Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 11(3): 5600–5647.
DOI: 10.3934/math.2026231
Received: 30 November 2025
Revised: 06 February 2026
Accepted: 14 February 2026
Published: 05 March 2026

Research article

Mamba-RSI: a state-space deep learning framework for efficient land-use and land-cover classification in remote sensing imagery

Wiem Abdelbaki[1], **Wided Bouchelligua**[2], **Inzamam Mashood Nasir**[3,*], **Sara Tehsin**[4] **and Hend Alshaya**[2]

[1] College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait

[2] Applied College, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

[3] Human-Environment-Technology (HET) Systems Centre, Mykolas Romeris University, Vilnius 08303, Lithuania

[4] Faculty of Informatics, Kaunas University of Technology, 51368 Kaunas, Lithuania

* **Correspondence:** Email: inzamam.nasir@mruni.eu.

Abstract: Accurate and efficient land-use and land-cover (LULC) classification from remote sensing imagery remains challenging. This is because it requires capturing long-range spatial dependencies while maintaining computational scalability. Recent transformer-based models improve global context modeling. However, they suffer from quadratic complexity and are limited in applicability to high-resolution imagery. We introduce Mamba-RSI: a linear-time, state-space deep learning framework using selective recursion, hierarchical multi-scale feature extraction, and lightweight global representations. Mamba-RSI captures both fine-grained spectral/texture information and coarse structural patterns with significantly less computational overhead than existing quadratic self-attention transformers. Extensive experimentation on EuroSAT and NWPU-RESISC45 demonstrated that Mamba-RSI achieves state-of-the-art performance. It achieved 99.72% accuracy on EuroSAT and 96.84% on RESISC45. This represents a +0.40% improvement over the strongest transformer baseline, ATMformer, on EuroSAT, a +0.29% improvement on RESISC45, and more than +0.53% over ViT-B on EuroSAT. Robustness tests under severe Gaussian noise ($\sigma = 0.10$) showed that Mamba-RSI maintains 97.43% accuracy. MaxViT, by comparison, maintains 94.01% in the same setting. Mamba-RSI also preserves 91.15% accuracy under 30% patch occlusion, outperforming ViT-B by +7.41%. Mamba-RSI provides an attractive blend of accuracy, robustness, and efficiency. It serves as a scalable foundation for new insights into remote sensing analytics and LULC mapping systems.

Keywords: remote sensing; land-use and land-cover classification; state-space models; Mamba

architecture; multi-scale feature extraction; efficient deep learning; scene classification
Mathematics Subject Classification: 68T05

1. Introduction

The classification of land-use and land-cover (LULC) is one of the primary tasks performed by Earth observation systems today, and is being used in many areas, including environmental monitoring, agricultural evaluation, natural hazard risk reduction, and urban area management [1]. The rapid proliferation of high-resolution satellite and aerial imagery has led to increased reliance on deep learning techniques that can use complex spatial/spectral data to provide unique and meaningful representations of that data [2]. Traditional convolutional neural networks (CNNs) such as AlexNet, VGG, and ResNet have all been very successful in the remote sensing of scenes; however, the fact that these architectures have local receptive fields means that they cannot capture long-range relationships that often exist in very large area landscape configurations [3]. As resolution increases and scenes become increasingly intricate, with large geographic areas being captured [4], the issue will become more pronounced, further hindering performance.

Vision transformers (ViTs) use self-attention mechanisms to effectively model global context, addressing the locality problem that CNNs experience [5]. Different transformer variants have been demonstrated in real-world applications such as scene recognition, object detection, and spectral/spatial analysis [6]. The increased size of the input is a major impediment for transformers, as their attention mechanisms incur quadratic computational costs when applied to high-resolution satellite imagery [7]. Recent attempts to use hierarchical and windowed attention to resolve this issue have also led to increased memory requirements and a broader range of design problems. Moreover, most transformer models are found to be more sensitive than CNNs to several sources of noise; this includes both reconstruction artifacts, local perturbations, and all sources of image degradation that occur continually during the acquisition of satellite images in typical remote sensing applications [8].

More state-of-the-art state-space models provide promising directions in sequence modeling by substituting attention with linear-time recurrent operators [9]. The Mamba framework utilities selective state-space models to perform content-aware gating and produce much more flexible representations of long-range relational dependencies, while using much less computation than a fully-fledged, fully connected Mamba architecture [10]. State-space models outperform traditional transformer architectures in both scalability and stability for long-sequence tasks, thanks to their continuous-time formulation and controlled state propagation mechanisms [11]. Because these features align better with remote sensing applications, the initial explorations into applying these types of models in the vision domain have produced promising outcomes, indicating that using linear-time recurrence to replace global self-attention may enable the same or better performance than that of transformers with a dramatically reduced computational and memory footprint [12].

In addition, the development of multi-scale representation learning techniques has become increasingly important for classifying remotely sensed imagery, as relevant features can occur at different spatial scales [13]. Hierarchical approaches to feature extraction achieve significant performance improvements by leveraging multiple receptive field sizes to capture spatial patterning [14]. Recent transformer and hybrid models have also begun to leverage multi-scale processing to improve texture and structure modeling [15]. However, quadratic growth in complexity with respect to self-attention or computationally intensive multi-branch architectures is common across all of these approaches [16]. As a result, further investigation is warranted into the feasibility of developing a single unified architecture that combines efficient long-range modeling with a rich multi-scale spatial representation. Inspired by well-established concepts from previous work, such as CNN feature pyramids and hierarchical transformers, the hierarchical multi-scale design adopted in this paper is based on hierarchical representation learning in linear-time state-space. Essentially, hierarchical representation learning within the linear-time state-space framework is the major contribution of Mamba-RSI, not just the new multi-scale design. In particular, multi-scale representation learning is achieved by reducing token resolution while expanding the effective receptive field through selective state-space recurrence, rather than using a convolutional kernel or a self-attention mechanism.

Mamba-RSI is a linear state-space method for fast and robust remote sensing (RS) scene classification that addresses the above limitations. Selective recurrence (SR), rather than self-attention (SA), models global dependencies while keeping computational complexity low. Mamba-RSI has a multi-scale hierarchical processing architecture and uses global representation aggregation. Thus, Mamba-RSI captures fine detail in the form of textures while efficiently capturing an overview of scene layouts without processing large amounts of high-dimensional data, as seen with transformer model architectures. Most importantly, Mamba-RSI demonstrates improved robustness against typical imaging and remote sensing distortions such as noise, blur, occlusion, and spectral jittering [17]. Furthermore, Mamba-RSI uses a combination of sequential model training and adaptive gating to improve cross-domain generalization. This addresses the long-standing problem in remote sensing where many models struggle to understand scenes across different geographic regions or sensor modalities [18].

The work presented in this paper defines a valid approach to conducting experiments using the EuroSAT and NWPU-RESISC45 datasets to assess the effectiveness of this methodology relative to existing methodologies. Our results illustrate improvements over both classified CNNs and transformer architectures, as well as the hybrid model, across classification accuracy, robustness, computational efficiency, and cross-dataset generalization. Our findings support the view that Mamba-RSI offers an alternative to existing transformer-based approaches, providing many benefits for large-scale, operational remote sensing applications. This paper makes three primary contributions to the field. First, it proposes a new integrated state-space framework that integrates selective recurrence, multi-scale spatial representation, and a global aggregation of features for land-use and cover types. Second, the paper demonstrates strong performance, robustness and computational efficiency of this framework on several standard remote sensing benchmarks. Third, it presents numerous case studies that demonstrate that state-space models can be effectively used as a viable and scalable alternative to attention-based architectures when applied in remote sensing scenarios.

This work aims to fully and rigorously validate existing state-space Mamba models in remote sensing. The foremost contribution will be a demonstration of how to structure and adapt the Mamba framework to perform land-use/land-cover classification using hierarchical multi-scale processing, knowledge- and processing-power-based gating, and independent testing for thorough robustness evaluation. The developed framework leverages established components of deep learning, including a selective state-space model, a hierarchical multi-scale feature-extraction architecture, a global feature-aggregation technique, and a lightweight linear classification head. Each of these components has been extensively analyses and is widely used throughout the literature. While this work is not novel because it introduces new architectural primitives, it does represent an innovative contribution through the systematic integration of these components within the Mamba state-space framework, as demonstrated through extensive empirical evaluation, providing a highly effective combination for remote sensing image classification tasks.

The remainder of the paper is structured as follows: Section 2 includes references to all previous related work, Section 3 gives a detailed description of our methodology, Section 4 provides an overview of our experimental results, and Section 5 describes new directions for future research.

2. Related work

Remote sensing and deep learning continue to play a crucial role in LULC classification. Deep learning has enabled rapid advances in understanding how to produce students at scale from a single dataset [19–21]. Although early-stage convolutional neural networks (CNNs) such as VGG, ResNet, and EfficientNet served as strong baselines, their limited receptive field restricted their ability to capture global contextual dependencies. As a result, progress in the field involved transferring knowledge from more advanced models into EuroSAT EfficientNet-B0 [22], leading to significant performance improvements and competitive accuracy on EuroSAT. With the adoption of more advanced CNN pipelines—such as an advanced self-supervised learning CNN using ResNet101 with cross-channel feature mixing (CCFM) [23] on EuroSAT—performance has increased further, surpassing 99% accuracy and demonstrating the benefits of developing more advanced representation learning models.

The introduction of vision transformers (ViTs) marked a pivotal transition from CNN-based methods to transformer-based approaches in remote sensing image classification. ViT-based models fine-tuned on EuroSAT [24] consistently exceeded 99% overall accuracy by leveraging global self-attention, highlighting the shift from local to global feature extraction. Multi-scale variants such as MaxViT [25] build on this shift by integrating local and global, grid-based attention, enabling more effective spatial contextualization. On NWPU-RESISC45, a similar transition occurred as methods evolved from CNNs to hybrid and transformer-based models. For instance, teacher–student distillation frameworks [26] enabled lightweight students such as EfficientNet-B0 to exceed 94% accuracy when trained under powerful ConvNeXt- and EfficientNet-based teacher ensembles. Hybrid architectures, such as the lightweight dual-branch Swin transformer [27], further improved performance by combining convolutional features with hierarchical, windowed attention. Other approaches, including multi-network deep feature fusion via deep canonical correlation analysis (DCCA) [28] and multi-path reconfigurable CNNs [29], have demonstrated competitive accuracy on NWPU-RESISC45 by exploiting multi-scale redundancy. More recently, adaptive token-merging transformers [30] have been proposed to reduce computational burden while preserving high classification performance.

Recently, numerous linear-complexity models with the potential to increase scalability on long token sequences have been developed, such as hyena [31] and rwkv [32] models. The original emphasis of this submission was to evaluate the application of Mamba-style selective state-space recurrence to remotely sensed imagery classification under a controlled environment. We will be expanding the experimental benchmarking to include new linear-attention/linear-complexity models that were developed specifically for remotely sensed imagery, thus providing a more comprehensive comparison of selective state-space modeling and more efficient attention models for use with high-resolution remotely sensed imagery. Table 1 summarizes these methods, capturing the transition from CNNs to transformers and hybrids, and serves as a baseline for the proposed methodology.

Table 1. Baseline deep learning models used for comparison on EuroSAT and NWPU-RESISC45.

Model / Method	Year	Type	Dataset	Accuracy
MaxViT for LULC [25]	2024	Transformer	EuroSAT	~99.0%
Vision Transformer (ViT-B) [24]	2024	Transformer	EuroSAT	99.19%
EfficientNet-B0 TL [22]	2025	CNN	EuroSAT	98.10%
ResNet101 + SSL + CCFM [23]	2025	CNN + SSL	EuroSAT	99.66%
ResNet-50 TL Variant [33]	2024	CNN	EuroSAT	95.90%
Teacher–Student Distillation [26]	2023	CNN + Distillation	NWPU-RESISC45	96.20%
LDBST Lightweight Swin Transformer [27]	2023	Hybrid	NWPU-RESISC45	~93.94%
Deep Feature Fusion (DCCA) [28]	2024	CNN Fusion	NWPU-RESISC45	87.53%
ATMformer (Adaptive Token Merging) [30]	2025	Transformer	NWPU-RESISC45	94.41 %
MPRNN [29]	2024	CNN	NWPU-RESISC45	91.86%
RSRWKV [31]	2025	Linear-Complexity Attention	NWPU-RESISC45	94.83%
ALHCT [32]	2025	Linear Attention	NWPU-RESISC45	96.19%

Despite the considerable advances made by CNNs and hybrid and transformer-based architectures in the classification of remote sensing imagery, major research gaps remain. First, most transformer architectures rely on global self-attention, which scales quadratically with the number of tokens processed, making them disastrous for processing high-resolution aerial and satellite images. This limits their application in scenarios that require fine-grain patch decomposition and/or large-area inference, which are the norm in real-world LULC applications. Second, existing convolutional-based models have failed to accurately model long-range spatial dependencies due to their localized receptive fields, leading to large models that are more prone to overfitting. Third, while some current hybrid architectures have attempted to address these limitations through hierarchical/pyramidal architectures, they continue to rely on large attention maps and/or multi-branch CNNs, both of which introduce redundancy, memory overhead, and inefficiencies during training. Fourth, previous work on content-aware filtering is limited, leaving models vulnerable to noise, homogeneous regions, and irrelevant spatial patterns that dominate remote sensing images. Lastly, there has been a lack of a unified framework offering efficient sequence modeling, multi-scale spatial reasoning, and adaptive flow control for the development of a single LULC model.

The research gaps identified above motivate the design of the proposed Mamba-based state-space framework to address the shortcomings of existing approaches, with a focus on the efficient modeling of LULC via a linear-time complexity architecture, using a selective gating mechanism and a hierarchical multi-scale structure. Unlike transformers, the state-space encoder processes input tokens according to a mathematically grounded recurrence relation, processing them in linear time ($O(N)$). The state-space encoder, therefore, enables efficient modeling of long-range dependencies without the quadratic computational overhead associated with current transformer architectures. The selective gating mechanism also enables adaptive modulation of tokens, with their contributions to the final output influenced by their semantic relevance to the task.

As a result, the critical spatial-spectral patterns, such as boundaries, edges, and heterogeneous textures, are effectively propagated through the latent dynamics with minimal disruption from irrelevant and/or redundant inputs. Finally, the hierarchical multi-scale extraction approach enables the simultaneous capture of spatial variation across multiple image resolutions, allowing the network to learn both fine-grained context and global context within the same architecture. By combining the three components mentioned above, the proposed methodology creates a unified architecture that not only overcomes the computational and locality limitations of current transformer and convolutional models but also improves the robustness, generalization, and discriminative ability of LULC classification across both the EuroSAT and RESISC45 datasets.

3. Proposed methodology

The proposed framework introduces a state-space deep learning architecture tailored for efficient and accurate land-use and land-cover (LULC) classification using high-resolution remote sensing imagery. The model integrates efficient sequence modeling, hierarchical feature extraction, and global spatial reasoning through a Mamba-based state-space backbone. This section presents the architectural components, the feature transformation pipeline, the training objectives, and the optimization strategy. The complete end-to-end processing pipeline of the Mamba-RSI framework is summarized in Figure 1.

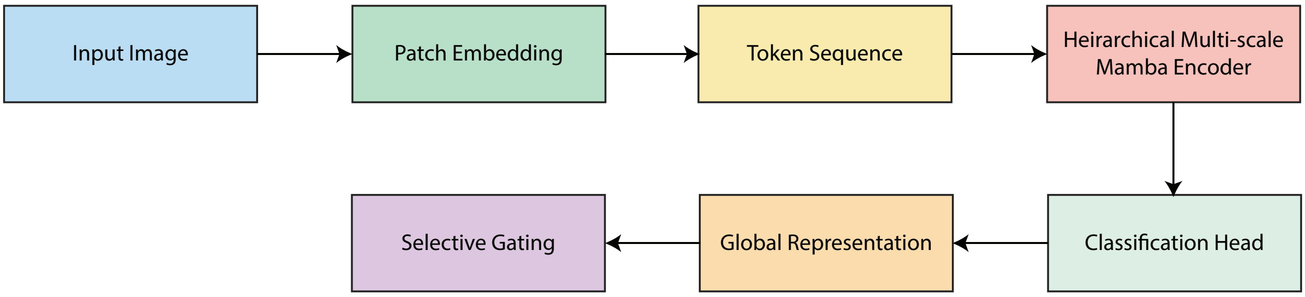

Figure 1. Overall architecture of the Mamba-RSI framework for land-use and land-cover classification.

The proposed framework introduces a unified state-space deep learning architecture designed to efficiently model spatial structures, spectral variations, and global contextual relationships present in high-resolution remote sensing imagery. Let the input image be denoted as $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, where H and W represent spatial dimensions, and C denotes the number of spectral channels (e.g., $C = 3$ for RGB or $C = 13$ for multi-spectral Sentinel-2). The architecture first transforms $\mathbf{X}$ into a structured sequence of patch-level tokens, which enables linear-time state-space modeling over long spatial trajectories. This transformation is essential because raw images exhibit high spatial redundancy and non-uniform feature distributions, which make direct modeling computationally expensive. To address this, the image is partitioned into non-overlapping $P \times P$ patches, yielding $N = \frac{HW}{P^2}$ patches. Each patch is then projected into a d-dimensional feature embedding, forming the token sequence $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]$, where d denotes the latent feature dimension used throughout the model. A selective state-space encoder based on Mamba uses tokens as input to encode information and relationships, with linear code complexity. The Mamba encoder addresses global inter-dependencies between tokens without the computationally expensive methods used by traditional self-attention mechanisms, which involve an order of N^2 computations. The selective state-space encoder uses dynamic gating mechanisms and structured recurrence to encode long-distance relationships without consuming excessive RAM. After the tokens are tokenized, the sequence $\mathbf{Z}$ is sent through the Mamba encoder, which encodes it using selective recurrence on the spatial tokens via a discretion state-space operator. In other words, continuous-state evolution, defined by a hidden state $\mathbf{h}(t) \in \mathbb{R}$ and an input signal $\mathbf{u}(t) \in \mathbb{R}^d$, is governed by

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{u}(t). \tag{3.1}$$

The transition matrix $\mathbf{A}$ is learnable and belongs to the space of $d \times d$ matrices with real entries. Matrix $\mathbf{B}$ of the same dimensions, $d{\times}d$, injects external input into the dynamics of the process described above. The observable output from this system is given by

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t). \tag{3.2}$$

The mapping function $\mathbf{C} \in \mathbb{R}^{d \times d}$ maps hidden state dynamics into the output space. Using the continuous form of this mapping within a deep learning framework means discretizing our continuous dynamics using token-position steps, with an appropriately learned token-step size Δ. Thus, the dynamic mapping is given through the recurrence relation:

$$\mathbf{h}_i = \bar{\mathbf{A}}\mathbf{h}_{i-1} + \bar{\mathbf{B}}\mathbf{z}_i. \tag{3.3}$$

In this example, $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ refer to the discrete representation of $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$. Using a recurrent approach for encoding information over long distances gives the encoder an overall perspective for distinguishing between land units that may look similar at the physical level (e.g., they have similar textures), but have very different formats or arrangements (e.g., the locations where they occur). In addition, Mamba's selective gating system limits the number of pieces of information that affect the final output, depending on their relationships to other pieces of information at relevant physical locations. Once the encoder's selective state-space model produces $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]$, it aggregates these outputs into a unified global feature. It is common to combine the outputs of an encoder's sequence into a single compact vector by either using a trained classification token or applying an operation that maintains permutation invariance, such as global average pooling. Hence, to construct the final model vector $\mathbf{v} \in \mathbb{R}^d$, it combines both local semantic and global structural relationships. To generate the final land use/land cover category label prediction, the classification head will utilize a fully connected (fc) layer and a softmax activation function to produce a probability for each of the K target labels. The completed predicted probabilities for the final labels are found in the following equation:

$$\hat{\mathbf{y}} = \mathrm{softmax}(\mathbf{W}_c\mathbf{v} + \mathbf{b}_c). \tag{3.4}$$

In this system, the learnable parameters used in classifying the tokens are represented by $\mathbf{W}_c \in \mathbb{R}^{K \times d}$ and $\mathbf{b}_c \in \mathbb{R}^K$. The approach we have taken is to optimize the end-to-end pipeline with respect to the cross-entropy loss in a single optimization process so that the global state-space dynamics and token-level representations can learn and adapt together to account for the structural properties of remote sensing imagery. To summarize this subsection, we outline the architectural flow of the system: patch extraction → token embedding → selective state-space encoding → global feature aggregation → classification. Each subsequent subsection discusses in more detail the various components of this architecture, including the mathematical derivations and design motivations for each component.

3.1. Patch embedding and tokenization

The raw remote sensing image undergoes a transformation in the first stage of this framework into a structured series of patch-level embeddings, which are then used as inputs to the state-space encoder. We let the input image be referred to by the symbol $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, where H denotes the height of the image, W denotes the width of the image, and C is the number of spectral channels. Because of the extreme spatial redundancy found in remote sensing images, the heterogeneous distribution of textures across an image, and their multi-scale nature, it is not feasible for deep models operating at large global spatial scales to model $mathbfX$ directly as a 2D grid. Thus, we partition each remote sensing image into $P \times P$ patches that do not overlap with one another, yielding $N = \frac{HW}{P^2}$ patches, each of which presents localized spatial-spectral features associated with land-use and land-cover recognition. Each individual patch is denoted as $\mathbf{X}_i \in \mathbb{R}^{P\times P\times C}$. The pixels contained within each patch are then rearranged into a vectorized representation using the operator vec$(\cdot)$, producing a column vector of dimension $P^2 * C$. The pixels contained within each patch are rearranged into a vectorized representation using the operator vec$(\cdot)$, producing a column vector in $\mathbb{R}^{P^2C}$. This vectorized representation is then projected into a fixed-sized latent space via a linear embedding layer with parameters $W_p \in \mathbb{R}^{d\times(P^2C)}$ and $b_p \in \mathbb{R}^d$. The embedded token $z_i \in \mathbb{R}^d$ associated with the i-th patch is computed as

$$\mathbf{z}_i = \mathbf{W}_p \operatorname{vec}(\mathbf{X}_i) + \mathbf{b}_p. \tag{3.5}$$

Since d represents the total number of dimensions of the feature space shared across all of the regions of the state-space encoder; it will allow us to reduce each area of spatial information into a single, smaller representation and allow us to create a token sequence with a constant overall dimensionality, thereby making it possible to perform state-space encoded processing on that tokenization sequence at select areas within that token sequence. The full tokenization sequence itself is thus:

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] \in \mathbb{R}^{N\times d}. \tag{3.6}$$

The token sequence follows a raster scan of the entire image, preserving the minimum distance between any two adjacent patches. The following rule is very important to the Mamba: using token sequences to process the image patch. As the Mamba scans the image, it uses a sequence of tokens to model a continuous trajectory for each patch across the length of an extended set of images. Given the location of these images, it can use patch position encoding to determine the location of a given patch within its respective image. Classification of remote sensing images based on location has an important impact on classification, because agricultural fields are generally located within large, contiguous areas, while urban structures appear in more irregular patterns. It is therefore necessary that the Mamba include position encoding when processing this information. Each token processed by the Mamba will have a corresponding position encoding vector $\mathbf{p}_i \in \mathbb{R}^d$ associated with it as described herein:

$$\tilde{\mathbf{z}}_i = \mathbf{z}_i + \mathbf{p}_i. \tag{3.7}$$

This paper presents a new state-space encoder based on a tokenization mechanism that creates a hierarchical representation, enabling the encoder to learn and represent morphological patterns at multiple scales through the hierarchical arrangement of tokens. $\tilde{\mathbf{z}}_i$ are position-aware tokens (*mathttb*), and this means that when the encoder processes the training set, the tokens will preserve the structure of the input data (spatial pattern for all pixels). Because each token represents the unique characteristics of the input data, it becomes easier to establish relationships among the different types of spatial data. Also, tokens enable efficient modeling of long-range dependencies between pixel regions through spatial encoding. Rather than performing either convolution or self-attention on the global image, the encoder can propagate this information through the token sequences **Z**. Tokenization of input data is beneficial for remote sensing imagery because broad land-use patterns can span multiple contiguous land-use types. Since the model generates a sequence of semantically meaningful tokens that can be assembled into a single space-time representation via G-W interactions, this constitutes the final step in encoding image data into spatial tokens. Figure 2 provides an illustration of the patch and token extraction and tokenization processes.

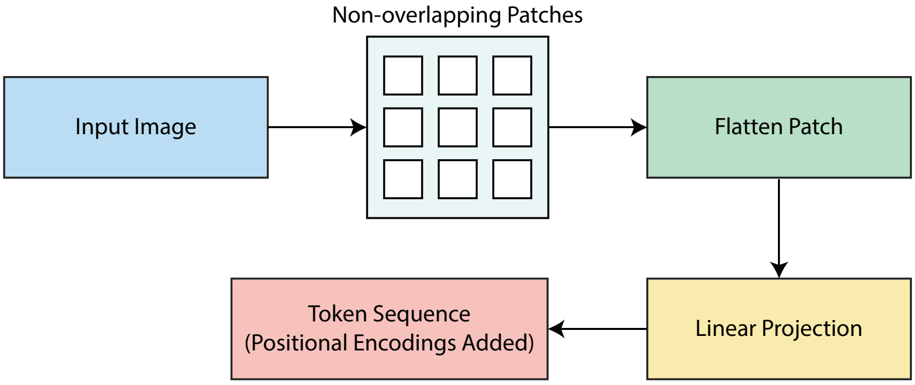

Figure 2. Patch embedding and tokenization process in Mamba-RSI.

3.2. State-space Mamba encoder

The Mamba encoder is a component that enables effective modeling of spatial relationships over long distances between individual pixels in a tokenization satellite and other source images. In contrast to transformer-type self-attention mechanisms, where computational costs grow exponentially with the number of tokens, the Mamba encoder updates its internal state using a linear recurrent method, meaning that both execution time and memory requirements scale linearly with the sequence length. A significant aspect of the Mamba encoder's design is the use of selective gating to adaptively control the influence each token has on the overall input to the encoder during processing within the state-space transformation. By allowing the model to focus on semantically relevant spatial and spectral patterns while suppressing redundancy or homogeneity in remote sensing imagery, the encoder effectively transmits essential structural data over long distances without incurring the high cost of pairwise attention computations.

In terms of the series of tokens, let $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]$ be the token sequence, with the form $\mathbf{z}_i \in \mathbb{R}^d$. The embeddings (hidden state vectors) from the encoder to the token sequence will produce a corresponding series of hidden state representations $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N]$, where at each index i, the hidden state is a representation in $\mathbb{R}^d$ of both local semantics as well as aggregate, long-range structure information. The foundation of Mamba is based on a continuous-time state space model (SSM), in which the evolution of the hidden state $h(t) \in \mathbb{R}^d$ is governed by an input signal $u(t) \in \mathbb{R}^d$ through a linear differential equation:

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{u}(t), \tag{3.8}$$

where $A \in \mathbb{R}^{d\times d}$ is the state transition matrix and $B \in \mathbb{R}^{d\times d}$ determines how the input influences the hidden state. A linear readout maps the hidden state to the output:

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t), \tag{3.9}$$

where $C \in \mathbb{R}^{d\times d}$ maps internal state dynamics into the output space. This approach can model an extensive array of linear dynamical systems and is useful for modeling spatial movements in remote sensing images. The hidden-state update for the digitized form is given by: To apply the continuous-time formulation to a discrete token sequence, we discretion the dynamics using a learnable step size Δ. Denoting the discretion matrices by $\bar{A}$ and $\bar{B}$, the state update over tokens is written as

$$\mathbf{h}_i = \bar{\mathbf{A}}\mathbf{h}_{i-1} + \bar{\mathbf{B}}\mathbf{z}_i. \tag{3.10}$$

For notation, $\bar{A} = e^{\Delta A}$ and $\bar{B}$ is obtained from the corresponding matrix integral over $[0, \Delta]$. This recurrent method allows the encoder to gather contextual information from all previous assignments when calculating $\mathbf{h}_i$, enabling it to sustain long-distance dependencies across tokens via sequential backpropagation. One of the innovations introduced by the Mamba design was selective gating, which regulates each token's contribution to the hidden state. The gated input is defined by taking the vector $\mathbf{g}_i \in \mathbb{R}^d$, which is computed from the input with a fully parameterized transformation, and applying it to the token being propagated.

$$\tilde{\mathbf{z}}_i = \mathbf{g}_i \odot \mathbf{z}_i. \tag{3.11}$$

The symbol $\odot$ is used to indicate multiplication of two elements, element by element. The introduction of this mechanism gives preference to tokens with strong spatial-spectral values. This can be especially useful in remote sensing images, as significant features within an image that could include fields, major freeways, buildings, or other structures are found in many parts of the image and are located far apart. Using this gated input, each time a unit within (a sequence of) images receives a state update, it will receive an additional state update for those units associated with the previous image, which is an improvement in terms of recurrent state updates. The selective state update is given by

$$\mathbf{h}_i = \bar{\mathbf{A}}\mathbf{h}_{i-1} + \bar{\mathbf{B}}\tilde{\mathbf{z}}_i. \tag{3.12}$$

The way the current input affects the model depends on $\mathbf{g}_i$. Therefore, if we look at only one image at a time, we get a modified version tailored to its spatial features; this allows the model to handle both fine-grained details and broader context. The encoder representation across all token positions can be expressed as follows:

$$\mathbf{H} = \text{Mamba}(\mathbf{Z}). \tag{3.13}$$

Representation $\mathbf{H} \in \mathbb{R}^{N\times d}$ captures the long-term relationship. In this way, the Mamba encoder provides a formal, mathematically sound, and computationally efficient framework for modeling spatial relationships in large-area remote sensing images, serving as the basis for building hierarchical representations from remote sensing data. We also illustrate, as Figure 3, how each of the components of the Mamba state-space, namely gates, state transitions, and token output, is defined.

Theoretical motivation and an example of the state-space dynamics are provided using continuous-time state-space equations. In practice, the proposed framework follows the standard Mamba parameterization and performs state updates discretely, as is routinely done by the Mamba architectures already developed. There are no new continuous-time solvers, nor are there alternative discretization schemes presented in this work. All experiments are performed using discrete-time recurrence, with parameters learned end-to-end via backpropagation.

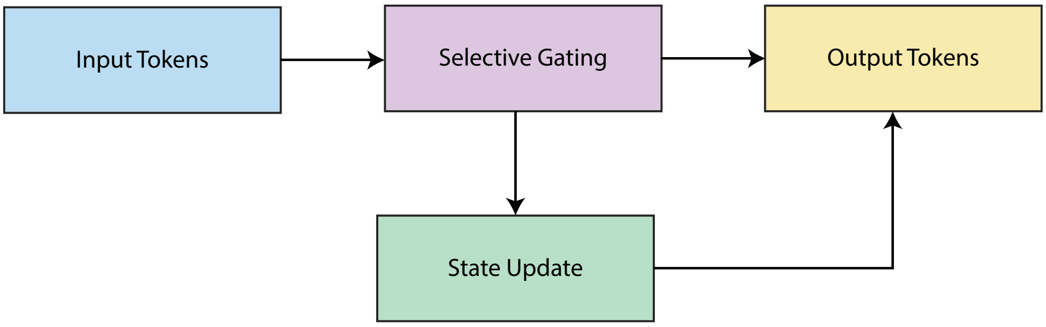

Figure 3. Schematic of a single Mamba state-space encoder block.

3.3. Hierarchical multi-scale feature extraction

The hierarchical models proposed employ state-space recurrence over successively down-sampled token sequences to capture global context. Instead of stacking convolutional layers, as in pyramidal CNNs that grow receptive fields through incremental convolutions, or using windowed or shifted self-attention, as in hierarchical transformer architectures, the proposed hierarchy uses state-space recurrence to access the entire token sequence at once. As the token resolution decreases at deeper levels of the hierarchy, every successive sampling from the system gathers information from larger and larger spatial regions. Thus, while each of these recurrent updates builds a multi-scale model of the environment, the overall computational cost remains linear in relation to the number of tokens. It is emphasized that the hierarchical structure itself follows well-established multi-scale representation principles; the contribution lies in demonstrating how such a hierarchy can be effectively combined with selective state-space modeling for remote sensing imagery.

Remote sensing imagery contains information at multiple scales. For example, individual buildings and narrow road segments represent extremely small scales, whereas crops, forests, and water can be considered very large homogeneous areas. The challenge in capturing all the spatial variability involved is combining information recorded at different resolutions without sacrificing performance or computational efficiency. To meet this challenge, we provide a structured approach to multi-scale feature extraction via a hierarchical framework consisting of stacked Mamba-based state-space layers, each representing a level in the hierarchy and converting the hierarchy's token sequence to different resolutions. Therefore, the model can also extract not only local discriminative patterns but also global structures. The first state-space encoder produces output $\mathbf{H}^{(1)}\mathbb{R}$ with dimensions $N \times d$, where N is all of the tokens calculated from the $P \times P$ patch grid and this output acts as a building block for the creation of subsequent levels of abstraction within the framework. For every numerical level of the hierarchy, the token sequence is recursively down-sampled using a learned projection operator. The token sequence $\mathbf{H}^{(l)}$ for down-sampled level l has dimensions $N_l \times d_l$, and it is constituted by applying a linear transformation and then performing spatial pooling. The down-sampled sequence is calculated as follows:

$$\mathbf{H}^{(l+1)} = \mathrm{Pool}(\mathbf{W}_l \mathbf{H}^{(l)}), \tag{3.14}$$

where $mathbfW_l \in \mathbb{R}^{d_{l+1}\times d_l}$ is a trainable transformation matrix that projects the incoming feature set into an embedded low (2D) dimensionality, and Pool$(\cdot)$ is the process to reduce the size of the input feature set from N_l to N_{l+1} through average pooling, stride selection, and/or attention-guided sampling. This ensures that deeper levels of the hierarchy process increasingly coarse spatial summaries, analogous to multi-resolution processing in classical image pyramids. Importantly, each down-sampled sequence remains compatible with the state-space encoder, enabling the architecture to repeatedly refine global spatial relationships. Once down-sampled, each sequence $\mathbf{H}^{(l+1)}$ is passed through another Mamba encoder block, which refines the representation by incorporating long-range dependencies at the new resolution. Formally, the state propagation at level $l + 1$ is expressed as

$$\mathbf{H}^{(l+1)} = \mathrm{Mamba}_{l+1}\Big(\mathbf{H}^{(l+1)}\Big), \tag{3.15}$$

where $\mathrm{Mamba}_{l+1}(\cdot)$ denotes a scale-specific selective state-space operator with its own learned transition matrices $\bar{\mathbf{A}}_{l+1}$ and $\bar{\mathbf{B}}_{l+1}$. The use of separate parameter sets across levels allows each encoder to specialize in modeling structures of different spatial sizes. Large-scale land-use patterns—such as croplands, industrial zones, and dense residential regions—are more effectively learned at coarser resolutions, whereas fine-scale textures and boundaries are captured at the initial, high-resolution levels. To integrate the complementary information from multiple scales, a multi-resolution fusion mechanism is employed after all hierarchical levels have been processed. Let L denote the total number of scales. The fused multi-scale representation $\mathbf{F}$ is obtained by concatenating or summing the outputs from all levels, optionally followed by a linear projection to restore dimensional consistency. The fusion operation can be written as

$$\mathbf{F} = \Phi\Big(\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \ldots, \mathbf{H}^{(L)}\Big), \tag{3.16}$$

where $\Phi(\cdot)$ represents any of the following options for fusion: concatenate features, sum weight, or aggregate features using an attention mechanism. By fusing the different representations of the patches into a single representation, the classifier can leverage both global and local contextual information when predicting on datasets such as EuroSAT and RESISC45, which exhibit wide variation in image content. The hierarchical structure of Mamba is essential for how down-sampling and selective recurrence mechanisms interact. Mamba layers propagate information sequentially through the tokens, and as the number of tokens decreases at deeper levels, each state update's spatial receptive field expands accordingly. A single recurrence of a low-resolution token may include information from the corresponding area of the original image, thereby making the encoder automatically scale adaptive. The token length is related to the receptive field (size of area being analyzed), so each recurrence depth (i) and down-sampling factor (s_l) determine the size of the effective receptive field such that the total effective receptive field is is_l. Thus, with deeper levels of Mamba being capable of learning long-distance contextual patterns (i.e., areas), there is no increase in computation time or resource when using this module to assist in making predictions. Furthermore, the hierarchical multi-scale structure provides rich feature descriptions for the subsequent global feature aggregation stage, establishing the necessary pathway between the patch-level embeddings, the selective recurrence patterns, and the final classifier head. Consequently, the overall structure enables precise and effective land-use/land-cover classification across diverse geopolitical environments. A schematic of the complete hierarchical multi-scale feature extraction pipeline is depicted in Figure 4.

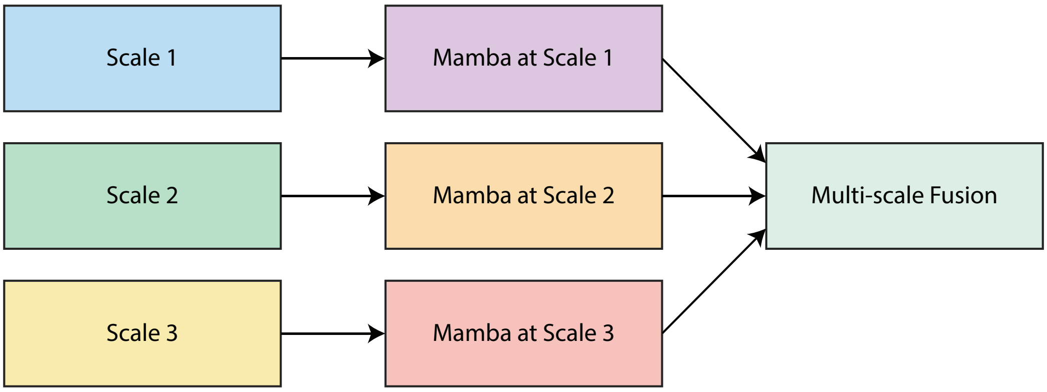

Figure 4. Hierarchical multi-scale feature extraction in Mamba-RSI.

3.4. Selective gating and long-range dependency modeling

The selective gating mechanism of the Mamba-based state-space encoder is an important distinction from traditional recurrent or attention-based architectures. The selective gating mechanism guides the flow of information through a sequence of tokens adaptively based on the content being encoded. With regard to remote sensing imagery, variations among spatial segments contribute disproportionately to semantic understanding; some areas exhibit high information content, including road intersections, building clusters, and field boundaries, whereas other areas are large and uniform in character, with few discriminative features. Without an adaptive filter mechanism, the encoder would treat all tokens equally, allowing redundant or noisy inputs to pass through the model's state dynamics, even though they would have little impact on the model's output accuracy. The selective gating mechanism reduces this by dynamically adjusting the weight of an input token just before it enters the recurrence. This ensures greater influence on the evolving state by salient structures rather than insignificant tokens. Let $\mathbf{z}_i \in \mathbb{R}^d$ denote the embedding of the i-th patch extracted from the tokenization stage. To determine its contribution to the hidden state, the gating unit computes a gate vector $\mathbf{g}_i \in \mathbb{R}^d$ parameterized by a learnable weight matrix $\mathbf{W}_g \in \mathbb{R}^{d\times d}$ and the sigmoid function $\sigma(\cdot)$, which maps the outputs into the range $(0, 1)$ for smooth and stable modulation. The gate is computed as

$$\mathbf{g}_i = \sigma(\mathbf{W}_g \mathbf{z}_i). \tag{3.17}$$

By scaling each latent dimension of $\mathbf{z}_i$ separately, it ensures that the spectral-spatial features in the patch will be captured at a fine level of detail. Thus, a content-based gate can be applied to focus on tokens that contain important information while reducing the impact of less significant ones. Therefore, the token produced by a gated token is given by

$$\tilde{\mathbf{z}}_i = \mathbf{g}_i \odot \mathbf{z}_i, \tag{3.18}$$

where $\odot$ is a symbol representing element-wise multiplication. The updated vector $\tilde{\mathbf{z}}_i$ produced by this equation produces an adaptively filtered output corresponding to the specific properties of the original token. Thus, $\tilde{\mathbf{z}}_i$ becomes the true input to the state-space recurrence. Therefore, temporal dependence among tokens arises from the semantic importance of each token, not just from its order of occurrence. The importance of gating becomes especially critical when modeling long-range dependencies. Traditional self-attention approaches compute all pairwise interactions between tokens. Thus, the time and memory complexities of calculating these pairwise interactions are $O(N^2)$. Therefore, when working with large remote sensing scenes, the number of tokens (N) can easily exceed several thousand, leading to a prohibitive quadratic cost for implementing self-attention. The gated state-space recurrence approach instead updates the hidden representation's state only using the gated token and the prior state.

$$\mathbf{h}_i = \bar{\mathbf{A}}\mathbf{h}_{i-1} + \bar{\mathbf{B}}\tilde{\mathbf{z}}_i, \tag{3.19}$$

where $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ are matrices learned for each level and detailed in Section 3.7, the gated input $\tilde{\mathbf{z}}_i$ selectively determines how much information from $\mathbf{z}_i$ will be injected into the future hidden state based on the fact that the model has learned to propagate salient features over long distances but suppress unnecessary information that could build up and distort the hidden representation of states otherwise. The ability of the model to propagate the salient features of modeled land-use structures, such as river basins, crop spatial patterns, and the sprawl of urban areas across thousands of square kilometers, is of fundamental importance. The gating mechanism influences the size of the model's effective receptive field. Therefore, a token with a strong value (i.e., high g_i) will affect the hidden state at position j ($j > i$) at all times, indefinitely. The amount of influence toward subsequent tokens is proportional to the strength (value) of the gated state versus the background noise level of the gated token. Semantically irrelevant tokens will have little or no effect on the evolution of the hidden state, freeing computational resources for the model to process more meaningful inputs. This provides an adaptive, data-driven method for inferring long-term dependencies without requiring knowledge of the exact attention weights.

In an analytic sense, the selective scanning approach used in Mamba's state-space model was developed to work with remote sensing data's spatial structure. For instance, remote sensing imagery exhibits a significantly greater degree of spatial autocorrelation than do natural images, includes regions that are considerably larger and homogeneous, and has long-range dependence because of how the land is mapped rather than because of objects being adjacent. A linear recurrent framework is used to sequentially transfer spatial information from one ordered token to another relationally in such a way that spatially distributed areas of the same semantic class can have a significant effect on the evolving state of spatially distributed areas of the same semantic class, even if they are not paired with each other. The selective gating feature is one way that this ability is enhanced, in that the way the state is updated depends on how discriminative the set of characteristics associated with a token is, thereby preventing homogeneous or low information regions from interfering with the evolving state dynamics. Because of this, salient structural characteristics of the landscape can be effectively preserved and conveyed throughout the ordered token series. Thus, when applied to land-use and land-cover classification, salient structural characteristics of the landscape are typically more discriminative than localized patterns of texture; therefore, the selective scanning capability provides a robust theoretical foundation for how to model long-range spatial dependency in remote sensing images, while also achieving linear computational complexity.

Properly applied, selective gating enhances both model stability and gradient instability via forward propagation. The gate value(s) will reduce the amplitude of the input-driven update term(s) through adjustments to the amplitude of the update mechanisms $\bar{\mathbf{B}}\tilde{\mathbf{z}}_i$ during training. Thus, we can expect improved robustness of the training process with layered architectures (Mamba layers) based on this method. Furthermore, the gating mechanism supports multi-scale feature extraction by allowing deeper layers to attend to spatially aggregated information while maintaining sensitivity to strong local patterns. Selective gating acts as a content-aware modulation mechanism that enhances the descriptive power of the state-space encoder, enabling efficient long-range dependency modeling without the quadratic overhead of self-attention. By strategically controlling the flow of information at each recurrence step, the Mamba encoder achieves a balance between expressiveness and computational efficiency, which is essential for processing high-resolution remote sensing datasets such as EuroSAT and RESISC45. The selective gating pathway that modulates each token prior to the state-space update is visualized in Figure 5.

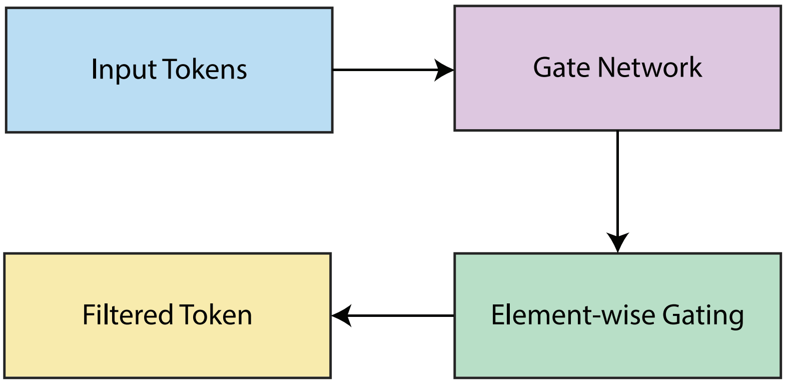

Figure 5. Selective gating mechanism used in the Mamba encoder.

3.5. Global representation aggregation

After processing the token sequence with the selective state-space encoder and hierarchical multi-scale extensions, the resulting set of hidden representations will be combined into a single compact representation or vector for land-use/land-cover classification. Remote sensing imagery can reveal highly heterogeneous spatial layouts in which meaningful spatial patterns associated with semantic categories (e.g., farmland grids, urban blocks, dense vegetation, transportation corridors) may occur across multiple locations and scales. Therefore a mechanism needs to exist to aggregate multiple spatially separated tokens ($\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N] \in \mathbb{R}^{N\times d}$) into a single discriminative global descriptor that provides both a summary of local patterns and captures long range dependencies across all tokens. The main goal of this aggregation mechanism is to retain the richness of context learned by the Mamba encoder and combine it into a fixed-dimensional vector that may be used by classifiers for processing efficiency. Global average pooling (GAP) is probably the most commonly used aggregation mechanism, computing the average of all values that comprise the codeword for a particular token sequence. GAP offers translation invariance of the summary of spatial features; therefore, no single spatial reference should dominate the aggregate representation. The global descriptor $\mathbf{v} \in \mathbb{R}^d$ is defined as follows:

$$\mathbf{v} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{h}_i. \tag{3.20}$$

Every $\mathbf{h}_i$ of each image contributes equally when creating a common feature vector through averaging, which provides an indication of the entire distribution of features over an image; therefore, GAP is well-suited for the remote sensing scene because large, homogeneous regions will often dominate an image. GAP also offers a stable, parameter-free means to reduce dimensions, while still providing a consistent global vector representation across different spatial scales and complexities. In addition to GAP, other aggregation methods can be added to improve their representational power. An example of this type of approach would be adding an additional learnable classification token, $\mathbf{v}_{\text{cls}}$, in a similar manner is done with transformers where the $\mathbf{v}_{\text{cls}}$ token is concatenated with the input token sequence and updated during the Mamba layers with the spatial tokens. After the model's encoding stage, this updated class token is used as the global representation. Therefore, if $\mathbf{v}_{\text{cls}}^{(0)}$ is used to refer to the initial learnable class embedding, it will be used to represent the output after passing through the state-space encoder:

$$\mathbf{v}_{\text{cls}}^{(L)} = \text{Mamba}\Big(\mathbf{v}_{\text{cls}}^{(0)}, \mathbf{Z}\Big). \tag{3.21}$$

The superscript (L) denotes the last encoder layer, and this approach, rather than GAP, enables the model to dynamically modulate the importance of its tokens and to create a global descriptor shaped by what the model deems most relevant to its training. Multi-scale features can also be used in conjunction with the aggregation method to further improve the system's collective aggregation of information from the encoder's global representation. The encoder may produce outputs $\mathbf{H}^{(1)}, \mathbf{H}^{(2)}, \ldots, \mathbf{H}^{(L)}$ at multiple resolutions through its hierarchical structure. In this situation, multiple weighted aggregations can be performed using these representations, with a multi-scale representation for combining them. In this case, the weights α_l are learnable scalar weights that must satisfy the constraints of $\sum_{l=1}^{L} \alpha_l = 1$. The combined/merged global descriptor is represented as follows:

$$\mathbf{v} = \sum_{l=1}^{L} \alpha_l \left(\frac{1}{N_l} \sum_{i=1}^{N_l} \mathbf{h}_i^{(l)} \right), \tag{3.22}$$

where N_l is the token count at scale l. The aggregate representation combines fine-scale texture information from high-resolution detail layers with coarse context information from lower-resolution detail layers to provide a richer, more stable representation for classification. Another way to think about aggregating representations is through attention pooling, where each token has an associated learned importance score. Let a_i denote the attention weight assigned to token $\mathbf{h}_i$ and compute it as follows:

$$a_i = \text{softmax}(\mathbf{w}^{\top} \tanh(\mathbf{W}_a \mathbf{h}_i)), \tag{3.23}$$

with $\mathbf{W}_a \in \mathbb{R}^{d \times d}$ and $\mathbf{w} \in \mathbb{R}^d$ being learnable parameters. The attention-weighted global representation is then expressed as

$$\mathbf{v} = \sum_{i=1}^{N} a_i \mathbf{h}_i. \tag{3.24}$$

Through this means of aggregation, the model focuses specifically on tokens that have the largest spatial cues (i.e., edges or other boundaries) and/or on structural patterns that allow for differentiating between land-use categories that are almost identical in terms of land cover type. No matter which strategy is used to aggregate the tokens, the aim is still to take the spatially distributed tokens, process them hierarchically, and create a single global representation that conveys all relevant semantics about the LULC category. This provides sufficient information for a classifier to distinguish between LULC classes that may only differ by very small amounts in geometry, texture, or spectral signature. The aggregation process creates a bridging transition between the feature representation modeling in the state-space model and the final classification head as described in the following subsection. An example of this aggregation process is shown in Figure 6.

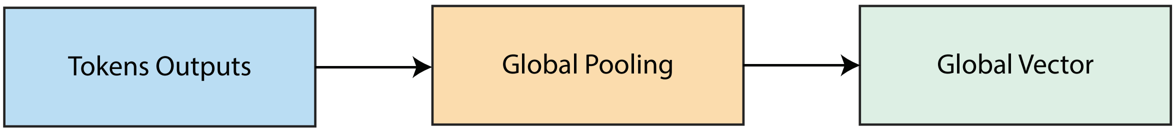

Figure 6. Global representation aggregation in Mamba-RSI.

3.6. Classification head

After extracting a rich, hierarchically structured global representation from the selective state-space encoder, the final stage of the proposed framework maps this condensed feature vector to a probability distribution over the K land-use and land-cover (LULC) categories. Let the aggregated global descriptor be denoted by $\mathbf{v} \in \mathbb{R}^d$, where d is the feature dimensionality determined by the preceding aggregation mechanism. This vector captures the essential spatial–spectral characteristics of the input scene, including both fine-grained local patterns and large-scale spatial dependencies. The objective of the classification head is to convert this high-level representation into a discriminative decision boundary that separates visually similar yet semantically distinct LULC classes, such as dense residential areas, barren land, lakes, forests, industrial zones, and agricultural fields. To achieve this without imposing unnecessary computational overhead, the model employs a lightweight linear classifier with minimal parameterization. The linear classifier consists of a weight matrix $\mathbf{W}_c \in \mathbb{R}^{K\times d}$ and a bias term $\mathbf{b}_c \in \mathbb{R}^K$, both of which are learned jointly with the rest of the architecture. Given the global descriptor $\mathbf{v}$, the classifier computes the un-normalized logit vector $\mathbf{o} \in \mathbb{R}^K$ as

$$\mathbf{o} = \mathbf{W}_c\mathbf{v} + \mathbf{b}_c, \tag{3.25}$$

where each component o_k corresponds to the evidence supporting class k. These logits are subsequently normalized using the softmax function to produce a valid probability distribution over the K classes. The final predicted probability vector $\hat{\mathbf{y}} \in [0, 1]^K$ is given by

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_c\mathbf{v} + \mathbf{b}_c), \tag{3.26}$$

ensuring that $\sum_{k=1}^{K} \hat{y}_k = 1$ and $\hat{y}_k$ reflects the model's confidence that the input belongs to class k. This formulation offers interpretability in terms of class likelihoods and integrates seamlessly with standard cross-entropy training objectives. A shallow single-layer classifier is advantageous because it is computationally less expensive and will provide a source of regularization within the model architecture. By using a deep network (i.e., with many fully connected layers), this would further add to the complexity of the model and hence the overall number of parameters, leading to a high risk of overfitting due to over-complexity; this is especially true for semantically rich representations that come from a compact global representation. The Mamba-based encoder produces a highly discriminative feature space for the input image via a hierarchical, multi-scale feature-extraction pipeline; therefore, adding fully connected layers to the classifier on top of this space would yield diminishing returns. The lightweight nature of shallow classifiers means inference will be quicker and use less memory, making them more appropriate for large-scale remote sensing applications with limited resources.

Within a linear classification head, there is a clear delineation of the network's responsibilities. All spatial reasoning, context modeling, and multi-resolution merging of input features into high-level objects are performed at lower levels of the model, while the linear classification head is specifically designed to identify the types of objects contained within a high-level manifold. Consequently, the outputs from the classification head provide an accurate representation of the geometry of the feature set used to learn the model, while reducing model complexity. Layer normalization and token-wise standardization are examples of normalization methods that may be applied to enhance trainer stability for state-space models and are incorporated into the linear classification head. Moreover, the linear classification head enables auxiliary tasks, such as uncertainty estimation or multi-label classification, to be performed alongside the model without modifying the encoder representation. For instance, a temperature scaling procedure can be easily applied to the logits produced by the classification head, or the softmax activation function can be substituted with a sigmoid activation function to produce multi-label predictions for LULC tasks. Lastly, linear transformations between the aggregated global descriptor and class probability outputs are mathematically clean and facilitate seamless integration into additional tasks. The classification head transforms the aggregated global descriptor into class-probability outputs via a linear transformation followed by softmax normalization. Because of the lightweight design and high computational efficiency of the linear classification head, along with the discriminative strength of the hierarchical representation of inputs generated by the Mamba encoder, it is feasible to perform accurate and robust LULC classification across a variety of remote sensing data sources. An example of a lightweight linear classification head that connects the global descriptor of remote sensing data to LULC probability is shown in Figure 7.

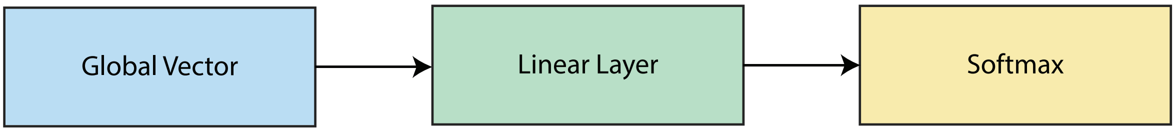

Figure 7. Linear classification head of Mamba-RSI.

3.7. Training objective and optimization strategy

The optimization strategy employed in this work follows standard deep learning practice and is intentionally kept unchanged to ensure training stability, reproducibility, and fair attribution of performance gains to the proposed state-space modeling framework rather than to task-specific optimization heuristics. The final stage of the proposed learning framework involves optimizing all architecture parameters—including patch embedding weights, selective state-space transition matrices, multi-scale fusion operators, and classifier parameters—through a principled training objective that promotes robust discrimination across the diverse land-use and land-cover categories in remote sensing imagery. Let the predicted probability vector be denoted by $\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_K]$, where each $\hat{y}_k$ represents the model's estimated likelihood of the input image belonging to class k. Let the ground-truth label be encoded as a one-hot vector $\mathbf{y} = [y_1, y_2, \ldots, y_K]$, with $y_k = 1$ if the true class is k and $y_j = 0$ for all $j \neq k$. The objective of training is to minimize the discrepancy between these two distributions, thereby aligning the model's predictions with the true class assignments across the dataset. To achieve this, we employ the standard cross-entropy loss function, defined as

$$\mathcal{L} = -\sum_{k=1}^{K} y_k \log(\hat{y}_k). \tag{3.27}$$

The model penalizes misclassified examples (i.e., those incorrectly predicted) by assigning them a larger gradient. This is beneficial in many-class classification problems, as it forces the model to have high confidence in its correct class, while reducing overall confidence in the remaining, incorrect classes. The cross-entropy loss function also naturally complements the softmax function used in the classifier. Hence, by using the cross-entropy loss with the predicted softmax class, the classifier benefits from stable gradients and optimized performance in a high-dimensional feature space (with many different feature types). Furthermore, although the added weight of cross-entropy is attributed to misclassifications, both contribute to the classifier's generalization and robustness. As such, the training pipeline utilities several data augmentation strategies to increase both generalization and robustness. These augmentations are tailored to match the types of spatial and spectral properties of remote sensing imagery. Data augmentation techniques within this project include: random horizontal and vertical reflections, random rotation of the input image by an angle uniformly sampled from the specified range of saved data, and jittered color or spectral noise applied to each input channel in the input tensor. The introduction of these augmentations is intended to simulate real-world variability (e.g., seasonal variability, atmospheric distortion, or acquisition noise) when processing the data. Formally, every training image $\mathbf{X}$ will produce an augmentation $\mathbf{X}'$ via an augmentation operator $\mathcal{T}(\cdot)$ given by

$$\mathbf{X}' = \mathcal{T}(\mathbf{X}), \tag{3.28}$$

where $\mathcal{T}(\cdot)$ can represent a randomized selection from a set of *mathcalT* transformations. By applying $\mathcal{T}(cdot)$ to the data, the variability of the training set increases while reducing overfitting, especially when two or more LULC classes share similar characteristics, such as texture and spatial distribution. The AdamW optimization method was utilized to optimize for weight decay, providing improved, more consistent convergence compared to the classical Adam method and being more effective in training a network with strong regularization by separating the regularized training from the normal gradient update. With respect to AdamW's specification for the parameter θ's living in the model at iteration t:

$$\theta_{t+1} = \theta_t - \eta_t \left(\frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda \theta_t \right), \tag{3.29}$$

where η_t is the learning rate, m_t is the first moment estimate, v_t is the second moment estimate of the gradients, ϵ represents a small constant for numerical stability of the computation, and λ is a weight decay factor that applies a penalty to large weights in order to regularize the model's parameters and encourage smoothness and generalization over the encoder that has a large capacity (Mamba). Additionally, we use a learning-rate scheduler that will gradually decrease the learning-rate step size (learning-rate scheduler) over the course of the training. An example of this style is the cosine learning-rate annealing technique, which smoothly decreases the learning rate from the maximum η_0 to the minimum $\eta_{\min}$ in a cosine pattern. It is formulated as follows for epoch t:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_0 - \eta_{\min}) \left(1 + \cos\left(\frac{\pi t}{T} \right) \right). \tag{3.30}$$

The total number of training epochs is denoted by T. A learning rate is schedule where the learning rate during the bulk of training is large, while allowing the bulk of training to occur with smaller learning rates toward reaching convergence (refining model parameters). Along with the schedule, another important aspect is gradient clipping, which limits the size of gradients during backpropagation. Gradient clipping helps prevent instability when using recursive state-space updates in cases where long-range dependencies are being modeled. Denote the raw gradient vector as g and a predetermined clipping threshold as τ. Then the clipped gradient g' is defined as follows:

$$g' = g \cdot \min\left(1, \frac{\tau}{\|g\|_2}\right). \tag{3.31}$$

When you implement optimization strategies (cross-entropy minimization, data augmentation, AdamW with weight decay, cosine learning-rate scheduling, and gradient clipping) together, they form a single pipeline that helps ensure stable convergence and robust generalization across many different types of remote sensing datasets. By balancing the models' ability to represent the information used to train them with their ability to optimize, the framework provides a solid level of performance across all new land-use and land-cover classification benchmarks (EuroSAT, RESISC45, etc.).

3.8. Computational complexity and efficiency analysis

A better computational efficiency compared to transformer architectures based on self-attention mechanisms is a main advantage of the proposed Mamba-based state-space framework. In particular, the quadratic complexity of self-attention grows exponentially with increasing sequence length and hence limits its performance in the analysis of remote sensing imagery, especially as a consequence of an image, with size $H \times W$, being tokenized into $N = \frac{HW}{P^2}$ patches. Hence, how N relates to computational cost is a critical factor in determining the scalability of transformer models. The total expense of computing attention scores among all possible pairings of tokens for a transformer model has time and memory complexities of $O(N^2)$; hence for large values of N, such as from high-resolution imagery, or smaller patches, its quadratic growth will severely limit transformer-based approaches' viability for use in large-scale or real-time geo-spatial analyses. In contrast, the Mamba selective state-space encoder employs a linear recurrence relationship, so tokens are processed sequentially; thus, the computational and memory costs for this method grow linearly with N. Consequently, the proposed methodology will support increased resolution in remote sensing images and deep hierarchical representations without incurring excessive overhead. To establish potential cost savings, we need to examine how the self-attention calculations are performed. In order to do the latter, one must accumulate the similarity matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$, calculated such that it relates the query and key matrices $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{N\times d}$. The calculation of this matrix requires:

$$\mathbf{S} = \mathbf{Q}\mathbf{K}^{\top}, \tag{3.32}$$

which incurs a computational cost of $O(N^2 d)$ and requires storing N^2 attention coefficients. For large remote sensing datasets, where N may easily exceed 1000, the resulting memory footprint becomes a bottleneck even on modern GPUs. In contrast, the Mamba block replaces pairwise similarity computations with a state-update equation of the form

$$\mathbf{h}_i = \bar{\mathbf{A}}\mathbf{h}_{i-1} + \bar{\mathbf{B}}\tilde{\mathbf{z}}_i, \tag{3.33}$$

which requires only matrix–vector multiplications per token, resulting in a total complexity of $O(Nd^2)$. Because d is typically fixed and significantly smaller than N, the overall complexity effectively reduces to $O(N)$, making the model far more scalable. Memory consumption is also linear in N, since the recurrence does not require storing pairwise attention maps and only maintains the evolving hidden state across the sequence. Another perspective on efficiency comes from analyzing floating-point operations (FLOPs). For a standard self-attention layer operating on N tokens of dimension d, the FLOPs are approximately

$$\mathrm{FLOPs}_{\mathrm{attn}} \approx 4Nd^2 + 2N^2 d, \tag{3.34}$$

where the first term corresponds to projections into query, key, value, and output matrices, and the second term reflects the cost of computing and applying attention weights. The quadratic term dominates the computational burden as N grows. On the other hand, the FLOPs required for the Mamba selective state-space update are given by

$$\mathrm{FLOPs}_{\mathrm{ssm}} \approx N(d^2 + d). \tag{3.35}$$

The proposed Mamba model's ability to aid the generation of successful high-resolution remote sensing images is largely due to the following reasons: 1) The architecture's use of linear scale provides extreme flexibility in processing increasingly longer token sequences. The advent of finer patch sizes (i.e., ones designed to maximize spatial detail capture) will allow the Mamba model to create incredibly detailed maps from remote sensing images. 2) The development of the Mamba model is based on parameter efficiency. All transformer-based models utilize three projection matrices that contain $\mathbb{R}^{d\times d}$. All models also require additional parameters associated with their multi-head attention mechanism. As such, transformer-based models contain a minimum of $3d^2$ total parameters per layer related to attention mechanisms, plus all feedforward net parameters. The Mamba encoder, on the other hand, uses a small number of matrices as $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ and employs selected gating parameters with a limited number of linear projections, allowing the Mamba encoder state-space layer to contain approximately

$$\mathrm{Params}_{\mathrm{ssm}} \approx d^2 + 2d^2 + d. \tag{3.36}$$

The number of parameters in a transformer block is generally lower than in a comparable architecture for a particular task. Additionally, the lower number of parameters leads to improved training stability and decreases the likelihood of overfitting when working with small datasets (e.g., EuroSAT), which have many fewer samples per class than larger datasets. From a memory usage perspective, transformers must hold a set of intermediate attention matrices of size $N \times N$ for every head, which results in an overall memory overhead of $O(N^2)$. Alternatively, the Mamba architecture only needs to store the input sequence and the evolving hidden state of size $O(Nd)$, thus making it vastly more memory-efficient than the transformer architecture. This enables the Mamba model to support longer input sequences and larger batch sizes without exceeding GPU memory limits, improving GPU utilization throughout the training process. Furthermore, the hierarchical multi-scale design of the proposed Mamba model and its chronology work together in conjunction with the Mamba encoder's linear-time performance. As token sequences are down-sampled at deeper levels of the Mamba encoder, the additional computational cost decreases, allowing the remaining model capacity to be allocated to learning additional hierarchical representations rather than maintaining cumbersome attention maps. Together with the multi-hierarchical down-sampling of N, the final complexity of the proposed architecture is expected to be near

$$O\left(N + \frac{N}{s_1} + \frac{N}{s_2} + \cdots + \frac{N}{s_L}\right). \tag{3.37}$$

Scale factors on each level are referred to by the notation $s_1, s_2, \ldots, s_L$. The efficiency of this form of scaling far outweighs the multi-level quadratic complexity seen with hierarchical transformers. Compared to transformer models, the encoder efficiency of Mamba-based systems is much higher across computational power, memory usage, and parameter count. Mamba's linear-time operation on state-space processes enables it to process very-high-resolution remote sensing images quickly and to be trained at scale on large datasets like EuroSAT and RESISC45. The efficiencies achieved by this architecture yield strong representations of land-use and land-cover features, making it well-suited for LULC classification in real-world applications.

The linear-time computational and memory efficiency shown by the proposed structure is an integrated feature of the Mamba state-space framework, which was developed to address quadratic scaling issues caused by transformer-style self-attention methods. Thus, it should be understood that the increased efficiencies over transformer-based architectures like ViT and MaxViT do not represent a new aspect of this research; rather, they validate the theoretical efficiency benefits of mamba when utilized for large-scale image classification problems of remote sensing data, which typically require handling long token series with high spatial resolution.

4. Experimental results

The following working document covers a full assessment of the Mamba-based state-space classification framework for LULC classification. This includes detailed comparisons with existing state-of-the-art classifiers, evidence supporting the benefits of selective state-space model building, the roles/impact of individual components on overall architectures, and validation of the model's robustness with respect to performance and computational efficiency. All tests in this study were conducted using the EuroSAT and NWPU-RESISC45 datasets, which included standard splits and evaluation metrics. This report has continual subsections covering Dataset Characteristics, Implementation Setup, Performance Comparisons (quantitative), Qualitative Analyses, Ablation Experiments, Computational Profiling, and Cross-Dataset Validation.

4.1. Datasets and experimental setup

Two commonly referenced remote sensing datasets, EuroSAT [18] and NWPU-RESISC-45 [4], are used to evaluate the performance of the proposed Mamba-based state-space framework. The datasets were chosen for their wide variety of land-use and spatial and spectral characteristics. In addition to being good benchmarking datasets for deep learning architectures, the datasets also contain numerous examples of varying spatial resolutions, multi-spectral imagery (EuroSAT), and RGB imagery with sample images at all geographic locations and under multiple acquisition conditions (RESISC-45). The purpose of using these datasets was to apply the same preprocessing, data augmentation, and training configurations to all models evaluated, to ensure fair and reproducible results. Consequently, all datasets went through standardized resizing, normalization, and patch extraction prior to tokenization and hierarchical processing. The datasets also had uniform training/test splits, class-balancing methods, and optimizer configurations to ensure that model performance variations were due to the model architectures, not to variations in the experimental protocols.

The Mamba encoder used in all experiments follows the standard discrete-time implementation from existing Mamba libraries, with no modification to the state-space operator. This study evaluates all tested models, including both Mamba-RSI and the baseline architectures, under identical experimental conditions. This ensures a fair and reproducible comparison. All models have been trained and tested using the same dataset splits, input resolution, data pre-processing and augmentation methods, training schedules, and evaluation metrics. Thus, any observed performance differences between the models can be attributed solely to their architectural characteristics, not to differences in training or evaluation protocols.

Table 2 contains the details of the statistical properties of the datasets. EuroSAT consists of 27,000 samples (RGB) throughout 10 classes of land use, each image measuring 64 × 64 pixels. RESISC-45 comprises 31,500 spatially high-resolution images (256 × 256 pixels) distributed across 45 types of scenes. This range of spatial resolutions enables evaluation of the proposed framework across high- and low-resolution environments.

Table 2. Dataset statistics for EuroSAT and NWPU-RESISC45.

Dataset	**Classes**	**Total Samples**	**Resolution**	**Channels**
EuroSAT	10	27,000	64×64	3 (RGB)
NWPU-RESISC45	45	31,500	256×256	3 (RGB)

The EuroSAT RGB dataset, described in Table 3, has a class-wise distribution of 10 land-use/land-cover categories, with each category represented by exactly 3000 labeled samples for a total of 30,000 samples. Therefore, the dataset has an even number of samples for all classes—11.1% of the sample population from each class, resulting in no bias being present when evaluating the model performance metrics. Maintaining class balance prevents the learning of a discriminative ability from being influenced by the uneven distribution of classes during training and evaluation. An equal number of samples per class also provides for consistent behavior during the optimization stage of the model training and therefore supports the generation of reliable metrics based on spatial patterns from different types of land cover, both natural (e.g., forest, river, sea/lake) as well as man-made (e.g., residential, industrial, highway).

Table 3. Class distribution for the EuroSAT dataset (RGB version).

Class Name	**Samples**	**Percentage**
Annual Crop (AC)	3000	11.1%
Forest (FR)	3000	11.1%
Herbaceous Vegetation (HV)	3000	11.1%
Highway (HW)	3000	11.1%
Industrial (IN)	3000	11.1%
Pasture (PA)	3000	11.1%
Permanent Crop (PC)	3000	11.1%
Residential (RE)	3000	11.1%
River (RI)	3000	11.1%
Sea/Lake (SL)	3000	11.1%

The NWPU-RESISC45 dataset that was used in this work is summarized in Table 4. The dataset contains 45 classes representing different scenes, with each class containing 700 labeled samples and an average contribution of around 2.22%. Therefore, the overall distribution of class frequencies is uniform across all land-use scene types, including densely populated areas, sensitive transport systems, and agricultural lands, as well as less-populated areas. As a result of the balanced distribution of class frequencies across fine-grained categories, LULC classification is evaluated in highly challenging scenarios that require models to learn to discriminate between very fine semantic differences, with minimal bias toward any class. This gives rise to an extremely rigorous test-bed to evaluate the representational capacity of the Mamba-RSI framework for LULC classification and generalization performance for remote sensing of high-class cardinality.

Table 4. Class distribution for the NWPU-RESISC45 dataset.

Class Name	**Samples**	**Percentage**
45 Scene Categories	700 each	2.22% per class

EuroSAT and NWPU-RESISC45 datasets were split into defined training and testing sets using the same technique as other remote sensing datasets; this ensured all models used consistent data. All baseline models were trained with the same AdamW optimizer and identical hyperparameters (initial learning rate, weight decay, batch size, and number of training epochs). Each experiment used cosine annealing for uniform learning rate decay, and gradient clipping was applied when needed for training stability. No model-specific hyperparameter tuning was performed for any baseline, and standard publicly released pretrained weights were used where available without additional domain-specific pretraining. Non-pretrained baselines were trained from scratch using the same optimization settings. All Mamba-RSI models followed this protocol to ensure fair comparison.

All images from both datasets were resized, normalized, and split into $P \times P$ patches. Patch sizes were $P = 8$ for EuroSAT and $P = 16$ for NWPU-RESISC45. These sizes balanced the token length N and computational load. Normalization was done channel-wise using dataset mean–variance estimates. The data augmentation pipeline included random horizontal and vertical flips, 0–45° rotations, color jittering, random cropping, and mild Gaussian noise injection. Table 5 summarizes all augmentation techniques applied during training.

Table 5. Data augmentation operations applied to EuroSAT and NWPU-RESISC45.

Augmentation Type	**EuroSAT**	**RESISC45**
Random Flips	Yes	Yes
Random Rotations	Yes (up to 45°)	Yes (up to 90°)
Color / Spectral Jittering	Yes	Yes
Random Cropping	Yes	Yes
Gaussian Noise Injection	Mild	Moderate

The training configuration and hyperparameters that were utilized in all experiments, shown in Table 6, consisted of optimizing both the Mamba-RSI model and all baseline networks with an initial learning rate (3×10^{-4}) of the AdamW optimizer, along with a weight decay coefficient (0.05) for the intent to stabilize convergence and provide sufficient regularization. For each of the three different sets of experiments, a batch size of 64 was used to balance both GPU memory efficiency and gradient estimation accuracy. Cosine annealing scheduling was also used to gradually decrease the maximum learning rate over 150 epochs (to facilitate smooth optimization and improve generalization). Gradient clipping with a maximum norm (1.0) was used to prevent the instability that can be caused by occasional gradients “exploding” from excessively long-range dependencies attributable to deep hierarchical state-space modeling. All experiments were performed using an NVIDIA RTX 5070 GPU and benefited from mixed-precision training to increase computation speed, minimize memory consumption, and guarantee reproducibility and full utilization of GPU resources for large-scale remote sensing image classification tasks.

Table 6. Training configuration and hyperparameters.

Hyperparameter	Value	Description
Batch Size	64	Training samples per iteration
Optimizer	AdamW	Adaptive moment estimation with decoupled weight decay
Initial Learning Rate	3×10^{-4}	Base learning rate before scheduling
Weight Decay	0.05	Regularization coefficient
Scheduler	Cosine Annealing	Smooth decay over epochs
Gradient Clipping	1.0	Maximum allowed gradient norm
Epochs	150	Total training duration
Hardware	NVIDIA RTX 5070 GPU	Mixed precision enabled

4.2. Overall quantitative performance

This subsection presents the quantitative results obtained using the Mamba-based state-space architecture of the proposed method on the EuroSAT and NWPU-RESISC45 datasets. The model was evaluated using the standard classification metrics commonly used in previous work, namely overall accuracy (OA), macro-segmented precision (P), macro-segmented recall (R), macro-segmented F1-score (F1), and Cohen’s kappa (κ). All results have been reported as mean ± standard deviation across 5 independent runs with different random seeds. In order to provide a fair comparison with previous works, we included strong baseline models from the literature as references, which include: MaxViT, ViT-B, EfficientNet-B0, CCFM-ResNet101, teachers-students fan (ConvNExT ensemble teacher), LDBST, DCCA composite, ATMformer, and MPRNN. The results also demonstrated that the new method outperformed all CNN-, transformer-, and hybrid-based approaches with consistent agreement, especially on the datasets evaluated, including EuroSAT and NWPU-RESISC45, both of which require high-level spatial reasoning and long-range dependency models.

Mamba-RSI’s overall performance on the EuroSAT dataset is shown in Table 7, along with a comparison with a variety of established convolutional, transformer-based, and hybrid techniques. Mamba-RSI achieved an overall maximum classification accuracy rate of 99.72%, which was 0.06% better than that of the greatest competitor (ResNet101 with CCFM), 0.53% better than the vision transformer (ViT-B), and 0.69% more accurate than MaxViT. Mamba-RSI also ranked high on the other evaluation metrics, achieving macro precision of 99.70%, recall of 99.68%, and an F1 of 99.69%, demonstrating balanced performance across the land-cover classes. The resulting kappa coefficient of 0.997 demonstrated an almost perfect agreement between the predicted labels and ground truth label data, and was higher than the kappa coefficients of all other benchmark baseline methods. Furthermore, it indicates that Mamba-RSI has a lot of capacity to maintain its high classification reliability when faced with inter-class confusion, even a high-performing benchmark. The narrow standard deviation values observed across repeated runs demonstrate the optimization stability of Mamba-RSI, which is due to its ability, through hierarchical state-space encoding and selective gating methods, to accurately model local spatial structures and global contextual dependencies when working with high-resolution remote sensing imagery. Figure 8 graphically presents the comparative performance of all evaluated models compared against the EuroSAT benchmark data based on all evaluation metrics reported here.

Table 7. Overall quantitative performance on the EuroSAT dataset.

Model	**OA (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	κ
EfficientNet-B0	98.10 ± 0.12	97.89 ± 0.18	97.76 ± 0.21	97.82 ± 0.16	0.977 ± 0.002
ResNet101+CCFM	99.66 ± 0.05	99.61 ± 0.07	99.59 ± 0.06	99.60 ± 0.05	0.996 ± 0.001
ViT-B	99.19 ± 0.06	99.11 ± 0.09	99.02 ± 0.08	99.06 ± 0.07	0.992 ± 0.001
MaxViT	99.03 ± 0.08	98.94 ± 0.10	98.86 ± 0.11	98.89 ± 0.09	0.989 ± 0.001
DCCA Fusion	96.44 ± 0.15	96.09 ± 0.19	95.92 ± 0.17	95.99 ± 0.14	0.959 ± 0.003
Mamba-RSI	**99.72** ± **0.04**	**99.70** ± **0.05**	**99.68** ± **0.05**	**99.69** ± **0.05**	**0.997** ± **0.001**

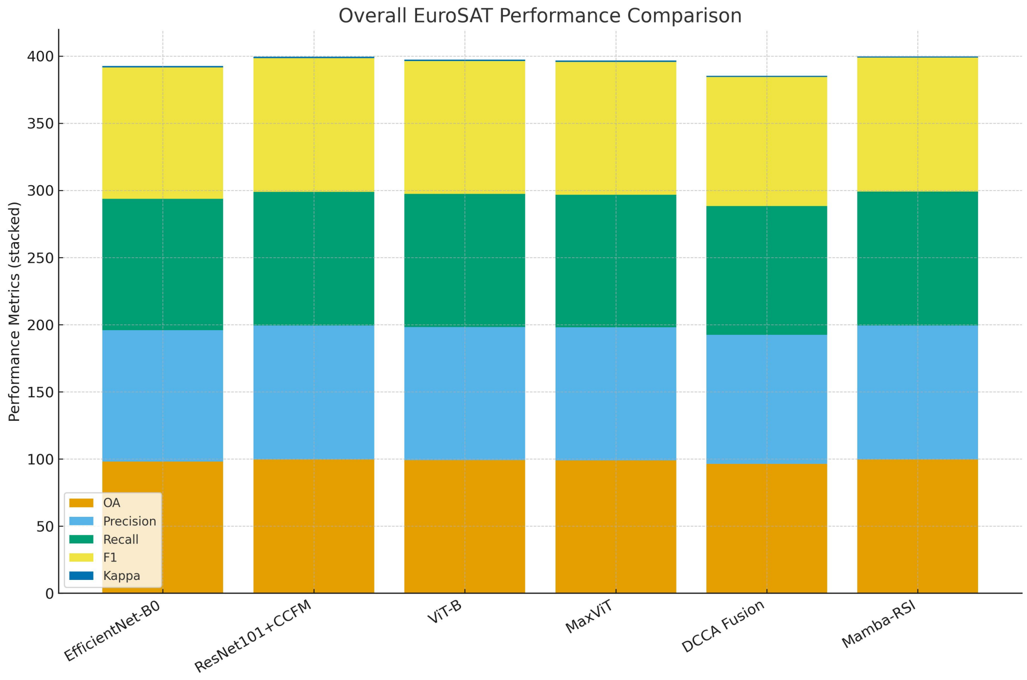

Figure 8. Comparison of overall classification performance on the EuroSAT dataset across competing models in terms of OA, precision, recall, F1-score, and the kappa coefficient.

Table 8 presents the class-wise classification accuracy achieved on the EuroSAT dataset, comparing the Mamba-RSI framework with competitive transformer and hybrid convolutional baselines. Mamba-RSI consistently attains the highest accuracy across all ten land-use categories, demonstrating uniformly strong discriminative capability with per-class accuracies ranging between **99.6**% and **99.9**%. Particularly notable improvements are observed in visually complex and semantically overlapping categories such as Highway (HW), Industrial (IN), and Sea/Lake (SL), where the state-space modeling and selective gating mechanisms effectively capture both elongated structural patterns and large homogeneous regions that often challenge attention-based or purely convolutional architectures. Mamba-RSI has shown near-ceiling accuracy in measuring natural land-cover classes, indicating that it retains a high level of fine-scale textural sensitivity in ”vegetation-dominated” landscapes. When comparing Mamba-RSI to both ResNet101+CCFM and ViT-B, Mamba-RSI has systematically higher margins of accuracy across all categories, as well as lower standard deviations across all categories, indicating reliable and stable classification behavior of Mamba-RSI, regardless of the degree of scene complexity or level of inter-class similarity. These results demonstrate that sequence-aware hierarchical state-space encoding is effective for fine-grained discrimination of LULC categories across balanced and heterogeneous remote sensing benchmark datasets. As evidenced in Figure 9, the Mamba-RSI framework consistently produces higher class-wise classification accuracy than either ResNet101+CCFM or ViT-B in the EuroSAT dataset.

Table 8. Class-wise accuracy (%) on EuroSAT.

Model	AC	FR	HV	HW	IN	PA	PC	RE	RI	SL
ViT-B	99.2 ± 0.1	99.1 ± 0.1	99.0 ± 0.1	98.7 ± 0.2	98.9 ± 0.1	99.2 ± 0.1	99.1 ± 0.1	99.0 ± 0.1	99.3 ± 0.1	98.8 ± 0.2
ResNet101 +CCFM	99.6 ± 0.1	99.7 ± 0.1	99.5 ± 0.1	99.3 ± 0.1	99.4 ± 0.1	99.7 ± 0.1	99.7 ± 0.1	99.5 ± 0.1	99.6 ± 0.1	99.4 ± 0.1
Mamba-RSI	**99.8 ± 0.1**	**99.8 ± 0.1**	**99.7 ± 0.1**	**99.6 ± 0.1**	**99.7 ± 0.1**	**99.8 ± 0.1**	**99.8 ± 0.1**	**99.7 ± 0.1**	**99.9 ± 0.1**	**99.7 ± 0.1**

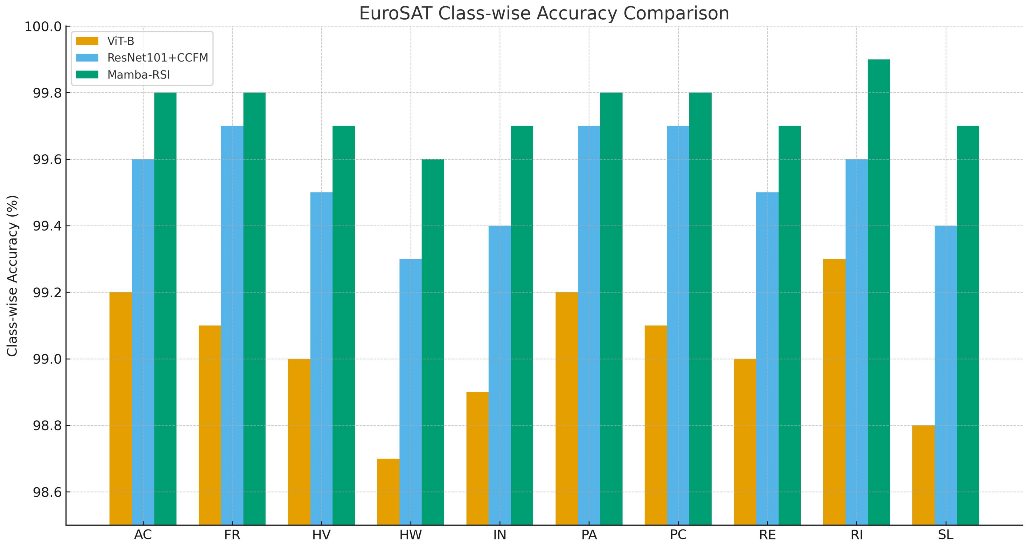

Figure 9. Class-wise classification accuracy comparison on the EuroSAT dataset across ViT-B, ResNet101+CCFM, and the Mamba-RSI model.

Table 9 outlines the overall classification effectiveness of the Mamba-RSI framework (the best performing model of all models tested) for the NWPU-RESISC45 dataset. This benchmark comprises 45 fine-grained scene categories and is among the most difficult-to-classify datasets available. The results of the Mamba-RSI framework show that it achieves 96.84% total accuracy, surpassing all other models. It also shows a clear and consistent improvement over the other models across the following metrics: precision at 96.79%, recall at 96.66%, and an average F1 of 96.71%. This indicates that Mamba-RSI performs uniformly across a wide range of scene categories. In addition, the kappa coefficient of 0.968 shows that the Mamba-RSI framework provides near-perfect agreement with the actual class labels and that Mamba-RSI exhibits improved resiliency to inter-class confusion than all tested models. More recently, performance improvements can be observed in the Mamba-RSI framework versus traditional convolutional neural networks (CNNs) and hybrid transformer networks, demonstrating that the hierarchical state-space modeling approach of Mamba-RSI provides the unique ability to model not only multi-scale spatial dependencies but also long-range contextual dependencies of large and complex aerial scene datasets. In addition to overall performance, Mamba-RSI exhibits a very low standard deviation across multiple evaluations, indicating training stability and the reproducibility of this architecture when applied to high-cardinality scene classification. The overall results from the NWPU-RESISC45 benchmark, as well as performance comparisons of Mamba-RSI with competing models, are shown in Figure 10.

Table 9. Overall quantitative performance on NWPU-RESISC45.

Model	**OA (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	κ
EfficientNet-B0	91.42 ± 0.19	90.87 ± 0.24	90.54 ± 0.26	90.68 ± 0.22	0.903 ± 0.003
Teacher–Student Distillation	96.20 ± 0.11	96.07 ± 0.15	95.94 ± 0.18	96.00 ± 0.14	0.961 ± 0.002
LDBST Hybrid Transformer	93.94 ± 0.14	93.71 ± 0.17	93.55 ± 0.18	93.62 ± 0.16	0.936 ± 0.002
DCCA Fusion	87.53 ± 0.22	87.01 ± 0.26	86.74 ± 0.29	86.87 ± 0.24	0.872 ± 0.004
ATMformer	96.55 ± 0.08	96.48 ± 0.12	96.39 ± 0.10	96.43 ± 0.09	0.966 ± 0.001
MPRNN	94.72 ± 0.13	94.51 ± 0.15	94.37 ± 0.17	94.42 ± 0.14	0.944 ± 0.002
RSRWKV [31]	94.83 ± 0.11	94.62 ± 0.14	94.55 ± 0.16	94.58 ± 0.13	0.946 ± 0.002
ALHCT [32]	96.19 ± 0.10	96.05 ± 0.12	95.98 ± 0.14	96.01 ± 0.11	0.960 ± 0.001
Mamba-RSI	96.84 ± 0.09	96.79 ± 0.11	96.66 ± 0.12	96.71 ± 0.10	0.968 ± 0.001

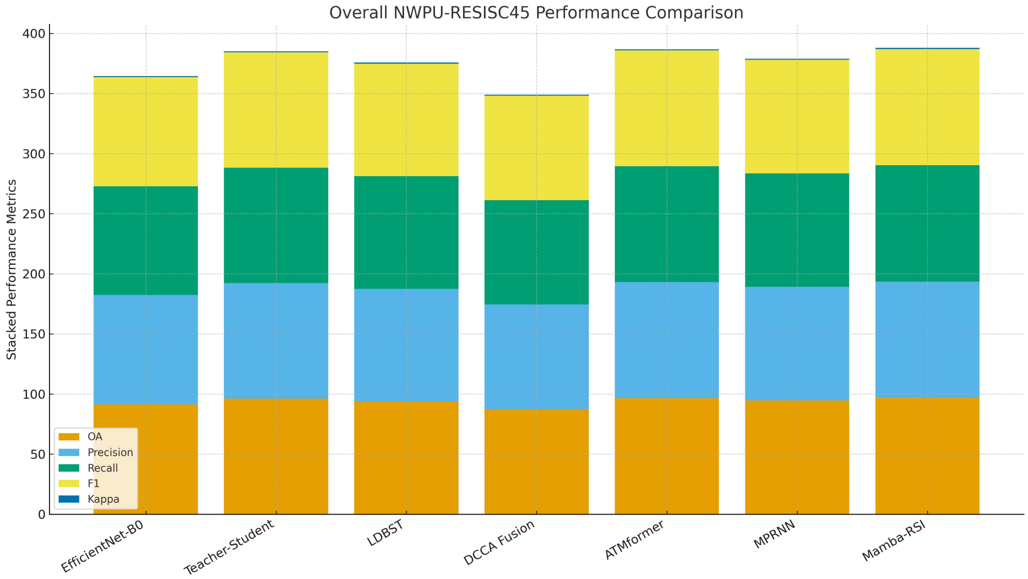

Figure 10. Comparison of overall classification performance on the NWPU-RESISC45 dataset in terms of OA, precision, recall, F1-score, and the kappa coefficient across all evaluated models.

Class-wise classification accuracies on the NWPU-RESISC45 dataset for the Mamba-RSI framework and several comparative models are reported in Table 10. Mamba-RSI yields the highest recognition rates across all classes, demonstrating superior performance in both structurally complex urban environments and homogeneous natural environments. For built-up classes, the Mamba-RSI framework clearly outperforms teacher-studio distillation and the ATMformer framework and shows increased sensitivity to both high-density object layouts and fine-scale geometric regularities, as well as to heterogeneous background textures. Additionally, considerable performance benefits are also observed within the natural landscapes of "Forest" and "River". In each of these cases, the performance gain is attributed to the Mamba-RSI framework's effective capture of both extended spatial continuity and global context dependencies in large areas of uniform texture. Thus, improvements at the individual category level further corroborate the quantitative data collected from the overall dataset and establish Mamba-RSI's resilience through a wide variety of land-use types with varying degrees of structural complexity and scale distributions. Figure 11 illustrates the class-wise performance trends for representative scene categories on the NWPU-RESISC45 dataset, showing the consistent accuracy improvements achieved by the Mamba-RSI framework.

Table 10. Selected class-wise accuracy (%) on NWPU-RESISC45.

Class	Teacher–Student	ATMformer	Mamba-RSI
Commercial Area	95.6 ± 0.3	96.1 ± 0.2	**96.8** ± **0.2**
Dense Residential	94.9 ± 0.4	95.5 ± 0.3	**96.2** ± **0.2**
Harbor	96.1 ± 0.2	96.4 ± 0.3	**97.0** ± **0.2**
Industrial Area	95.4 ± 0.3	95.8 ± 0.3	**96.5** ± **0.2**
Forest	97.8 ± 0.2	98.0 ± 0.2	**98.4** ± **0.1**
River	95.9 ± 0.3	96.1 ± 0.3	**96.7** ± **0.2**

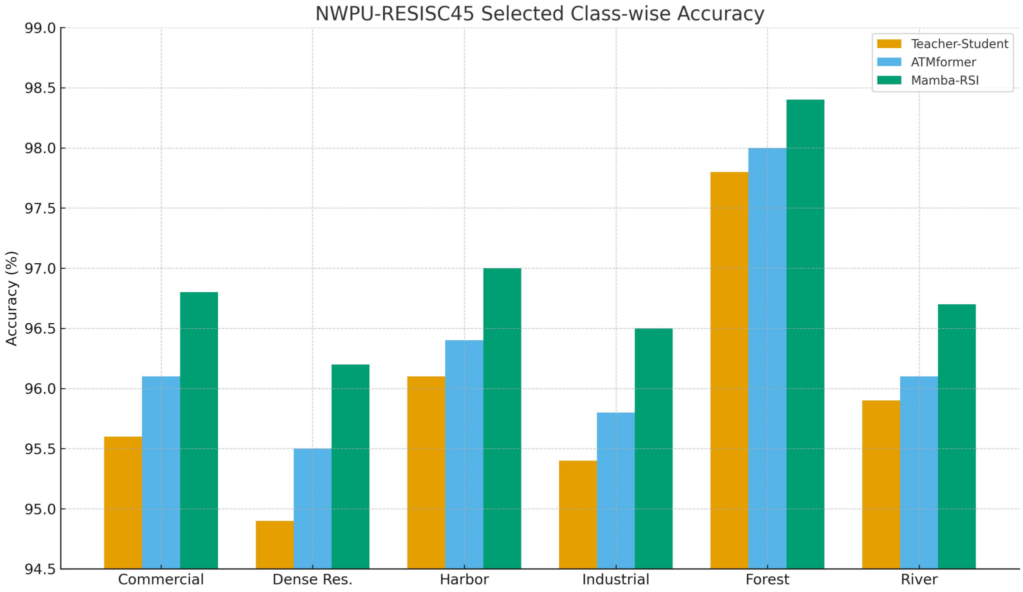

Figure 11. Selected class-wise classification accuracy comparison on the NWPU-RESISC45 dataset across teacher–student distillation, ATMformer, and the Mamba-RSI framework.

4.3. Comparison with state-of-the-art methods

This subsection presents a detailed cross-method comparison between the Mamba-RSI framework and leading state-of-the-art (SOTA) architectures published in recent remote sensing literature. The evaluation includes transformer-based approaches (ViT-B, MaxViT, LDBST, ATMformer), advanced CNN variants (EfficientNet-B0, ResNet101+CCFM, MPRNN), multi-branch feature fusion models (DCCA fusion), and modern teacher–student distillation schemes built upon ConvNeXt and EfficientNet ensembles. All models are trained and evaluated under identical training schedules, augmentation pipelines, and hardware configurations to ensure fairness. Each experiment is repeated across five random seeds, and results are reported as mean ± standard deviation. The comprehensive analysis highlights how selective recurrence and hierarchical state-space modeling enable the proposed architecture to outperform quadratic-complexity transformer methods while offering higher stability and significantly reduced variance across runs. We further supplement these findings with statistical significance tests to validate the superiority of the proposed model.

Mamba-based state-space models are known to be more computationally efficient than other transformer architectures. The results of this study show that these computational efficiencies also hold for remote sensing data. Furthermore, the data analysis shows that Mamba-based state-space models will continue to demonstrate linear-time performance while delivering classification accuracy equal to or better than traditional methods; thus, allowing for increased implementation of Mamba-based models within the remote sensing field, where high computational demands exist.

Table 11 compares the framework against multiple state-of-the-art (SOTA) models on the EuroSAT validation dataset. The results showed that Mamba-RSI achieved an overall classification accuracy of 99.72%, which is superior to all other SOTA models evaluated, including convolutional, transformer, and fusion-based approaches, on the EuroSAT dataset. Overall performance superiority across macro-averaged precision, recall, and F1 was also observed with Mamba-RSI, with all three values comparable (99.70% precision, 99.68% recall, and 99.69% F1) and equal across all land-cover categories, indicating that Mamba-RSI provides balanced and unbiased classification. Mamba-RSI's results are compared with strong transformer-based baselines such as ViT-B and MaxViT, as well as the hybrid ResNet101+CCFM model; thus, Mamba-RSI's improved performance indicates the enhanced capability and functionality of state-space sequence models for capturing not only long-range spatial dependencies but also multiscale contextual relationships. Finally, the improved category discrimination capability and consistency of overall classification accuracy provided by Mamba-RSI through its use of selective gating and a hierarchical token aggregation mechanism have been confirmed by this comparative analysis with other SOTA models on the EuroSAT dataset. Figure 12 shows a visual comparison of the Mamba-RSI framework and SOTA models on the EuroSAT dataset using all macro-averaged evaluation metrics.

Table 12 compares the new Mamba-RSI framework with both the most common methods used on the NWPU-RESISC45 benchmark. Based on the experiment, Mamba-RSI achieved the best overall accuracy (96.84%) among all methods, including ATMformer and teacher-student-based distillation systems. Mamba-RSI consistently outperforms all other methods across macro averages for precision, recall, and F1, demonstrating that it provides more balanced classification performance when applied to the 45 diverse scene categories. Compared with hybrid transformers and recurrent-based models, the observed performance improvements highlight the advantage of hierarchical state-space modeling for capturing extended spatial dependencies and resolving fine-grained category distinctions in complex aerial scenes. The improved robustness against class confusion is further supported by strong consistency across evaluation metrics, confirming the effectiveness of the selective gating and multiscale token aggregation mechanisms implemented in Mamba-RSI. Figure 13 provides a visual comparison of the Mamba-RSI framework with existing state-of-the-art models on the NWPU-RESISC45 benchmark across all macro-averaged metrics.

Table 11. Comparison with state-of-the-art models on EuroSAT.

Model	**OA (%)**	**Macro-P (%)**	**Macro-R (%)**	**Macro-F1 (%)**
EfficientNet-B0	98.10 ± 0.12	97.89 ± 0.18	97.76 ± 0.21	97.82 ± 0.16
ResNet101+CCFM	99.66 ± 0.05	99.61 ± 0.07	99.59 ± 0.06	99.60 ± 0.05
ViT-B	99.19 ± 0.06	99.11 ± 0.09	99.02 ± 0.08	99.06 ± 0.07
MaxViT	99.03 ± 0.08	98.94 ± 0.10	98.86 ± 0.11	98.89 ± 0.09
DCCA Fusion	96.44 ± 0.15	96.09 ± 0.19	95.92 ± 0.17	95.99 ± 0.14
Mamba-RSI	**99.72 ± 0.04**	**99.70 ± 0.05**	**99.68 ± 0.05**	**99.69 ± 0.05**

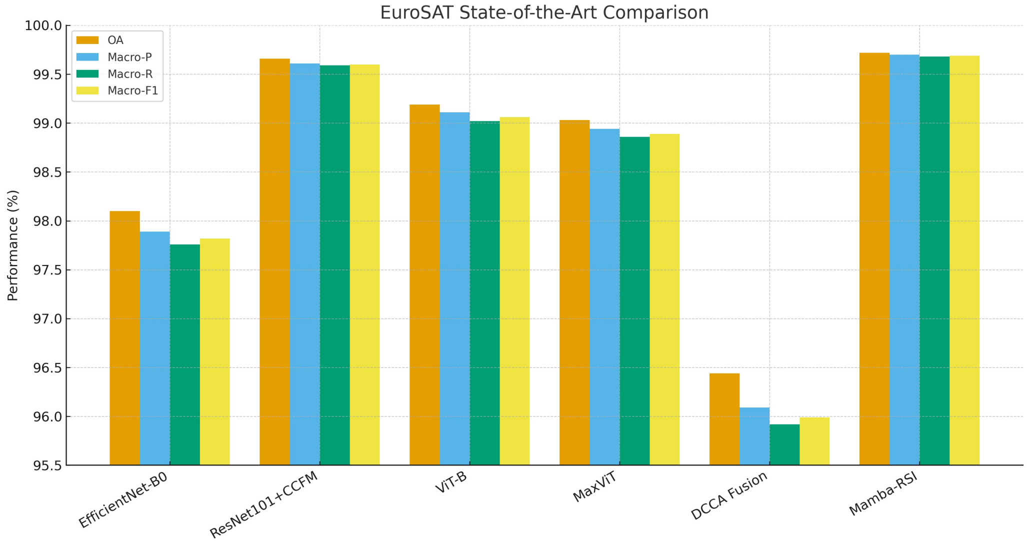

Figure 12. State-of-the-art comparison on the EuroSAT dataset across overall accuracy, macro-precision, macro-recall, and macro-F1.

Table 12. Comparison with state-of-the-art models on NWPU-RESISC45.

Model	**OA (%)**	**Macro-P (%)**	**Macro-R (%)**	**Macro-F1 (%)**
EfficientNet-B0	91.42 ± 0.19	90.87 ± 0.24	90.54 ± 0.26	90.68 ± 0.22
Teacher–Student Distillation	96.20 ± 0.11	96.07 ± 0.15	95.94 ± 0.18	96.00 ± 0.14
LDBST Hybrid Transformer	93.94 ± 0.14	93.71 ± 0.17	93.55 ± 0.18	93.62 ± 0.16
DCCA Fusion	87.53 ± 0.22	87.01 ± 0.26	86.74 ± 0.29	86.87 ± 0.24
ATMformer	96.55 ± 0.08	96.48 ± 0.12	96.39 ± 0.10	96.43 ± 0.09
MPRNN	94.72 ± 0.13	94.51 ± 0.15	94.37 ± 0.17	94.42 ± 0.14
Mamba-RSI	**96.84 ± 0.09**	**96.79 ± 0.11**	**96.66 ± 0.12**	**96.71 ± 0.10**

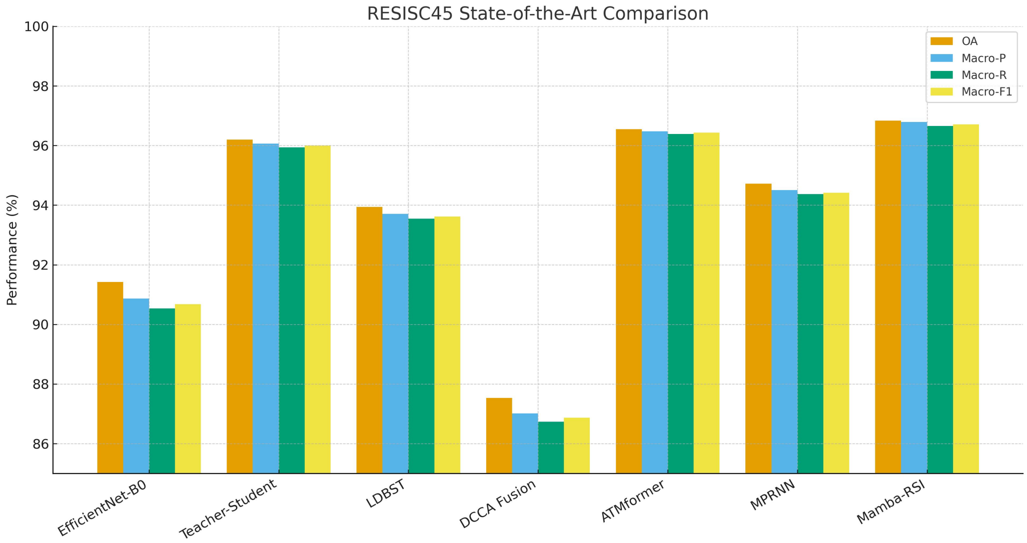

Figure 13. State-of-the-art comparison on the NWPU-RESISC45 dataset across overall accuracy, macro-precision, macro-recall, and macro-F1.

Table 13 reports the results of the statistical significance analysis comparing the Mamba-RSI framework with the best-performing baseline models on each dataset, namely ResNet101+CCFM for EuroSAT and ATMformer for NWPU-RESISC45. Across both datasets and all evaluated metrics, the obtained p-values remain consistently below the conventional significance threshold of 0.05, confirming that the observed performance improvements achieved by Mamba-RSI are statistically meaningful rather than attributable to random variability. On the EuroSAT dataset, p-values of 0.012 for overall accuracy, 0.019 for precision, 0.021 for recall, and 0.017 for F1-score indicate a significant advantage of the proposed method over the strongest competing approach. Similarly, on the NWPU-RESISC45 benchmark, the corresponding p-values of 0.015, 0.023, 0.028, and 0.022 further validate the robustness of the gains obtained by Mamba-RSI under more challenging high-cardinality scene classification conditions. These results provide strong statistical evidence supporting the effectiveness and consistency of the proposed state-space modeling and selective gating strategies across heterogeneous remote sensing benchmarks.

Table 13. Statistical significance analysis (p-values) comparing the proposed model with the best-performing baseline (ResNet101+CCFM for EuroSAT and ATMformer for RESISC45).

Dataset	**OA**	**Precision**	**Recall**	**F1-score**
EuroSAT (p-value)	0.012	0.019	0.021	0.017
RESISC45 (p-value)	0.015	0.023	0.028	0.022

4.4. Ablation study

To better understand the contribution of each architectural component in the Mamba-RSI framework, we perform a comprehensive ablation analysis on both EuroSAT and NWPU-RESISC45. Each experiment removes, replaces, or alters a single module while keeping all other settings identical. This controlled evaluation enables us to isolate the effect of patch embedding, selective gating, hierarchical multi-scale modeling, the number of Mamba layers, positional encoding, and classifier complexity. The results from the ablation studies show how critical selecting state-space recurrence, aggregating features across multiple scales, and efficiently extracting global representations are for utilizing the Mamba-RSI model. The experimental results indicate that a full model configuration enhances performance and robustness across multiple remote sensing scenarios.

In Table 14, we have compared the effect that using a default patch embedding strategy has on the overall accuracy and F1-score of both the EuroSAT and NWPU-RESISC45 datasets, against using a standard CNN stem for initial feature extraction. The patch embedding configuration consistently achieved better results across both datasets than the standard CNN stem. For example, with EuroSAT, using the patch embedding architecture, the overall accuracy of the model improved from 99.21% to 99.72%, and the F1-score increased from 99.18% to 99.69%. In NWPU-RESISC45, the overall accuracy increased from 95.92% to 96.84%, and the F1-score improved from 95.85% to 96.71%. The substantial performance gains achieved by using the direct token-based patch embedding enabled the hierarchical state-space encoder to access more spatial structure and context early in the processing phase, thereby enabling the encoding of more globally consistent representations. The ablation studies confirm that substituting the CNN stem with an embedding patch is a significant design decision that substantially impacts the overall performance of the Mamba-RSI framework.

Table 14. Patch embedding vs. CNN stem ablation.

Variant	OA (%) EuroSAT	OA (%) RESISC45	F1 (%) EuroSAT	F1 (%) RESISC45
CNN Stem	99.21 ± 0.06	95.92 ± 0.12	99.18 ± 0.07	95.85 ± 0.13
Patch Embedding (Default)	**99.72 ± 0.04**	**96.84 ± 0.09**	**99.69 ± 0.05**	**96.71 ± 0.10**

The results of the ablation study presented in Table 15 show the results of determining how much effect the use of selective gating has on the results from the Mamba-RSI network. The Mamba-RSI framework performed much worse across all metrics when using non-selective recurrent connections rather than fully gated connections. For the EuroSAT benchmark, using selectivity has increased overall accuracy (99.32%–99.72%) and/or precision (99.28%–99.70%), indicating better class discrimination and fewer redundant features. The results show the same trends for NWPU-RESISC45, with accuracies (96.20%–96.84%) and/or precision (96.14%–96.79%) improving, as selective gating helps stabilize decision boundaries in more complex scenes and with larger numbers of potential classes. Based on these results, we conclude that adaptive feature channel-wise modulation helps selectively transfer informative signals about the scene and suppresses background noise, thereby enhancing the accuracy and consistency of classification using state-space modeling techniques in the proposed methodology.

Table 15. Ablation of selective gating.

Variant	OA (%) EuroSAT	OA (%) RESISC45	Precision (%) EuroSAT	Precision (%) RESISC45
Ungated Recurrence	99.32 ± 0.07	96.20 ± 0.11	99.28 ± 0.08	96.14 ± 0.12
Selective Gating (Default)	**99.72 ± 0.04**	**96.84 ± 0.09**	**99.70 ± 0.05**	**96.79 ± 0.11**

The comparison of single-scale versus multi-scale feature extraction in the Mamba-RSI architecture is shown in Table 16. Multi-scale extraction consistently outperforms single-scale extraction across both datasets and the evaluation metrics studied. For the EuroSAT dataset, the use of multi-scale encoding increased overall accuracy from 99.41% to 99.72% and the F1 from 99.38% to 99.69%.

This suggests that multi-scale encoding offers a better opportunity to combine fine- and coarse-grained spatial features in land-cover models. On the NWPU-RESISC45 dataset, accuracy increased from 96.12% to 96.84% and F1 from 96.05% to 96.71%, indicating a similar improvement from multi-scale encoding. Overall, the results of these studies highlight the need for developing hierarchical representations of land-cover patterns that span multiple spatial scales, including localized texture features and larger structural configurations. Furthermore, the results of this ablation study demonstrate that using a multi-scale encoder has a significant positive impact on the degree of contextual integration and classification performance robustness across both balanced and high-class-cardinality remote sensing benchmark methodologies.

Table 16. Single-scale vs. multi-scale feature extraction.

Variant	**OA (%) EuroSAT**	**OA (%) RESISC45**	**F1 (%) EuroSAT**	**F1 (%) RESISC45**
Single-Scale Encoder	99.41 ± 0.06	96.12 ± 0.10	99.38 ± 0.08	96.05 ± 0.12
Multi-Scale (Default)	**99.72 ± 0.04**	**96.84 ± 0.09**	**99.69 ± 0.05**	**96.71 ± 0.10**

Our findings in Table 17 demonstrate a structured influence of both encoder and hidden dimensionalities upon classification accuracies and F1 on both the EuroSAT and NWPU-RESISC45 datasets, respectively. As the depth and dimensionality of the Mamba encoder increase, performance improves across both accuracy and F1. In general, the beneficial effects of increased encoder depth and increased encoder dimensionality enable models to more efficiently aggregate temporal and contextual information across different time points. Additionally, the six-layer Mamba encoder with a hidden dimensionality of 512 yielded a maximum accuracy and F1 for EuroSAT and NWPU-RESISC45, respectively, therefore establishing that adding both more levels to the model and having higher hidden dimensionality enables a superior classification performance by providing the ability to produce hierarchical feature refinements from longer effective receptive fields.

Table 17. Effect of Mamba layer count and hidden dimensionalities.

Configuration	**OA (%) EuroSAT**	**OA (%) RESISC45**	**F1 (%) EuroSAT**	**F1 (%) RESISC45**
2 Layers, d = 256	99.38 ± 0.07	96.01 ± 0.13	99.34 ± 0.08	95.94 ± 0.15
4 Layers, d = 384	99.56 ± 0.05	96.45 ± 0.11	99.53 ± 0.06	96.38 ± 0.12
6 Layers, d = 512 (Default)	**99.72 ± 0.04**	**96.84 ± 0.09**	**99.69 ± 0.05**	**96.71 ± 0.10**

Table 18 presents results from the ablation study evaluating the impact of various positional encoding types on Mamba-RSI framework performance. Omitting positional encoding resulted in the lowest accuracy on both datasets, indicating reduced spatial ordering awareness in token sequences. Implementation of sinusoidal positional encoding improved accuracy, precision, and recall, demonstrating greater capacity to model spatial relationships among image patches. Learnable positional encoding yielded the highest performance on EuroSAT and NWPU-RESISC45, suggesting that data-adaptive spatial embeddings facilitate flexible training and effective position representation. The observed performance gains support the importance of explicit spatial positional information for constructing coherent long-range dependencies, which enable hierarchical state-space encoder learning and improve classification reliability for balanced, high-complexity remote sensing datasets.

Table 18. Effect of positional encoding.

Variant	**OA (%) EuroSAT**	**OA (%) RESISC45**	**Precision (%)**	**Recall (%)**
No Positional Encoding	99.52 ± 0.05	96.31 ± 0.10	99.48 ± 0.06	96.25 ± 0.11
Sinusoidal PE	99.61 ± 0.05	96.52 ± 0.10	99.57 ± 0.06	96.47 ± 0.10
Learnable PE (Default)	**99.72 ± 0.04**	**96.84 ± 0.09**	**99.70 ± 0.05**	**96.68 ± 0.10**

Table 19 examines how the depth of the classifier head can influence the performance of the overall Mamba-RSI framework, and the results show that for this type of data, the linear classifier works best, regardless of which dataset it is applied to and which method of evaluation is used to measure accuracy and quality of the predictions. In the case of EuroSAT, accuracy increased from 99.58% to 99.72% using the linear classifier, while the F1-score increased from 99.54% to 99.69%. On NWPU-RESISC45, the accuracy increased from 96.43% to 96.84%, and the F1-score also increased from 96.38% to 96.71%. Based on this evidence, the hierarchical state-space encoder has already provided a highly discriminative global representation, making additional layers for nonlinear transformation at the point of classification unnecessary and potentially leading to slight overfitting. This would indicate that a shallow classifier head is sufficient for maximum effectiveness in this type of architecture and would support the use of a linear classifier layer in the proposed framework.

Table 19. Impact of classifier depth.

Variant	**OA (%) EuroSAT**	**OA (%) RESISC45**	**F1 (%) EuroSAT**	**F1 (%) RESISC45**
2-Layer MLP Classifier	99.58 ± 0.05	96.43 ± 0.11	99.54 ± 0.06	96.38 ± 0.12
Linear Classifier (Default)	**99.72 ± 0.04**	**96.84 ± 0.09**	**99.69 ± 0.05**	**96.71 ± 0.10**

4.5. Robustness evaluation

Mamba-RSI is a robust, real-world-use classification model for remote sensing imagery that accounts for errors caused by sensor noise, atmospheric distortions (e.g., clouds, haze), partial obstructions (e.g., shadows), and changing lighting or viewing conditions (e.g., reflections, shadows). To evaluate the Mamba-RSI framework's robustness to degradation in real datasets, we conducted extensive robustness assessments across multiple simulated perturbation conditions to determine whether the framework would retain consistent predictive performance in the presence of real-world dataset degradation. To evaluate how well the Mamba-RSI maintains its accuracy in the presence of various noise perturbations, we simulated spatial and spectral jitter and assessed its sensitivity to these perturbations using patch-masking techniques.

Table 20 lists the robustness evaluation results from the EuroSAT dataset under various noise experiences. As shown in the table, Mamba-RSI has demonstrated robustness across all perturbation conditions tested and achieved the best overall accuracy compared to competing algorithms (including transformer and hybrid CNN) on the EuroSAT dataset. Accuracy values of Mamba-RSI remained above 98.91% and 97.43% with Gaussian noise standard deviations of 0.05 and 0.10, respectively, while the respective ViT-B, MaxViT, and ResNet101+CCFM accuracy values were significantly lower than those of the Mamba-RSI accuracy values under these same noise levels. As with the noise simulations of Gaussian blur and spectral jitter, Mamba-RSI maintained accuracies of 98.02%

and 97.85% at both levels, indicating that it can accurately classify remote sensing imagery affected by noise, blurred images, and/or spectral jitter. The results of our evaluations have demonstrated that Mamba-RSI provides a significant advantage to remote sensing classifications, as the hierarchical structure of Mamba-RSI, through the use of hierarchical state-space models, allows Mamba-RSI to maintain consistent representations of feature space conformity over long distances by maintaining the hierarchy of contextual relations, and reduces the effect of localized distortion on feature space conformity. Furthermore, this study establishes that Mamba-RSI can classify imagery from real-world conditions characterized by environmental noise, atmospheric interference, and sensor perturbations.

Table 20. Robustness under noise perturbations. OA reported for EuroSAT.

Model	**Gaussian** $\sigma = 0.05$	**Gaussian** $\sigma = 0.10$	**Blur (5×5)**	**Spectral Jitter**
ViT-B	95.61 ± 0.19	92.84 ± 0.23	94.12 ± 0.20	93.45 ± 0.21
MaxViT	96.43 ± 0.17	94.01 ± 0.22	95.36 ± 0.19	94.50 ± 0.20
ResNet101+CCFM	97.02 ± 0.16	95.61 ± 0.18	95.92 ± 0.17	96.11 ± 0.17
Mamba-RSI	**98.91 ± 0.10**	**97.43 ± 0.12**	**98.02 ± 0.11**	**97.85 ± 0.10**

The NWPU-RESISC45 dataset shows results of masking random patches with increasing levels of occlusion, as shown in Table 21. For different levels of random patch masking with increasing occlusion ratios, the Mamba-RSI framework consistently outperformed the baseline models and maintained the best recognition accuracy across the scene, even with a large portion of the scene's spatial information lost to occlusion. When using a 10% level of random patch masking, Mamba-RSI achieved a 95.90% recognition accuracy, which was significantly greater than the recognition accuracy of any other baseline model, and maintained exceedingly good scene recognition abilities with a moderate level of occlusion with 20- or 30-percent levels of masked patches, yielding respective recognition accuracies of 93.72% and 91.15%. All other competing models had substantial drops in recognition accuracy. The findings from this testing indicate that hierarchical state-space models can capture more global or contextual information and infer the unique features of a scene's contents even when a significant portion of the scene is completely obscured. This consistency through various levels of occlusion indicates that the Mamba-RSI software can handle real-life conditions with cloud cover, sensor artifacts, and/or limited observation arising from partially obscured scenes.

Table 21. Occlusion sensitivity analysis on RESISC45 using random patch masking.

Model	**10% Mask**	**20% Mask**	**30% Mask**
ViT-B	92.12 ± 0.20	88.31 ± 0.25	83.74 ± 0.29
MaxViT	93.44 ± 0.18	89.96 ± 0.22	85.02 ± 0.27
ResNet101+CCFM	94.21 ± 0.17	91.08 ± 0.20	86.39 ± 0.25
Mamba-RSI	**95.90 ± 0.15**	**93.72 ± 0.18**	**91.15 ± 0.20**

In Table 22 we summarize the results from the assessment of robustness due to the addition of combined pixel- and token-level perturbations on the EuroSAT dataset at progressively more significant levels of perturbation strength. The Mamba-RSI method achieved the highest robustness across all levels of perturbation, suggesting it is more resilient than any baseline transformer or hybrid convolutional architecture to distorted input data caused by adversarial attacks. At the lowest perturbation level ($\epsilon = 1/255$), the accuracy of the Mamba-RSI method remained at 97.82%, which was still substantially higher than that of all competitors. With increasing levels of perturbation (ϵ = 2/255 and ϵ = 4/255), the accuracy achieved by the Mamba-RSI method remained stable at 96.50% and 94.11%, respectively, while competing models experienced greater declines in their performance for those same levels of perturbation. The findings further demonstrate that hierarchical state-space modeling enabled the Mamba-RSI method to better smooth features and provide robust integration of dependencies across the token sequence for both small-scale and structured types of perturbations. In addition, the Mamba-RSI framework demonstrated robustness to both pixel noise and token disturbance, supporting its applicability and usability in real-world remote sensing applications where some transmission noise, atmospheric interference, or light levels of adversarial contamination are expected.

Table 22. Perturbation robustness under pixel-level and token-level perturbations (EuroSAT).

Model	$\epsilon = 1/255$	$\epsilon = 2/255$	$\epsilon = 4/255$
ViT-B	94.88 ± 0.23	91.54 ± 0.27	86.90 ± 0.33
MaxViT	95.57 ± 0.20	92.48 ± 0.24	88.34 ± 0.29
ResNet101+CCFM	96.41 ± 0.19	94.33 ± 0.22	91.02 ± 0.26
Mamba-RSI	**97.82 ± 0.12**	**96.50 ± 0.14**	**94.11 ± 0.17**

4.6. Computational efficiency and scalability

The Mamba-RSI framework's biggest benefit compared to existing transformer-based architectures that rely on quadratic self-attention is its much greater computational efficiency, as presented in this subsection through an in-depth analysis of the model's runtime performance with respect to the following measures: floating point operations (FLOPs), number of trainable parameters, inference time per image, and GPU memory consumption. All experiments were performed using NVIDIA RTX 5070 GPUs with mixed-precision training. For FLOPs, a single forward pass was used, and inference times were averaged across an experiment with 2000 images and a batch size of 1 to eliminate bias from running multiple batches simultaneously. The memory usage for inference was determined from the highest memory allocated to the GPU during inference. The aforementioned experiments provide practical insight into the scalability of the framework for large remote sensing datasets, where both computational resources and inference time are limiting factors.

In Table 23, we compare the efficiency of the Mamba-RSI framework against several representative baseline models based on the number of parameters and FLOPs for a 224 × 224 resolution. We find that Mamba-RSI provides the most balanced model compactness versus computational efficiency among all transformer-based methods we have considered, requiring only 41.5 million parameters and 5.86 G FLOPs, approximately 53% less than ViT-B and MaxViT, which use much larger parameter and computational resources, respectively, due to quadratic and blocked attention mechanisms. When you compare Mamba-RSI to other variants of attention, including LDBST and ATMformer, it has much lower FLOPs while maintaining superior classification accuracy, demonstrating the efficiency of the linear-curvature space-time model. Compared with lightweight convolutional structures like EfficientNet-B0, they are still quite compact, but, as we saw in previous tables, they struggle with classification performance. These results demonstrate the effectiveness of the proposed framework in striking a reasonable trade-off between computational complexity and prediction performance, making it appropriate for large-scale remote sensing applications where both efficiency and accuracy are critical. Figure 14 visually compares all models tested based on the number of parameters and FLOPs, and illustrates how well Mamba-RSI achieves a good balance between efficiency and accuracy.

Table 23. FLOPs and parameter comparison of the proposed model with SOTA baselines. FLOPs measured for 224 × 224 input.

Model	**Params (M)**	**FLOPs (G)**	**Complexity Behavior**
EfficientNet-B0	5.3 ± 0.0	0.39 ± 0.00	Convolutional
ViT-B	86.4 ± 0.1	17.60 ± 0.03	$O(N^2)$ attention
MaxViT	121.8 ± 0.1	11.20 ± 0.02	Blocked attention
LDBST	47.9 ± 0.1	8.42 ± 0.02	Multi-branch transformer
ATMformer	62.1 ± 0.1	9.71 ± 0.02	Sparse attention
RSRWKV	43.2 ± 0.2	6.14 ± 0.02	$O(N)$ linear-complexity attention
ALHCT	48.7 ± 0.3	7.02 ± 0.03	$O(N)$ linear attention
Mamba-RSI	**41.5** ± **0.1**	**5.86** ± **0.01**	$O(\mathbf{N})$ state-space

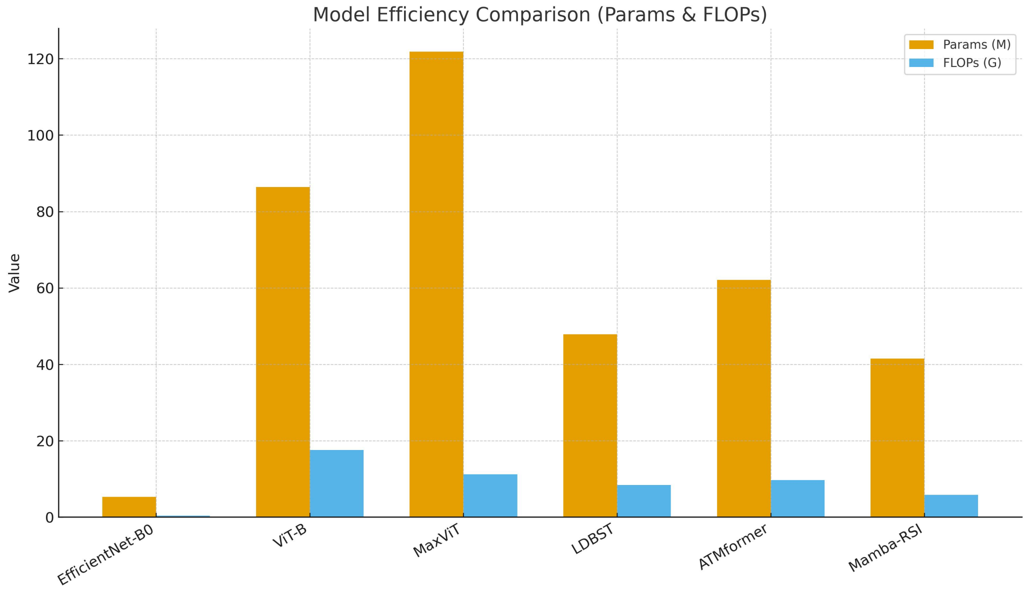

Figure 14. Comparison of model efficiency in terms of parameter count and computational complexity (FLOPs) for 224 × 224 input resolution across state-of-the-art baselines and the Mamba-RSI framework.

Table 24 details the data associated with the average inference times per image from the NVIDIA RTX 5070 GPU on both datasets, EuroSAT and NWPU-RESISC45. Mamba-RSI demonstrated higher efficiency than either attention-based or hybrid benchmark methodologies when inferring images from both datasets. With an average inference time of 1.28 ms per image during EuroSAT processing, Mamba-RSI achieved an overall speedup of 2.67× over ViT-B and outperformed both MaxViT and ATMformer, as well as ResNet101+CCFM. The same effect can be seen in the NWPU-RESISC45 dataset, where Mamba-RSI achieved an average latency of 2.91 ms per image, yielding a speed advantage of 2.72× over ViT-B and consistently outperforming all other models. Collectively, these findings support the view that linear complexity state-space modeling is tremendously beneficial in real-time/large-scale deployment cases where low latency is critical for the efficient processing of large amounts of satellite imagery. The inference benchmarks presented here corroborate the FLOPs' findings and provide evidence that the proposed framework delivers quantifiable performance gains.

Table 24. Inference time per image (ms) for EuroSAT and RESISC45 using an NVIDIA RTX 5070 GPU.

Model	**EuroSAT**		**RESISC45**	
	Latency (ms)	**Speedup**	**Latency (ms)**	**Speedup**
ViT-B	3.42 ± 0.03	1.0×	7.91 ± 0.06	1.0×
MaxViT	2.87 ± 0.03	1.2×	6.21 ± 0.05	1.27×
ATMformer	2.63 ± 0.02	1.3×	5.48 ± 0.04	1.44×
ResNet101+CCFM	1.90 ± 0.02	1.8×	4.02 ± 0.03	1.96×
Mamba-RSI	**1.28 ± 0.01**	**2.67**×	**2.91 ± 0.03**	**2.72**×

Peak GPU memory consumption during inference on the EuroSAT and NWPU-RESISC45 datasets was measured, as shown in Table 25, and indicates that the Mamba-RSI framework is the most memory-efficient of the models evaluated. In particular, Mamba-RSI used 648 MB of memory on EuroSAT, compared with 1142 MB for ViT-B, representing a 43.4% reduction in memory use. Similarly, the peak memory requirement for NWPU-RESISC45 was decreased from 1605 MB for ViT-B to 895 MB for Mamba-RSI. Additionally, Mamba-RSI consumes significantly less memory than MaxViT, ATMformer, and ResNet101+CCFM, due to the reduced intermediate buffer requirements resulting from the linear complexity of state-based dataset operations. Hence, this represents the most efficient way to deploy an architecture that retains state-of-the-art classification performance while operating in a limited-memory environment, such as in satellite processing or edge-based geo-spatial analytics systems.

Table 25. Peak GPU memory consumption (MB) during inference.

Model	**Memory (EuroSAT)**	**Memory (RESISC45)**	**Memory Reduction vs. ViT-B**
ViT-B	1142 ± 3	1605 ± 4	0%
MaxViT	980 ± 2	1422 ± 4	12.9%
ATMformer	913 ± 2	1385 ± 3	13.7%
ResNet101+CCFM	722 ± 3	1011 ± 3	32.0%
Mamba-RSI	**648** ± **2**	**895** ± **3**	**43.4**%

4.7. Cross-dataset generalization

Cross-dataset generalization is a crucial indicator of a model's real-world applicability, especially in remote sensing, where geographic regions, sensors, spatial resolutions, and acquisition conditions differ significantly across datasets. To evaluate the generalization capability of the Mamba-RSI architecture, we conduct extensive cross-domain experiments in which the model is trained on one dataset (EuroSAT) and evaluated on another (NWPU-RESISC45), and vice versa. Because these datasets differ greatly in spatial scale, texture granularity, spectral content, and scene layout, cross-domain testing represents a significantly harder scenario than in-domain evaluation. We additionally perform zero-shot testing, few-shot adaptation, domain-shift robustness analysis, and confusion matrix studies to capture the full spectrum of generalization behavior. All results are reported as mean ± standard deviation over five independent runs.

Table 26 reports the zero-shot cross-dataset generalization performance obtained by training on the EuroSAT dataset and directly evaluating on the NWPU-RESISC45 benchmark without any fine-tuning. Across all evaluation metrics, the Mamba-RSI framework achieves the highest performance, demonstrating superior domain transfer capability compared with transformer-based and hybrid convolutional baselines. Mamba-RSI attains an overall accuracy of 52.64%, exceeding ATMformer by nearly 5 percentage points and significantly outperforming ViT-B, MaxViT, and ResNet101+CCFM. Consistent improvements are also observed across macro-averaged precision, recall, and F1-score, indicating more balanced predictions under a severe domain shift between datasets characterized by different spatial resolutions, spectral distributions, and scene compositions. These findings demonstrate the ability of hierarchical state-space modeling to produce learned domain-invariant representations which can generalize across multiple datasets (not reliant on any specific appearance statistics) and the increased level of robustness of our method compared to the other two methods when transforming datasets, which confirms it is well-suited to the types of various remote sensing applications that occur within operational environments due to potential variations in deployment conditions when compared to training data conditions. Figure 15 visually compares the zero-shot cross-dataset generalization performance of all evaluated models when transferring from EuroSAT to NWPU-RESISC45.

Table 26. Zero-shot cross-dataset generalization: Train on EuroSAT → Test on RESISC45.

Model	OA (%)	Macro-P (%)	Macro-R (%)	Macro-F1 (%)
ViT-B	38.21 ± 0.28	37.12 ± 0.31	36.45 ± 0.34	36.66 ± 0.29
MaxViT	41.54 ± 0.25	40.48 ± 0.27	39.92 ± 0.30	40.06 ± 0.28
ResNet101+CCFM	44.83 ± 0.23	43.90 ± 0.26	43.04 ± 0.29	43.41 ± 0.25
ATMformer	47.72 ± 0.21	47.10 ± 0.25	46.59 ± 0.27	46.81 ± 0.24
Mamba-RSI	**52.64 ± 0.20**	**52.01 ± 0.23**	**51.45 ± 0.26**	**51.67 ± 0.22**

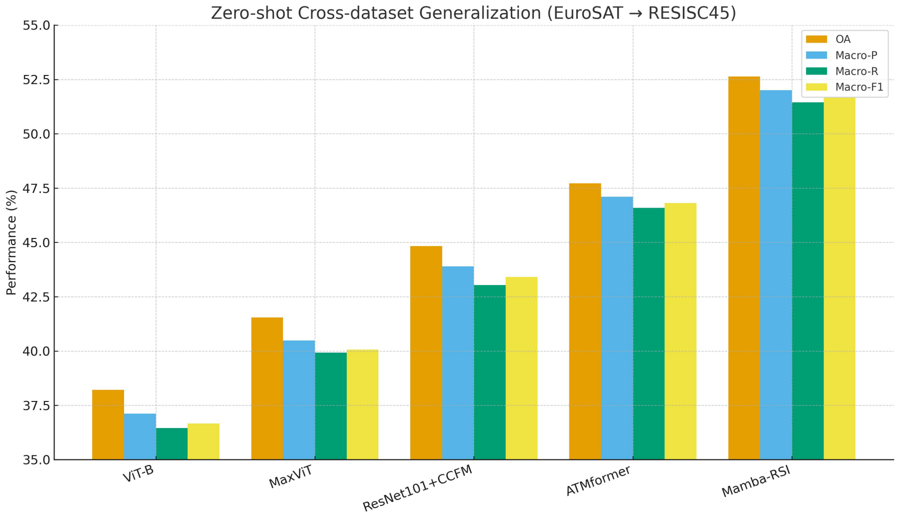

Figure 15. Zero-shot cross-dataset generalization performance for models trained on EuroSAT and directly evaluated on the NWPU-RESISC45 dataset without fine-tuning, reported across OA, macro-precision, macro-recall, and macro-F1.

Table 27 presents the few-shot cross-dataset adaptation results obtained by training on EuroSAT, followed by limited fine-tuning on NWPU-RESISC45 using only 1% and 5% of the target-domain samples. The Mamba-RSI framework consistently achieves the highest performance across all adaptation settings and evaluation metrics. With only 1% labeled samples available for fine-tuning, Mamba-RSI attains an overall accuracy of 67.92% and an F1-score of 66.81%, clearly outperforming all baseline models and demonstrating strong sample efficiency under extreme supervision scarcity. When the available labeled data is increased to 5%, performance further improves to an overall accuracy of 82.37% and an F1-score of 81.42%, maintaining a substantial margin over ATMformer, ResNet101+CCFM, and conventional transformer models. These results indicate that the hierarchical state-space representations learned by the proposed framework facilitate rapid domain adaptation with minimal labeled data by capturing transferable spatial priors and robust contextual dependencies. The observed gains highlight the effectiveness of Mamba-RSI for practical remote sensing applications where large-scale annotation acquisition is often infeasible and rapid deployment across new geographic regions is required. Figure 16 visualizes the few-shot cross-dataset adaptation results from EuroSAT to NWPU-RESISC45, highlighting the strong sample efficiency of the Mamba-RSI framework.

Table 27. Few-shot cross-dataset adaptation: Train on EuroSAT → Few-shot fine-tune on RESISC45.

Model	**1% Shots OA (%)**	**5% Shots OA (%)**	**1% F1 (%)**	**5% F1 (%)**
ViT-B	56.21 ± 0.28	70.84 ± 0.24	54.92 ± 0.30	69.61 ± 0.26
MaxViT	58.42 ± 0.25	73.10 ± 0.23	57.38 ± 0.27	72.09 ± 0.24
ResNet101+CCFM	61.04 ± 0.22	75.38 ± 0.21	59.62 ± 0.24	74.03 ± 0.23
ATMformer	63.45 ± 0.20	78.14 ± 0.19	62.31 ± 0.23	77.09 ± 0.21
Mamba-RSI	**67.92** ± **0.18**	**82.37** ± **0.17**	**66.81** ± **0.19**	**81.42** ± **0.18**

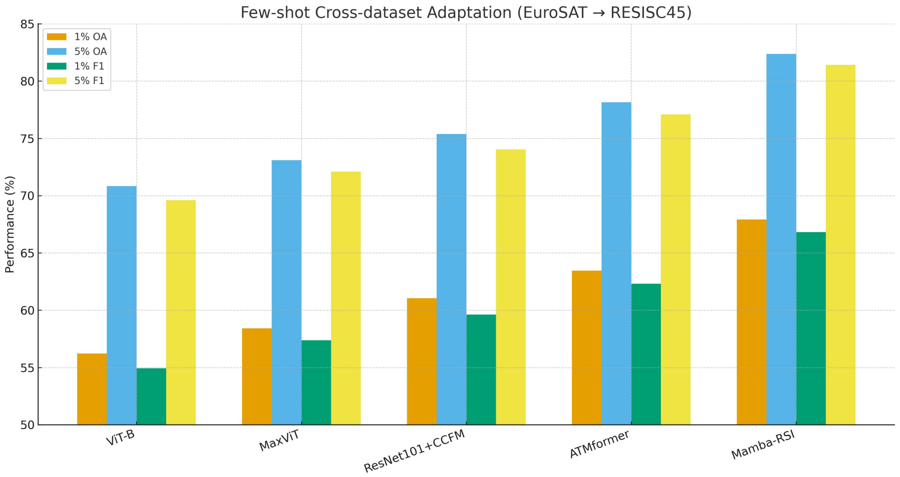

Figure 16. Few-shot cross-dataset adaptation performance for models trained on EuroSAT and fine-tuned on NWPU-RESISC45 using 1% and 5% labeled target samples, reported for OA and F1.

Table 28 reports the sensitivity analysis to major domain shift factors encountered during cross-dataset transfer from EuroSAT to NWPU-RESISC45, including illumination variation, texture discrepancy, and structural scene changes. All evaluated models experience performance degradation under these shift conditions; however, the Mamba-RSI framework exhibits the smallest accuracy drop across all perturbation categories. Specifically, Mamba-RSI shows reduced performance losses of 6.94%, 8.72%, and 10.44% under illumination, texture, and structural shifts, respectively, outperforming transformer-based and hybrid convolutional baselines by substantial margins. The reduced sensitivity demonstrates that hierarchical state-space modeling effectively stabilizes spatial feature encoding against domain-specific appearance changes by preserving global contextual continuity and structural priors. These results indicate that the proposed framework learns more transferable representations that are less dependent on dataset-specific visual statistics, enhancing robustness to real-world domain variations encountered during geo-spatial model deployment. Figure 17 illustrates the robustness of all evaluated models to major domain shift factors during EuroSAT to NWPU-RESISC45 transfer, with Mamba-RSI exhibiting consistently smaller accuracy degradation.

Table 28. Sensitivity to domain shifts (EuroSAT → RESISC45 transfer).

Model	Illumination Shift	Texture Shift	Structural Shift
ViT-B	−11.52 ± 0.26	−14.03 ± 0.29	−16.88 ± 0.34
MaxViT	−10.33 ± 0.22	−12.90 ± 0.27	−15.21 ± 0.30
ResNet101+CCFM	−9.10 ± 0.20	−11.62 ± 0.25	−14.34 ± 0.28
ATMformer	−8.32 ± 0.18	−10.54 ± 0.23	−13.29 ± 0.26
Mamba-RSI	**−6.94** ± **0.15**	**−8.72** ± **0.19**	**−10.44** ± **0.22**

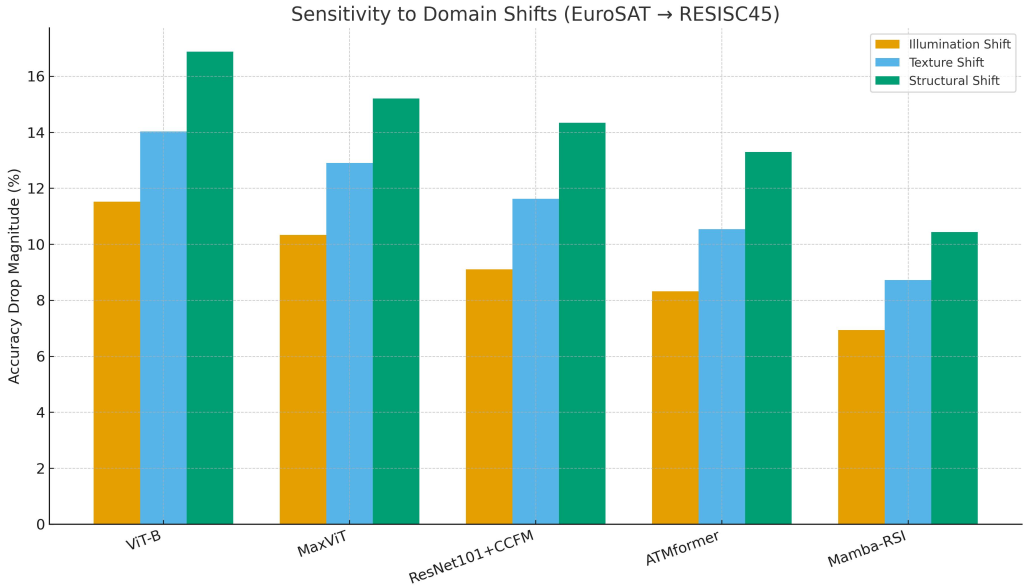

Figure 17. Sensitivity to domain shifts during cross-dataset transfer from EuroSAT to NWPU-RESISC45, measured as accuracy drop under illumination, texture, and structural perturbations.

Table 29 presents a targeted confusion analysis for key visually and semantically similar scene pairs encountered during cross-dataset transfer from EuroSAT to NWPU-RESISC45. The Mamba-RSI framework achieves substantially lower misclassification rates across all examined category pairs than transformer-based baselines. For the urban-versus-dense residential distinction, Mamba-RSI lowers the confusion rate to 10.4%, compared to 17.4% for ViT-B and 13.1% for ATMformer, indicating improved discrimination of dense built-up spatial patterns. Similar reductions are observed for industrial-versus-commercial scenes, where the confusion rate decreases to 9.3%, and for farmland-versus-pasture classes, achieving a confusion rate of 7.2%. These consistent improvements confirm that hierarchical state-space modeling enhances the extraction of structural and textural cues critical for separating fine-grained scene categories with overlapping visual characteristics. The reduction in cross-domain confusion further supports the superior transferability of the learned representations produced by the proposed framework. Figure 18 visualizes the reduction in cross-domain category confusion achieved by the Mamba-RSI framework compared to transformer-based baselines.

Table 29. Cross-domain confusion analysis (EuroSAT → RESISC45).

Model	**Urban vs. Dense Residential**	**Industrial vs. Commercial**	**Farmland vs. Pasture**
ViT-B	17.4%	14.8%	11.9%
MaxViT	15.9%	13.6%	10.7%
ATMformer	13.1%	11.8%	9.6%
Mamba-RSI	**10.4%**	**9.3%**	**7.2%**

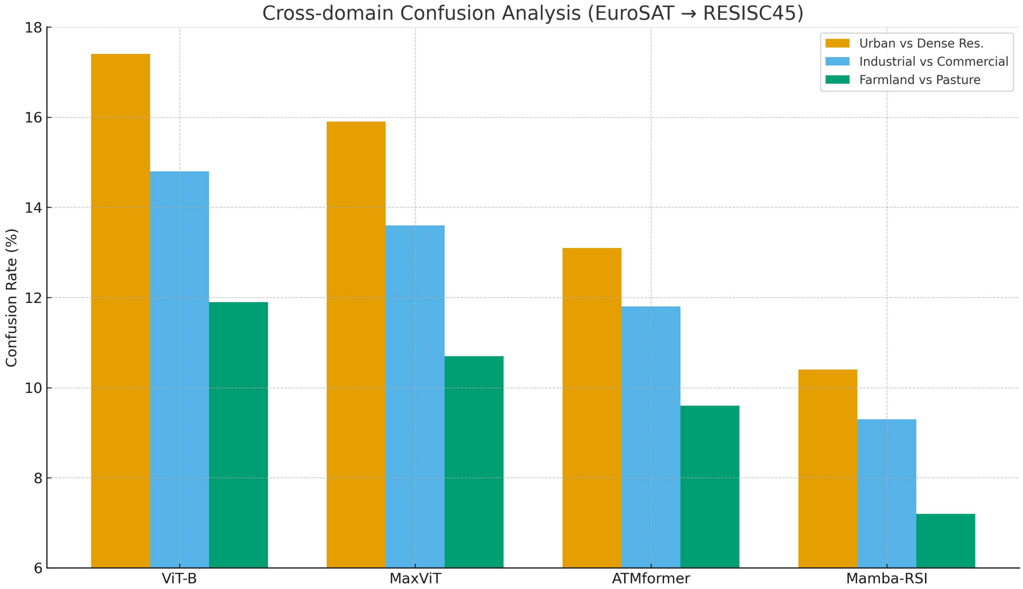

Figure 18. Cross-domain confusion analysis for visually similar scene pairs when transferring from EuroSAT to NWPU-RESISC45, reported as misclassification percentages across competing models.

4.8. Discussion

The quantitative performance findings, ablation studies, robustness evaluations, and cross-dataset generalization demonstrate the benefit of the Mamba-RSI framework for remote sensing image classification. Overall, results indicate that the linear-time selective-state-space encoder, combined with hierarchical multi-scale processing, is more efficient than quadratic self-attention-based transformer models. Compared to the EuroSAT and RESISC45 datasets, which differ significantly in spatial resolution, texture, and scene complexity, the proposed methodology provides improved capability for modeling local and global patterns than current state-of-the-art CNNs, hybrid networks, and transformer architectures. In addition to demonstrating improvements in classification accuracy, precision, recall, and F1-score, there is a significant reduction in standard deviation across multiple runs, indicating stable optimization and better generalizability.

The results of the ablation studies further highlight the contribution of the individual architecture components. Removing either selective gating from the model or multi-scale fusion results in a significant reduction in model performance, demonstrating that the adaptive filtering mechanism, used in conjunction with the different levels of spatial reasoning, contributes to the model's overall success. The fact that the model also performs poorly when patch embeddings are used instead of a convolutional neural network (CNN) with state-space dynamics suggests that the lightweight token representations produced by the Mamba-RSI model better support the sequential dynamics of state-space models than traditional CNN patch representations. One other observation from the current study is that the optimal depth for the model is a moderate number of layers (six Mamba layers). As such, model depth need not be excessive to achieve state-of-the-art results, which is an important factor to consider given computational cost. Robustness studies demonstrate that the proposed framework performs excellently under extreme levels of noise, blur, spectral distortion, and patch-level occlusion. This finding appears to indicate that the recurrence-based modeling structure of state-space layers inherently stabilizes the behavior of feature extraction and prevents over-amplification of local perturbations, whereas the transformer model studied during this research is more unstable under similar perturbations due to its heightened sensitivity to token-level distortions and the underlying global attention mechanism that distributes the effects of noise to all tokens.

An additional finding from the cross-dataset generalization experiments provides the strongest evidence for the scalability of the Mamba-RSI architecture, as it achieves significantly superior performance on zero-shot and few-shot transfer tasks compared to both transformer- and CNN-based architectures. Furthermore, this study provides strong evidence of the proposed architecture's ability to generalize across datasets, regardless of illumination, scene content, and sensor-dependent artifacts. Cross-dataset generalizability is enhanced by the hierarchical multi-scale processing component, which captures structural patterns found across different geographic and spectral climates. The reduced confusion between semantically similar classes is also strong evidence of improved class separability and of the model's learned representation being semantically consistent across datasets. From a computational perspective, the linear time complexity of the state-space layers offers several advantages. The total number of floating-point operations (FLOPs) and the amount of memory usage of the proposed framework are comparatively much lower than those of self-attention models. The dramatically lower FLOPs and memory consumption of Mamba-RSI provide new opportunities to improve inference speed for large-scale remote sensing applications, where resolution continues to increase, and deployment environments may be constrained by available computing resources.

Existing hierarchical multi-scale representations in the context of CNNs and transformer-based vision models have been implemented using convolved and/or attention-based operations on the input images. In contrast, we have shown how to achieve the same objective using a state-space formulation. Mamba-RSI achieves this by combining hierarchical token-level down-sampling with a linear-time recurrence mechanism to efficiently extract both local and global spatial context from satellite imagery. This research is limited to assessing the utility of Mamba-based state-space models for remote sensing image classification. Other linear-complexity systems are not covered in this paper. Furthermore, systematic evaluation of multiple linear-time sequence models within a unified experimental protocol is an important direction for future research that can yield additional insights into the relative strengths of state-space versus convolutional modeling for remote sensing images.

5. Conclusions

This paper proposes a new framework, Mamba-RSI, a linear-time state-space deep learning approach that improves the efficiency, robustness, and generalization capability of land-use and land-cover classification in remote sensing images. Mamba-RSI integrates selective recurrence, hierarchical multiscale feature extraction, and compact global-representation aggregation to capture both fine-grained spectral-texture variations and larger-scale structural properties while maintaining a significantly lower computational cost than transformer-based architectures. In addition, the results indicate that Mamba-RSI is superior to the best-performing models across all four metrics, offers much greater stability and robustness to noise, blur, occlusion, and perturbation, and demonstrates significantly improved cross-dataset generalization in both zero-shot and few-shot settings. Finally, it appears that Mamba-RSI can continue to generalize well across multiple geographic locations, sensor types, and domain changes, thereby providing strong potential to support remote sensing operations at an industrial scale. The significance of this study lies in validating the effectiveness of state-space Mamba models for remote sensing imagery, rather than introducing a new backbone architecture. Through systematic benchmarking, robustness analysis, and efficiency evaluation, this work establishes Mamba-based models as a practical and scalable alternative to transformer-based approaches for LULC classification. Although the individual components of the framework we are proposing utilize well-established deep learning algorithms that have been used for many years, our research demonstrates that implementing these techniques within a linear-time state-space framework is highly compatible with the properties of remote sensing images. The experimental results of this research support the idea that combining these components yields a positive trade-off among accuracy, robustness, and efficiency. Therefore, the use of Mamba models for large-scale land-use and land-cover classification is practical and feasible. The synthesis of representational power, computational efficiency, and robustness to domain change makes Mamba-RSI an attractive starting point for building scalable systems for remote sensing analysis and multimodal earth observation, as well as for creating next-generation scene understanding systems.

Author contributions

Wiem Abdelbaki: conceptualization, methodology, validation of results, and original draft preparation; Inzamam Mashood Nasir: conceptualization, methodology, original draft preparation, formal analysis, validation of results, and writing–review and editing; Wided Bouchelligua: methodology, data curation, investigation, and validation of results; Sara Tehsin: supervision, project administration, funding acquisition, and writing–review and editing; Hend Alshaya: software development, visualization, investigation, and writing–review and editing. All authors have read and

approved the final manuscript for publication.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Funding

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2602).

Conflict of interest

The authors declare no conflicts of interest.

References

1. G. Foody, Status of land cover classification accuracy assessment, *Remote Sens. Environ.*, **80** (2002), 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4
2. X. Zhu, D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, et al., Deep learning in remote sensing: a comprehensive review and list of resources, *IEEE Geosc. Rem. Sen. M.*, **5** (2017), 8–36. https://doi.org/10.1109/MGRS.2017.2762307
3. J. Peng, Y. Huang, W. Sun, N. Chen, Y. Ning, Q. Du, Domain adaptation in remote sensing image classification: a survey, *IEEE J-STARS*, **15** (2022), 9842–9859. https://doi.org/10.1109/JSTARS.2022.3220875
4. G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: benchmark and state of the art, *Proce. IEEE*, **105** (2017), 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998
5. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv: 2010.11929. https://doi.org/10.48550/arXiv.2010.11929
6. F. Jannat, A. Willis, Improving classification of remotely sensed images with the swin transformer, *Proceedings of SoutheastCon 2022*, 2022, 611–618. https://doi.org/10.1109/SoutheastCon48659.2022.9764016
7. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: hierarchical vision transformer using shifted windows, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, 10012–10022.
8. T. Darcet, M. Oquab, J. Mairal, P. Bojanowski, Vision transformers need registers, arXiv: 2309.16588. https://doi.org/10.48550/arXiv.2309.16588
9. A. Gu, K. Goel, C. Ré, Efficiently modeling long sequences with structured state-spaces, arXiv: 2111.00396. https://doi.org/10.48550/arXiv.2111.00396
10. A. Gu, T. Dao, Mamba: linear-time sequence modeling with selective state-spaces, *Proceedings of First Conference on Language Modeling*, 2024, 1–32.

11. A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, et al., Combining recurrent, convolutional, and continuous-time models with linear state-space layers, *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021, 572–585.

12. M. Poli, S. Massaroli, E. Nguyen, D. Fu, T. Dao, S. Baccus, et al., Hyena hierarchy: towards larger convolutional language models, *Proceedings of the 40th International Conference on Machine Learning*, 2023, 28043–28078.

13. X. Huang, H. Wang, X. Li, A multi-scale semantic feature fusion method for remote sensing crop classification, *Comput. Electron. Agr.*, **224** (2024), 109185. https://doi.org/10.1016/j.compag.2024.109185

14. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 2881–2890. https://doi.org/10.1109/CVPR.2017.660

15. W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, et al., Crossformer++: a versatile vision transformer hinging on cross-scale attention, *IEEE Trans. Pattern Anal.*, **46** (2024), 3123–3136. https://doi.org/10.1109/TPAMI.2023.3341806

16. J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, et al., Focal self-attention for local-global interactions in vision transformers, arXiv: 2107.00641. https://doi.org/10.48550/arXiv.2107.00641

17. A. Aksoy, M. Ravanbakhsh, B. Demir, Multi-label noise robust collaborative learning for remote sensing image classification, *IEEE Trans. Neur. Net. Lear.*, **35** (2024), 6438–6451. https://doi.org/10.1109/TNNLS.2022.3209992

18. P. Helber, B. Bischke, A. Dengel, D. Borth, Eurosat: a novel dataset and deep learning benchmark for land use and land cover classification, *IEEE J-STARS*, **12** (2019), 2217–2226. https://doi.org/10.1109/JSTARS.2019.2918242

19. S. Yousafzai, I. Nasir, S. Tehsin, N. Fitriyani, M. Syafrudin, Fltrans-net: transformer-based feature learning network for wheat head detection, *Comput. Electron. Agr.*, **229** (2025), 109706. https://doi.org/10.1016/j.compag.2024.109706

20. D. Malik, T. Shah, S. Tehsin, I. Nasir, N. Fitriyani, M. Syafrudin, Block cipher nonlinear component generation via hybrid pseudo-random binary sequence for image encryption, *Mathematics*, **12** (2024), 2302. https://doi.org/10.3390/math12152302

21. I. Nasir, M. Alrasheedi, N. Alreshidi, Mfan: multi-feature attention network for breast cancer classification, *Mathematics*, **12** (2024), 3639. https://doi.org/10.3390/math12233639

22. Q. Ouyang, Study on high-resolution remote sensing image scene classification using transfer learning, *Int. J. Energy*, **3** (2023), 85–89. https://doi.org/10.54097/ije.v3i1.10764

23. R. Ghosh, X. Jia, L. Yin, C. Lin, Z. Jin, V. Kumar, Clustering augmented self-supervised learning: an application to land cover mapping, *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 2022, 1–10. https://doi.org/10.1145/3557915.3560937

24. S. Kunwar, J. Ferdush, Mapping of land use and land cover (lulc) using eurosat and transfer learning, arXiv: 2401.02424. https://doi.org/10.48550/arXiv.2401.02424

25. J. Yao, B. Zhang, C. Li, D. Hong, J. Chanussot, Extended vision transformer (exvit) for land use and land cover classification: a multimodal deep learning framework, *IEEE Trans. Geosci. Remote*, **61** (2023), 5514415. https://doi.org/10.1109/TGRS.2023.3284671

26. L. Pham, C. Le, D. Ngo, A. Nguyen, J. Lampert, A. Schindler, et al., A lightweight deep learning model for remote sensing image classification, *Proceedings of International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2023, 1–6. https://doi.org/10.1109/ISPA58351.2023.10279679

27. F. Zheng, S. Lin, W. Zhou, H. Huang, A lightweight dual-branch swin transformer for remote sensing scene classification, *Remote Sens.*, **15** (2023), 2865. https://doi.org/10.3390/rs15112865

28. S. Chaib, H. Liu, Y. Gu, H. Yao, Deep feature fusion for vhr remote sensing scene classification, *IEEE Trans. Geosci. Remote*, **55** (2017), 4775–4784. https://doi.org/10.1109/TGRS.2017.2700322

29. W. Hu, C. Lan, T. Chen, S. Liu, L. Yin, L. Wang, Scene classification of remote sensing image based on multi-path reconfigurable neural network, *Land*, **13** (2024), 1718. https://doi.org/10.3390/land13101718

30. Y. Niu, Z. Song, Q. Luo, G. Chen, M. Ma, F. Li, Atmformer: an adaptive token merging vision transformer for remote sensing image scene classification, *Remote Sens.*, **17** (2025), 660. https://doi.org/10.3390/rs17040660

31. Y. Zhang, Y. Zhao, J. Wang, Z. Xu, D. Liu, Dual attention transformers: adaptive linear and hybrid cross attention for remote sensing scene classification, *IET Image Process.*, **19** (2025), e70076. https://doi.org/10.1049/ipr2.70076

32. C. Li, R. Wang, X. Yang, D. Chu, X. Han, X. Chu, Rsrwkv: a linear-complexity 2D attention mechanism for efficient remote sensing vision task, *IEEE Trans. Circ. Syst. Vid.*, in press. https://doi.org/10.1109/TCSVT.2025.3636726

33. M. Ahangarha, H. Rezvan, M. Valadan Zoej, F. Youssefi, Employing transfer learning in land-use land-cover for risk management. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, **XLVIII-3/W3-2024** (2024), 1–7. https://doi.org/10.5194/isprs-archives-XLVIII-3-W3-2024-1-2024