. HSI-MFormer: integrating Mamba and transformer experts for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2025, 63: 1–16
154 Wang C, Tsepa O, Ma J, et al. Graph-mamba: towards long-range graph sequence modeling with selective state spaces. 2024. ArXiv:2402.00789
155 Li L, Wang H, Zhang W, et al. Stg-mamba: spatial-temporal graph learning via selective state space model. 2024. ArXiv:2403.12418
156 Liu C, Lin J, Wang J, et al. Mamba4rec: towards efficient sequential recommendation with selective state space models. 2024. ArXiv:2403.03900
157 Yang J, Li Y, Zhao J, et al. Uncovering selective state space model's capabilities in lifelong sequential recommendation. 2024. ArXiv:2403.16371
158 Luan X, Fan H, Wang Q, et al. FMambaIR: a hybrid state-space model and frequency domain for image restoration. IEEE Trans Geosci Remote Sens, 2025, 63: 1–14
159 Zhuang P, Zhang X, Wang H, et al. FAHM: frequency-aware hierarchical Mamba for hyperspectral image classification. IEEE J Sel Top Appl Earth Observations Remote Sens, 2025, 18: 6299–6313
160 Li D, Bhatti U A, Huang M, et al. Hyperfusionmamba: hybrid spectral-spatial fusion Mamba for efficient hyperspectral image classification with few-shot learning. IEEE Trans Consum Electron, 2025, doi: 10.1109/TCE.2025.3645064
161 Yue Z, Xu J, Yan Y, et al. TFFNet: transform fusion fuzzy network for multimodal remote sensing classification. IEEE Geosci Remote Sens Lett, 2025, 22: 1–5
162 Wu L, Peng J, Yang B, et al. Ssmif: enhanced spatial-spectral Mamba interactive fusion network for hyperspectral change detection. In: Proceedings of IEEE International Conference on Signal, Information and Data Processing (ICSIDP), 2024. 1–5
163 Ming R, Chen N, Peng J, et al. Semantic tokenization-based Mamba for hyperspectral image classification. IEEE J Sel Top Appl Earth Observations Remote Sens, 2025, 18: 4227–4241
164 Zhang G, Zhang Z, Deng J, et al. S^2CrossMamba: spatial-spectral cross-Mamba for multimodal remote sensing image classification. IEEE Geosci Remote Sens Lett, 2024, 21: 1–5
165 Li J, Liu Z, Liu S, et al. MBSSNet: a Mamba-based joint semantic segmentation network for optical and SAR images. IEEE Geosci Remote Sens Lett, 2025, 22: 1–5
166 Zhu C, Deng S, Song X, et al. Mamba collaborative implicit neural representation for hyperspectral and multispectral remote sensing image fusion. IEEE Trans Geosci Remote Sens, 2025, 63: 1–15
167 He Y, Tu B, Jiang P, et al. IGroupSS-Mamba: interval group spatial-spectral Mamba for hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2024, 62: 1–17
168 Wang R, Zhang J, Lu X, et al. JM-guided sentinel 1/2 fusion and lightweight APM-UNet for high-resolution soybean mapping. Remote Sens, 2025, 17: 3934
169 Li Z, Zhao L, Lu Y, et al. Mamba for remote sensing: architectures, hybrid paradigms, and future directions. Remote Sens, 2026, 18: 243
170 Chen B, Chen K, Yang M, et al. SeG-SR: integrating semantic knowledge into remote sensing image super-resolution via vision-language model. IEEE Trans Geosci Remote Sens, 2025, 63: 1–15
171 Ma H, Zhang C, Bian Y, et al. Fairness-guided few-shot prompting for large language models. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 36: 43136–43155
172 Jiang M, Bao K, Zhang J, et al. Item-side fairness of large language model-based recommendation system. In: Proceedings of the ACM on Web Conference 2024, 2024. 4717–4726
173 Cheng G, Huang Y, Li X, et al. Change detection methods for remote sensing in the last decade: a comprehensive review. Remote Sens, 2024, 16: 2355
174 Chen W, Gao F, Gan Y, et al. Wavelet-assisted Mamba for satellite-derived sea surface temperature super-resolution. IEEE Trans Geosci Remote Sens, 2025, 63: 1–12
175 Chen K, Liu C, Chen B, et al. Dynamicvis: an efficient and general visual foundation model for remote sensing image understanding. 2025. ArXiv:2503.16426
176 Ma X, Wu Q, Zhao X, et al. SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints. IEEE Trans Geosci Remote Sens, 2024, 62: 1–16
177 Kirillov A, Mintun E, Ravi N, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 4015–4026
178 Girdhar R, El-Nouby A, Liu Z, et al. Imagebind: one embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 15180–15190
179 Wang X, Wang W, Cao Y, et al. Images speak in images: a generalist painter for in-context visual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 6830–6839
180 Ma X, Xu X, Zhang X, et al. Adjacent-scale multimodal fusion networks for semantic segmentation of remote sensing data. IEEE J Sel Top Appl Earth Observations Remote Sens, 2024, 17: 20116–20128
181 Lüddecke T, Ecker A. Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 7086–7096
182 Chen X, Wang X, Changpinyo S, et al. Pali: a jointly-scaled multilingual language-image model. 2022. ArXiv:2209.06794
183 Yang J, Jin H, Tang R, et al. Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. ACM Trans Knowl Discov Data, 2024, 18: 1–32
184 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
185 Yang J, Liu T, Zhang L, et al. Effective road segmentation with selective state-space model and frequency feature compensation. IEEE Trans Geosci Remote Sens, 2025, 63: 1–13
186 Li Y, Ren J, Fu H, et al. GASSM: global attention and state space model based end-to-end hyperspectral change detection. J Franklin Institute, 2025, 362: 107424
187 Ma X, Zhang X, Ding X, et al. Decomposition-based unsupervised domain adaptation for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens, 2024, 62: 1–18
188 Chen K, Liu C, Chen B, et al. RSRefSeg 2: decoupling referring remote sensing image segmentation with foundation models. 2025. ArXiv:2507.06231
189 Chen B, Liu L, Zou Z, et al. Target detection in hyperspectral remote sensing image: current status and challenges. Remote Sens, 2023, 15: 3223
190 Chen B, Chen K, Yang M, et al. Heterogeneous mixture of experts for remote sensing image super-resolution. IEEE Geosci Remote Sens

Lett, 2025, 22: 1–5
191 Otter D W, Medina J R, Kalita J K. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learn Syst, 2020, 32: 604–624
192 Duc C M, Fukui H. Satmamba: development of foundation models for remote sensing imagery using state space models. 2025. ArXiv:2502.00435
193 Devalal S, Karthikeyan A. Lora technology—an overview, in 2018 Second International Conference on Electronics. In: Proceedings of Communication and Aerospace Technology (ICECA), 2018. 284–290
194 Shanmuga Sundaram J P, Du W, Zhao Z. A survey on LoRa networking: research problems, current solutions, and open issues. IEEE Commun Surv Tutorials, 2020, 22: 371–388
195 Chen Z, Duan Y, Wang W, et al. Vision transformer adapter for dense predictions. 2022. ArXiv:2205.08534
196 Xing Z, Ye T, Yang Y, et al. Segmamba: long-range sequential modeling Mamba for 3d medical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, 2024. 578–588
197 Wang C, Zheng W, Huang Y, et al. V2m: visual 2-dimensional Mamba for image representation learning. 2024. ArXiv:2410.10382
198 Wei R, Xu S, Zhong L, et al. Lightmamba: efficient Mamba acceleration on FPGA with quantization and hardware co-design. In: Proceedings of Design, Automation & Test in Europe Conference (DATE), 2025. 1–7
199 Mattern J, Hohr K. Mamba-chat. 2023. https://github.com/havenhq/mamba-chat
200 Hu E J, Shen Y, Wallis P, et al. Lora: low-rank adaptation of large language models. 2021. ArXiv:2106.09685
201 Wang X, Wang Z, Jiang J, et al. Treat stillness with movement: remote sensing change detection via coarse-grained temporal foregrounds mining. 2024. ArXiv:2408.08078
202 Chen Z, Shamsabadi E A, Jiang S, et al. Vision Mamba-based autonomous crack segmentation on concrete, asphalt, and masonry surfaces. 2024. ArXiv:2406.16518
203 Fu D Y, Epstein E L, Nguyen E, et al. Simple hardware-efficient long convolutions for sequence modeling. In: Proceedings of International Conference on Machine Learning, 2023. 10373–10391
204 Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models. 2023. ArXiv:2302.13971
205 Brown T B. Language models are few-shot learners. 2020. ArXiv:2005.14165
206 Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter efficient transfer learning for NLP. In: Proceedings of International Conference on Machine Learning, 2019. 2790–2799
207 Lin Z, Madotto A, Fung P. Exploring versatile generative language model via parameter-efficient transfer learning. 2020. ArXiv:2004.03829
208 Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 22199–22213
209 Xiao G, Lin J, Seznec M, et al. Smoothquant: accurate and efficient post-training quantization for large language models. In: Proceedings of International Conference on Machine Learning, 2023. 38087–38099
210 Zhang H, Guo H, Chen K, et al. FoBa: a foreground-background co-guided method and new benchmark for remote sensing semantic change detection. IEEE Trans Geosci Remote Sens, 2025, 63: 1–19
211 Ma C, Ji Y, Ye J, et al. Towards interpretable counterfactual generation via multimodal autoregression. 2025. ArXiv:2503.23149
212 Rao V M, Hla M, Moor M, et al. Multimodal generative AI for medical image interpretation. Nature, 2025, 639: 888–896
213 Hu Y, Wang L, Liu X, et al. Simulating the real world: a unified survey of multimodal generative models. 2025. ArXiv:2503.04641
214 Li J, Zhang C, Zhu W, et al. A comprehensive survey of image generation models based on deep learning. Ann Data Sci, 2025, 12: 141–170
215 Wei H, Sun Y, Li Y. DeepSeek-OCR: contexts optical compression. 2025. ArXiv:2510.18234
216 Duan Y, Che L, Cheng X, et al. EMambaEdgeNet: featuring feature decoupling and geometry-aware fusion with Mamba architecture. In: Proceedings of International Conference on Neural Information Processing, 2025. 517–531