remote sensing

MDPI

Article

MambaHSINet: A Dual-Branch Bidirectional State Space Network for Hyperspectral Tree Species Classification

Xinying Liu [1,†], Yanfeng Zhang [1,†], Junyang Wu [1], Tianyu Cai [1], Yumeng Li [1], Xinran Wang [2] and Xinwei Li [1,*]

[1] School of Sciences, Beijing Forestry University, Beijing 100083, China; liuxinying726@bjfu.edu.cn (X.L.); e578851062@bjfu.edu.cn (Y.Z.); wjyang@bjfu.edu.cn (J.W.); caitianyu@bjfu.edu.cn (T.C.); liyumeng@bjfu.edu.cn (Y.L.)

[2] School of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, China; wangxinran@bjfu.edu.cn

* Correspondence: lixinwei@bjfu.edu.cn; Tel.: +86-18810457603

† These authors contributed equally to this work.

Highlights

What are the main findings?

- A full-image input dual-branch bidirectional State Space Network, MambaHSINet, is proposed for hyperspectral tree species classification. This network synergistically extracts features through spectral-spatial dual branches and models long-range context with linear complexity. It outperforms existing mainstream approaches on the self-built tree species classification dataset, and achieves favorable adaptability and competitive performance on public datasets for crop classification and urban land-cover classification as well.
- We construct a new UAV-borne hyperspectral tree species dataset *ZJK*, and publicly release its two finely annotated subsets. These subsets cover typical forest-farmland transitional landscapes that contain complex spatial patterns and spectrally similar land-cover classes, serving as a challenging benchmark for fine-grained tree species classification research.

What are the implications of the main findings?

- The study indicates that Mamba-based State Space Models can efficiently model long-range spatial-spectral dependencies in hyperspectral images with linear computational complexity, providing a more efficient alternative to Transformers for hyperspectral image processing.
- Using the two publicly released finely annotated subsets of the ultra-high resolution UAV hyperspectral dataset *ZJK*, MambaHSINet achieves reliable performance for fine-grained tree species classification. Its competitive results on multiple public benchmarks further verify the strong generalization and cross-task applicability of the proposed model.

Abstract

Hyperspectral remote sensing provides rich spectral information and has been widely used in fine-grained land-cover classification and forest monitoring. However, accurate tree species classification remains challenging due to subtle interspecific spectral differences, similar spatial structures among related species, redundant spectral bands, and the limited ability of existing methods to model long-range spatial–spectral dependencies efficiently. In addition, many existing hyperspectral image classification methods rely on patch-based inputs and sliding-window inference, which often lead to redundant computation and insufficient utilization of global image context. To address these issues, this paper

Check for updates

Academic Editors: Jiangjun Peng, Xiangyong Cao, Shuang Xu, Jing Yao and Gemine Vivone

Received: 13 May 2026
Revised: 14 June 2026
Accepted: 10 July 2026
Published: 16 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

proposes MambaHSINet, a dual-branch bidirectional state space network for full-image pixel-wise hyperspectral classification. Specifically, the proposed network employs a spectral branch and a spatial branch to explicitly extract complementary spectral responses and spatial structural features. Subsequently, a bidirectional Mamba global modeling module based on selective state space modeling is adopted to capture long-range contextual dependencies in both forward and backward directions with linear computational complexity. Unlike conventional patch-based methods, MambaHSINet takes the entire hyperspectral image as input and produces full-resolution pixel-wise classification maps, thereby avoiding repeated cropping and redundant sliding-window inference. We also construct two well-annotated subsets of a UAV-borne hyperspectral dataset dedicated to tree species classification. Experimental results on self-collected and public hyperspectral datasets demonstrate that the proposed method achieves excellent classification accuracy, inference efficiency, and generalization performance. It exhibits great potential for practical tree species classification and other general hyperspectral application scenarios.

Keywords: hyperspectral image classification; tree species classification; state space model; Mamba

1. Introduction

Hyperspectral remote sensing has emerged as a cornerstone of modern Earth observation, offering unparalleled capabilities for forestry applications, such as forest structural inversion, tree species mapping, biomass estimation, and forest health monitoring [1–3]. By acquiring hundreds of contiguous narrow spectral bands, Hyperspectral image (HSI) provides a wealth of information that enables the identification of subtle land-cover variations. Owing to its fine spectral resolution, it can characterize minor reflectance differences that are imperceptible to conventional multispectral sensors, thereby providing a robust foundation for characterizing complex vegetation canopies and precise forest resource monitoring [4,5].

In the context of tree species classification, HSI presents distinct advantages as it captures unique spectral signatures dictated by leaf structure, pigment content, water status, and biochemical properties [6,7]. Compared with traditional field surveys, RGB-based visual interpretation and multispectral imagery, HSI offers enhanced separability for tree species identification within complex forest ecosystems [8–11]. Consequently, hyperspectral tree species classification has evolved into a pivotal research domain. Accurate tree species identification not only contributes to forest inventory and ecological assessment, but also provides critical data support for biodiversity conservation, carbon cycle modeling, and sustainable forest management.

The methodological evolution of hyperspectral image classification (HSIC) has progressed from traditional machine learning to deep learning models. Classical classifiers, such as support vector machine (SVM), random forest (RF), k-nearest neighbor (KNN), and sparse representation classification (SRC), generally rely on hand-crafted features and have limited capability in extracting deep discriminative representations [12–15]. Although these methods perform adequately on simple datasets, their generalization capability is usually constrained when addressing complex scenes and fine-grained class categories.

To overcome these limitations, Convolutional Neural Networks (CNNs), including 1D-, 2D-, and 3D-CNNs, have been extensively deployed to learn spectral, spatial, and spectral–spatial features [16–18]. In general, 1D-CNNs focus on spectral signatures extraction, 2D-CNNs mainly capture spatial texture information, and 3D-CNNs jointly

exploit spectral and spatial correlations. These methods have significantly improved HSIC performance compared with traditional machine learning approaches. However, CNNs are inherently limited by local receptive fields, rendering them less effective at modeling long-range dependencies, particularly in scenes with large homogeneous regions or intricate contextual structures.

To deal with the issue of local context, Transformer architectures have been introduced, leveraging self-attention mechanisms to capture global correlations [19–21]. By establishing pairwise relationships among tokens, Transformer-based models can effectively model long-range contextual dependencies and have achieved promising performance in HSIC. Nevertheless, their quadratic complexity, $O(N^2)$, results in high computational and memory costs for hyperspectral long-sequence data, which limits their efficiency and scalability in high-resolution hyperspectral classification tasks.

Recently, state space models (SSMs)-based methods, particularly Mamba, have revolutionized long-sequence modeling by achieving linear complexity, $O(N)$, while maintaining global modeling capabilities [22]. Given the sequential nature of HSI, Mamba-based models are well-suited for HSIC. Several studies have integrated Mamba with convolutions, spatial–spectral attention, or dual-path structures to leverage local and global features synergistically [23–27]. These methods demonstrate that Mamba can serve as an efficient alternative to self-attention for large-scale sequence modeling [28,29]. Nevertheless, the majority of existing Mamba-based methods are mainly designed for general land-cover classification, and their effectiveness for fine-grained tree species classification remains insufficiently explored.

Practical deployment of hyperspectral tree species classification faces several persistent challenges. First, closely related species often exhibit subtle spectral differences and similar spatial textures, making fine-grained discrimination extremely difficult. Second, existing methods usually lack explicit and efficient interaction between spectral and spatial features, leading to insufficient utilization of their complementarity [30]. Third, patch-based inference may introduce redundant computation and fail to fully exploit global contextual information [31]. In addition, ultra-high-resolution UAV-borne hyperspectral datasets for realistic fine-grained tree species classification remain limited, which also restricts the development and evaluation of advanced methods in this field [32,33].

To address these gaps, this study constructs a self-collected, manually annotated hyperspectral tree species dataset, named *ZJK*, and proposes a full-image pixel-wise classification network, MambaHSINet. The proposed network adopts a dual-branch architecture to separately extract spectral and spatial features, and employs a bidirectional Mamba module for efficient global contextual modeling. Unlike patch-based methods, MambaHSINet avoids repeated cropping and sliding-window inference while maximizing the utilization of global context. In this way, the proposed method aims to achieve a superior balance between classification accuracy and inference efficiency for practical hyperspectral tree species, as well as other HSIC tasks.

The main contributions of this study are summarized as follows:

1. An integrated bidirectional Mamba dual-branch network (MambaHSINet) with full-image end-to-end inference is proposed. It decouples spectral and spatial feature extraction to handle the challenge of spectrally similar land-cover classes. The bidirectional state space modeling realizes long-range dependency learning at linear complexity $O(N)$. Compared with conventional patch and sliding-window approaches, our scheme eliminates overlapping computation and enhances global context utilization. The combination of three core designs enables the model to excel in UAV-based fine-grained tree species classification and general hyperspectral tasks, with well-balanced accuracy and inference efficiency.

2. Two finely annotated subsets of the newly built *ZJK* UAV-borne hyperspectral tree species dataset are publicly released. These subsets cover typical forest-farmland transition scenes, which enrich labeled data for fine-grained forest classification and act as useful benchmarks for UAV forest monitoring.

The remainder of this paper is organized as follows. Section 2 introduces the related foundations. Section 3 describes the proposed MambaHSINet. Section 4 presents the experiments and analysis. Section 5 provides a detailed discussion of the experimental results and analyzed phenomena. Finally, Section 6 concludes this paper.

2. Related Work

This section provides a brief overview of algorithms related to Mamba. Regarding Mamba, we first introduce the State-Space Models (SSMs). Subsequently, Structured SSMs (S4) are introduced through discretization, followed by the addition of a selective mechanism to form S6. Finally, the Mamba model is constructed based on S6.

2.1. State-Space Models (SSMs)

State-space models (SSMs) are a class of systems that relate an input function $x(t) \in \mathbb{R}^L$ to an output response $y(t) \in \mathbb{R}^L$ through a set of latent state variables $\xi(t) \in \mathbb{R}^N$. Inspired by modern control theory, these models have recently been reintroduced to deep learning as an efficient alternative for sequence modeling. Generally, a continuous-time SSM is formulated as a linear first-order differential equation:

$$\dot{\xi}(t) = \mathbf{A}\xi(t) + \mathbf{B}x(t) \tag{1}$$

$$y(t) = \mathbf{C}\xi(t) + \mathbf{D}x(t) \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the evolution matrix that governs the system dynamics, $\mathbf{B} \in \mathbb{R}^{N\times L}$ and $\mathbf{C} \in \mathbb{R}^{L\times N}$ are the projection matrices, and $\mathbf{D} \in \mathbb{R}^{L\times L}$ represents the direct feedthrough component (usually skipped or handled as a residual connection in practical implementations). In HSI processing, this framework provides a rigorous mathematical basis for characterizing the spectral evolution of different tree species. By treating contiguous spectral bands as a continuous sequence, SSMs can capture the complex physical and biochemical signatures of forest canopies with linear computational complexity.

2.2. Structured SSMs (S4)

To facilitate the processing of discrete digital HSI data, the continuous system must be transformed into a discrete-time representation. By introducing a timescale parameter $\Delta \in \mathbb{R}^+$, the continuous matrices $(\mathbf{A}, \mathbf{B})$ are discretized into $(\overline{\mathbf{A}}, \overline{\mathbf{B}})$ via the zero-order hold (ZOH) mechanism:

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}) \tag{3}$$

$$\overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \tag{4}$$

The Structured SSM (S4) incorporates a specifically structured transition matrix (e.g., HiPPO initialization) to mitigate the vanishing gradient problem in long-range modeling. For HSI cubes with hundreds of spectral bands, S4 enables the extraction of global dependencies with $O(L \log L)$ computational complexity. This provides a significant efficiency advantage over Transformer-based architectures, which suffer from a quadratic $O(L^2)$ memory footprint when dealing with high-dimensional remote sensing data.

2.3. Selective Structured SSMs (S6)

Despite the efficiency of S4, its time-invariant nature means the transition parameters remain constant regardless of the input content, which limits its adaptability to complex land-cover types. The selective structured SSM (S6), as the core of the Mamba architecture, overcomes this by parameterizing **B**, **C**, and Δ as functions of the input x:

$$\mathbf{B} = \text{Linear}_B(x), \quad \mathbf{C} = \text{Linear}_C(x), \quad \Delta = \text{Softplus}(\text{Parameter} + \text{Linear}_\Delta(x)) \tag{5}$$

This content-aware selection mechanism allows MambaHSINet to adaptively focus on discriminative spectral features (e.g., the red-edge position or chlorophyll absorption pits) while suppressing redundant noise.Such a property is particularly vital for fine-grained tree species classification, as well as other HSIC tasks.

2.4. Mamba Block and Bidirectional Modeling

The Mamba block integrates the S6 engine with gated linear units and depth-wise convolutions to exploit both local and global features. However, traditional Mamba blocks are designed for causal sequences, which is suboptimal for HSI data where spatial-spectral correlations are inherently non-causal and multi-directional. In MambaHSINet, we implement a bidirectional modeling strategy to scan the spatial-spectral tokens from both forward and backward directions. This approach ensures that the contextual information of the forest canopy is captured holistically. By merging the efficiency of SSMs with bidirectional feature interaction, the Mamba block enables comprehensive feature representation of HSI, thus improving classification accuracy for complex forest ecosystems and diverse remote sensing scenarios.

3. Proposed Method

This section introduces the proposed MambaHSINet. We first present the overall architecture and then describe its main components in detail, including the spectral branch, spatial branch, dual-branch feature fusion, bidirectional Mamba global modeling module, classification head, and training strategy. The proposed method takes the full HSI as input and jointly performs local spectral–spatial feature extraction and global contextual dependency modeling with relatively low computational cost.

3.1. Overall Framework

As shown in Figure 1, the proposed MambaHSINet mainly consists of five parts: (1) input preprocessing, (2) spectral branch, (3) spatial branch, (4) feature fusion and bidirectional Mamba global modeling, and (5) classification head. Unlike traditional patch-based classification methods, the proposed framework directly takes the whole HSI as input and outputs pixel-wise classification results, thereby avoiding repeated cropping and redundant inference while making better use of full-image contextual information [34].

Let the input HSI be denoted as

$$\mathbf{X} \in \mathbb{R}^{H \times W \times B}, \tag{6}$$

where H and W denote the spatial height and width, respectively, and B is the number of spectral bands. First, layer normalization is applied along the spectral dimension to obtain the normalized feature

$$\mathbf{X}_n = \text{LN}(\mathbf{X}). \tag{7}$$

Then, $\mathbf{X}_n$ is fed into the spectral branch and the spatial branch to extract complementary spectral-domain and spatial-structure representations. The two features are concate-

nated and fused, rearranged into a sequence, and then sent into a bidirectional Mamba module for long-range contextual modeling. Finally, a classification head is used to output the category prediction for each pixel.

The whole framework can be formulated as

$$\mathbf{Y} = F_{\text{cls}}\Big(F_m\big(F_{\text{fus}}([F_{\text{spe}}(\mathbf{X}_n), F_{\text{spa}}(\mathbf{X}_n)])\big)\Big), \tag{8}$$

where $F_{\text{spe}}(\cdot)$, $F_{\text{spa}}(\cdot)$, $F_{\text{fus}}(\cdot)$, $F_m(\cdot)$, and $F_{\text{cls}}(\cdot)$ denote the spectral branch, spatial branch, fusion module, bidirectional Mamba module, and classification head, respectively; $[\cdot,\cdot]$ denotes channel-wise concatenation; and $\mathbf{Y}$ is the final pixel-wise classification output.

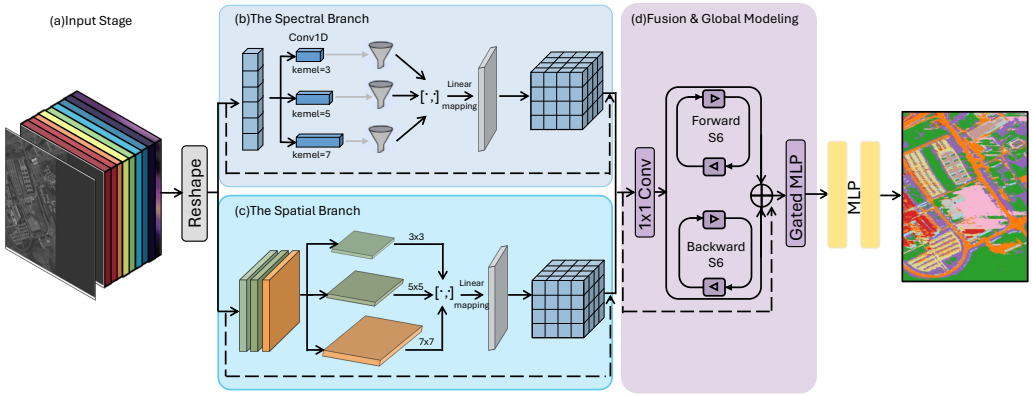


Figure 1. Overall Framework of the Proposed Method.

3.2. Spectral Branch

Each pixel in a HSI possesses a high-dimensional spectral response, where subtle but decisive inter-class variances are often embedded within the continuous spectral curves. To effectively mine these intra-pixel correlations while mitigating the risk of feature degradation, we design a lightweight multi-scale spectral branch with an explicit emphasis on original information preservation, as shown in Figure 2.

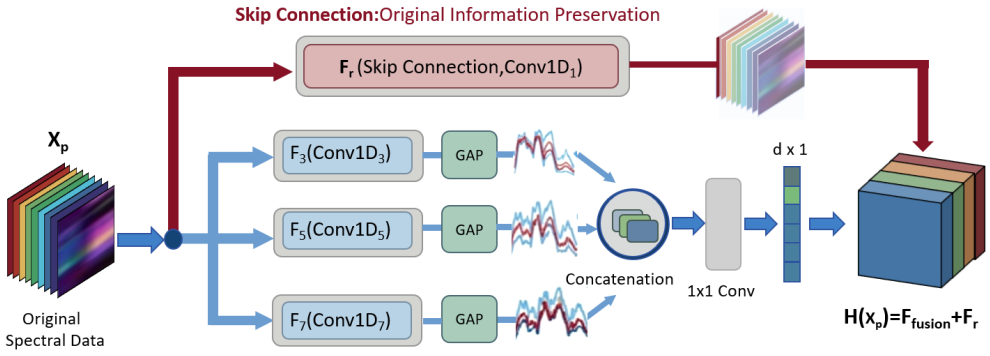


Figure 2. Spectral Branch Structure Diagram.

Given the input feature $\mathbf{X}_n \in \mathbb{R}^{H\times W\times B}$, it is first reshaped into a pixel-wise spectral sequence $\mathbf{X}_p \in \mathbb{R}^{(HW)\times 1\times B}$, where B denotes the number of spectral bands. To capture local spectral dependencies at various scales, three parallel 1D convolution branches with kernel sizes of 3, 5, and 7 are employed [35]. Crucially, a 1×1 residual projection branch is integrated as a core component to serve as an "identity-like" anchor, ensuring that the intrinsic spectral signatures are not smoothed out or lost during multi-scale feature extraction. These branches are formally expressed as:

$$\mathbf{F}_3 = \text{Conv1D}_3(\mathbf{X}_p), \tag{9}$$

$$\mathbf{F}_5 = \mathrm{Conv1D}_5(\mathbf{X}_p), \tag{10}$$

$$\mathbf{F}_7 = \mathrm{Conv1D}_7(\mathbf{X}_p), \tag{11}$$

$$\mathbf{F}_r = \mathrm{Conv1D}_1(\mathbf{X}_p), \tag{12}$$

where $\mathbf{F}_r$ represents the residual path that directly propagates the primary spectral response.

To maintain computational efficiency, global average pooling is applied along the spectral dimension of each branch to generate compact descriptors. The resulting features are concatenated and projected onto a unified d-dimensional space via linear mapping, yielding $\mathbf{X}_{\mathrm{spe}}$, which is then rearranged back to $\mathbb{R}^{H\times W\times d}$. This architecture ensures that the model can exploit high-level multi-scale correlations without sacrificing the fidelity of the raw spectral data, which is vital for distinguishing tree species with highly overlapping spectral profiles.

3.3. Spatial Branch

Besides spectral information, HSIs also contain strong spatial correlations among neighboring pixels. In particular, local textures, boundary structures, and neighborhood patterns are important for improving discriminability in complex scenes. Therefore, a multi-scale spatial branch is further introduced to extract local spatial contextual features.

First, the normalized input $\mathbf{X}_n$ is converted into a convolution-friendly tensor:

$$\mathbf{X}_s \in \mathbb{R}^{1\times B\times H\times W}. \tag{13}$$

Then, three parallel 2D convolution branches with kernel sizes of 3×3, 5×5, and 7×7 are adopted to extract local spatial information under different receptive fields. Meanwhile, a 1×1 convolution residual projection branch is introduced to complement the original spatial representation. The corresponding computations are

$$\mathbf{S}_3 = \mathrm{Conv2D}_{3\times 3}(\mathbf{X}_s), \tag{14}$$

$$\mathbf{S}_5 = \mathrm{Conv2D}_{5\times 5}(\mathbf{X}_s), \tag{15}$$

$$\mathbf{S}_7 = \mathrm{Conv2D}_{7\times 7}(\mathbf{X}_s), \tag{16}$$

$$\mathbf{S}_r = \mathrm{Conv2D}_{1\times 1}(\mathbf{X}_s). \tag{17}$$

After that, the four feature maps are concatenated along the channel dimension and projected to the unified feature dimension d through a 1×1 convolution, resulting in the spatial feature $\mathbf{X}_{\mathrm{spa}}$. The output is finally rearranged into a tensor of size $H\times W\times d$.

Through this design, the spatial branch can sufficiently capture local texture patterns, edge structures, and neighborhood information at multiple scales, while the residual path helps preserve shallow spatial priors and improves training stability.

3.4. Dual-Branch Feature Fusion

The spectral branch focuses on intra-pixel spectral correlations, whereas the spatial branch emphasizes local structural patterns in neighboring regions. Since the two branches are highly complementary, an effective feature fusion strategy is required.

First, the two branch features are concatenated along the channel dimension:

$$\mathbf{X}_{\mathrm{cat}} = [\mathbf{X}_{\mathrm{spe}}, \mathbf{X}_{\mathrm{spa}}] \in \mathbb{R}^{H\times W\times 2d}. \tag{18}$$

Then, $\mathbf{X}_{\text{cat}}$ is transformed into a convolutional tensor and passed through a 1×1 convolution for cross-channel interaction and dimensionality reduction, yielding the fused feature

$$\mathbf{X}_{\text{fus}} \in \mathbb{R}^{H \times W \times d}. \tag{19}$$

Compared with simple weighted summation, this fusion strategy enables more sufficient semantic interaction between spectral and spatial features at relatively low cost. On the one hand, the 1×1 convolution explicitly models channel-wise correlations across branches; on the other hand, the channel number is compressed back to a unified dimension, which is beneficial for the subsequent global sequence modeling.

3.5. Bidirectional Mamba Global Modeling Module

When flattening 2D spatial feature maps into 1D sequences, common scanning strategies include row-first, column-first, serpentine, and space-filling curve scans. Row-first scanning aligns with mainstream deep learning framework memory layouts, offering the highest computational efficiency; non-continuous scans like Hilbert curves require additional index mapping. Given that our dual-branch architecture theoretically suppresses spurious adjacency introduced by scanning order, row-first scanning becomes the natural choice.

Row-first scanning erroneously connects row-end and row-start pixels from adjacent rows as neighbors in the sequence. Our dual-branch architecture mitigates this from two aspects. First, the spatial branch performs multi-scale convolution on local spatial neighborhoods before flattening, with residual connections preserving raw spatial information, so each token entering the Mamba layer carries context validated through real spatial interactions, diluting the effect of artificial adjacency. Second, bidirectional Mamba processes the sequence in both forward and reverse orders simultaneously; the two outputs are added to form a near-"ring" cross-dependency structure, where pseudo-influences from row-end to row-start in the forward pass are counterbalanced by compensatory information from row-start to row-end in the reverse pass.

Although multi-scale convolution is effective for local spectral–spatial feature extraction, its receptive field is still limited by the kernel size and thus insufficient for modeling long-range contextual dependencies. To further enhance the representation of full-image semantic relationships, a bidirectional Mamba module is introduced after feature fusion, as shown in Figure 3.

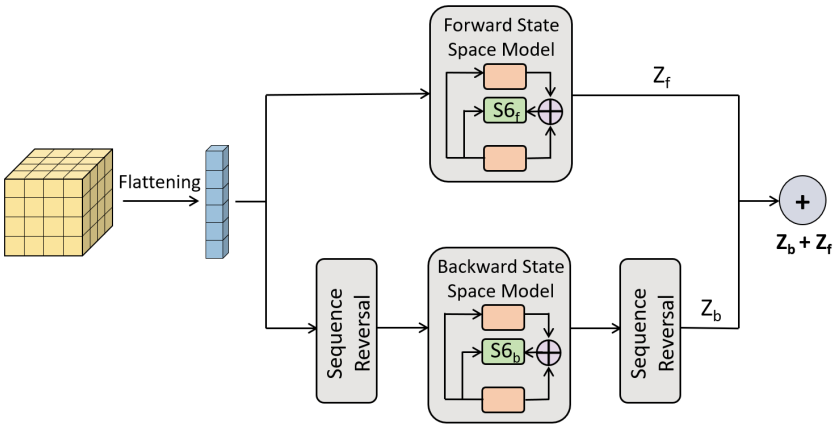


Figure 3. Bidirectional Mamba Modeling Module.

Given the fused feature

$$\mathbf{X}_{\text{fus}} \in \mathbb{R}^{H \times W \times d}, \tag{20}$$

It is first flattened into a sequence of length

$$L = H \times W, \tag{21}$$

resulting in

$$\mathbf{Z} \in \mathbb{R}^{1 \times L \times d}. \tag{22}$$

To capture both forward and backward contextual information, two independent S6 state space modules are used for forward and reverse scanning [36]. The forward feature is defined as

$$\mathbf{Z}_f = \mathrm{S6}_f(\mathbf{Z}), \tag{23}$$

and the backward feature is defined as

$$\mathbf{Z}_b = \mathrm{Flip}\big(\mathrm{S6}_b(\mathrm{Flip}(\mathbf{Z}))\big), \tag{24}$$

where $\mathrm{Flip}(\cdot)$ denotes sequence reversal. The two directional features are then fused by summation and combined with normalization, nonlinear activation, dropout, and residual connection to obtain the final globally enhanced feature:

$$\mathbf{Z}_m = \mathbf{Z} + G(\mathbf{Z}_f + \mathbf{Z}_b), \tag{25}$$

where $G(\cdot)$ denotes a nonlinear transformation composed of Layer Normalization, GELU activation, and Dropout.

The main advantages of this module are threefold. First, the state space model captures long-range dependencies with linear complexity. Second, the bidirectional scanning alleviates the information bias caused by unidirectional sequence modeling. Third, the residual connection and pre-normalization strategy improve both training stability and feature representation capability. Therefore, this module enables more complete pixel-wise contextual modeling over the whole image.

3.6. Classification Head

After obtaining the globally enhanced feature $\mathbf{Z}_m$, a lightweight multilayer perceptron is used as the classification head to map each feature vector to its category. The classification head consists of two fully connected layers with Layer Normalization, GELU activation, and Dropout inserted between them to improve nonlinear representation ability and suppress overfitting. It can be written as

$$\mathbf{P} = \mathrm{FC}_2\big(\mathrm{Drop}(\mathrm{GELU}(\mathrm{LN}(\mathrm{FC}_1(\mathbf{Z}_m))))\big), \tag{26}$$

where $\mathbf{P}$ denotes the sequence-form prediction result. Finally, $\mathbf{P}$ is rearranged back to the image space to obtain the pixel-wise classification output:

$$\mathbf{Y} \in \mathbb{R}^{H \times W \times C}, \tag{27}$$

where C is the number of categories.

3.7. Loss Function and Training Strategy

Since the proposed framework adopts a full-image pixel-wise classification paradigm, the supervision loss is computed only on labeled pixels during training. Let

$$\mathbf{M} \in \{0,1\}^{H \times W} \tag{28}$$

be the training mask, where only pixels satisfying $M(i,j) = 1$ participate in loss computation. For the pixel at location (i,j), let $\hat{\mathbf{y}}_{ij}$ denote the predicted class probability and y_{ij} denote the ground-truth label. The training objective is defined as a weighted cross-entropy loss:

$$\mathcal{L} = -\sum_{(i,j)\in\Omega} w_{y_{ij}} \log \hat{\mathbf{y}}_{ij}^{(y_{ij})}, \tag{29}$$

where Ω is the set of labeled training pixels and $w_{y_{ij}}$ is the class weight used to alleviate class imbalance [37].

The model is trained in an end-to-end manner using gradient-based optimization. Gradient clipping is further employed to suppress gradient explosion and improve training stability. During inference, the model outputs the full-image prediction in a single forward pass, thereby avoiding the repeated computation introduced by traditional patch-based sliding-window inference.

4. Experiments

4.1. Datasets

To comprehensively evaluate the classification performance and generalization ability of the proposed model, both self-built and public hyperspectral datasets are adopted for validation. Two subsets, namely HuaiL_1 (*HL_1*) and HuaiL_2 (*HL_2*), are derived from the self-constructed Zhangjiakou (*ZJK*) tree species dataset for fine-grained forest classification evaluation. In addition, several public benchmark datasets with diverse typical scenarios are employed for cross-task validation: *Pavia Centre* and *Houston* for urban land-cover classification, *WHU-Hi-LongKou* for fine-grained farmland classification. Table 1 lists all datasets adopted in this study.

Table 1. Summary of the datasets used in this study.

No.	Dataset	Location	Size (Pixel × Pixel)	Spatial Resolution (m)	Classes	Bands	Scenario
1	HL_1	China	3465 × 130	0.027	7	164	agroforest
2	HL_2	China	1240 × 265	0.027	6	164	agroforest
3	Pavia Centre	Italy	1096 × 715	1.3	9	102	urban
4	Houston	USA	601 × 2384	1.0	7	144	urban
5	WHU-Hi-LongKou	China	550 × 400	0.463	9	270	farmland

4.2. Zhangjiakou (ZJK)

The *ZJK* dataset is a high-spatial and high-spectral benchmark dataset established by our team, specifically designed for fine-grained tree species identification in complex urban-fringe and "forest-agriculture" mosaic ecosystems.

4.2.1. Study Area Overview and Environmental Background

The *ZJK* dataset was collected in Donghuayuan Town, Huailai County, Zhangjiakou City, Hebei Province, as shown in Figure 4, with geographical coordinates of 40°17′N, 115°53′E and an elevation of 470–480 m. The region is situated within the Huailai Basin and is characterized by a typical temperate continental monsoon climate, featuring four distinct seasons and concurrent rain and heat. The total area of the study site is approximately 400 mu, with flat micro-topography. The land cover encompasses artificial forests, nurseries, and seasonal agricultural crops, presenting typical transitional landscape characteristics. This provides a complex and diverse ecological background for tree species identification.

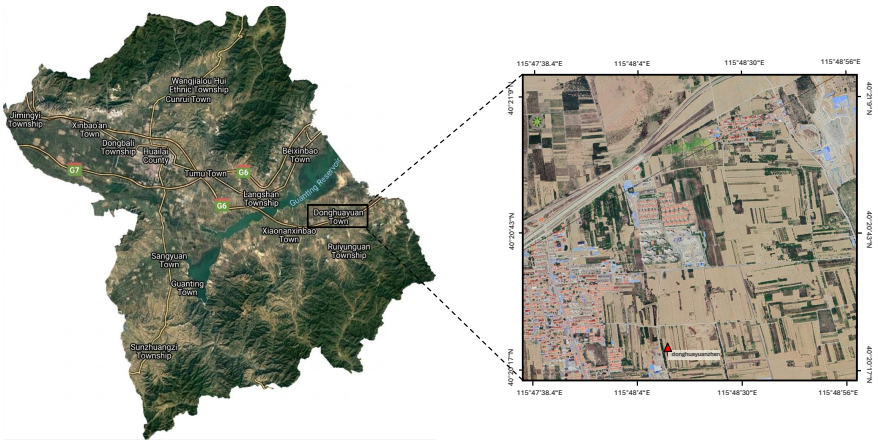
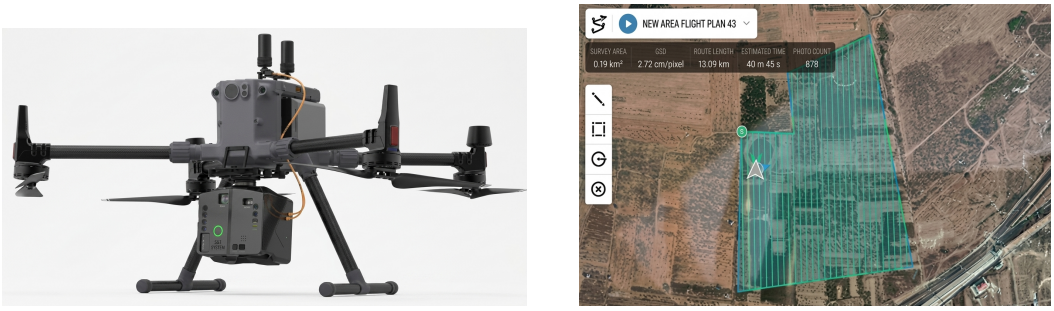


Figure 4. Geographical location and environmental setting of the study area: (**left**) topographic map of Huailai County, Hebei Province; (**right**) schematic diagram of the *ZJK* dataset site in Donghuayuan Town.

4.2.2. Spatial-Spectral Data Acquisition Platform

The imagery was acquired in October 2025 using an ultra-high-resolution multi-source remote sensing platform. A DJI Matrice 350 RTK UAV, equipped with an X20P-LIR integrated multi-source imaging system, served as the flight platform, as shown in Figure 5. To minimize shadow interference and ensure a high signal-to-noise ratio (SNR), flight missions were conducted between 14:00 and 15:00 under clear, cloudless weather conditions.

The UAV flight altitude was set at 80 m. The acquired HSI covers a spectral range of 350–1000 nm (consisting of 164 bands) with an image size of 940 × 475 pixels and a spatial resolution of approximately 0.027 m.



(**a**) (**b**)

Figure 5. The data acquisition platform and flight planning: (**a**) X20P-LIR system mounted on the DJI M300/350 RTK multi-rotor UAV; (**b**) Predefined flight route map.

4.2.3. Data Preprocessing and Field-Based Label Construction

In accordance with standard remote sensing processing specifications, the raw HSI was subjected to systematic preprocessing and high-quality ground truth dataset construction to support subsequent fine-grained tree species classification.

First, radiometric and geometric corrections were conducted to eliminate data distortions and ensure high spatial and spectral accuracy. The raw digital number (DN) values were converted to surface reflectance using laboratory calibration coefficients and the empirical line method with standard reference panels. Geometric distortions were further rectified by integrating high-precision GNSS/IMU observations and digital elevation model (DEM) data, enabling sub-decimeter spatial registration. These preprocessing steps ensured the geometric fidelity and spectral reliability of the extracted spectral curves.

Second, comprehensive field surveys were carried out over the entire study area to record field land cover types and tree species information. Afterwards, two representative subsets, *HL_1* and *HL_2*, were selected for manual labeling. Pixel-level land-cover classification of the two subsets was accomplished using LabelMe, and the annotation results were exported as JSON files. Finally, standardized ground truth labels in MAT format were produced via unified post-processing to facilitate model training and validation.

4.2.4. *HL_1* and *HL_2* Datasets

The *HL_1* dataset characterizes an urban-fringe landscape with mixed forest and agricultural land, covering seven land-cover categories (eight classes including the background) and 450,450 annotated pixels (299,271 valid terrestrial pixels and 151,179 background pixels). The mixed distribution of broadleaf and coniferous trees, as well as the interleaved layout of forest and agricultural areas, poses great challenges for fine-grained classification.

Characterized by a typical plantation-agriculture transition zone, the *HL_2* dataset contains six valid terrestrial land-cover categories (seven classes, including the background) and a total of 328,600 annotated pixels, consisting of 259,704 terrestrial pixels and 68,896 background pixels. The main classification challenges of this dataset stem from the blurred boundaries between forest litter and woodland, as well as the interleaved distribution of forest and agricultural land.

For all datasets in this work, stratified sampling is utilized when partitioning samples into the training set and testing set. Unless otherwise specified, 1% of the labeled samples of *HL_1* and *HL_2* are randomly selected with a fixed random seed for training, and the remaining samples are used for testing. The detailed sample distributions of *HL_1* and *HL_2* are presented in Tables 2 and 3, and their pseudo-color images and ground truth maps are shown in Figures 6 and 7.

Table 2. Land cover classes and training/testing sample statistics for the *HL_1* dataset.

No.	Class Name	Train	Test	Total
C1	Crabapple	566	56,023	56,589
C2	Chinese pine	207	20,523	20,730
C3	Grassland	560	55,428	55,988
C4	Mono maple	67	6612	6679
C5	Corn stubble	430	42,533	42,963
C6	Corn field	680	67,298	67,978
C7	Chinese scholar tree	483	47,861	48,344
BG	Background	1512	149,667	151,179
Total	/	4505	445,945	450,450

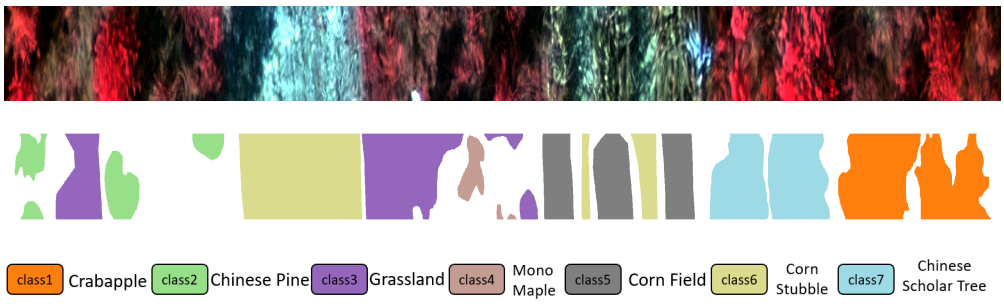


Figure 6. *HL_1* dataset. pseudo-RGB image. Ground truth map and color legends.

Table 3. Land cover classes and training/testing sample statistics for the *HL_2* dataset.

No.	Class Name	Train	Test	Total
C1	Crabapple	374	37,072	37,446
C2	Populus tomentosa	309	30,601	30,910
C3	Grassland	429	42,440	42,869
C4	Leaf litter	834	82,540	83,374
C5	Corn field	445	44,009	44,454
C6	Road	207	20,444	20,651
BG	Background	689	68,207	68,896
Total	/	3287	325,313	328,600

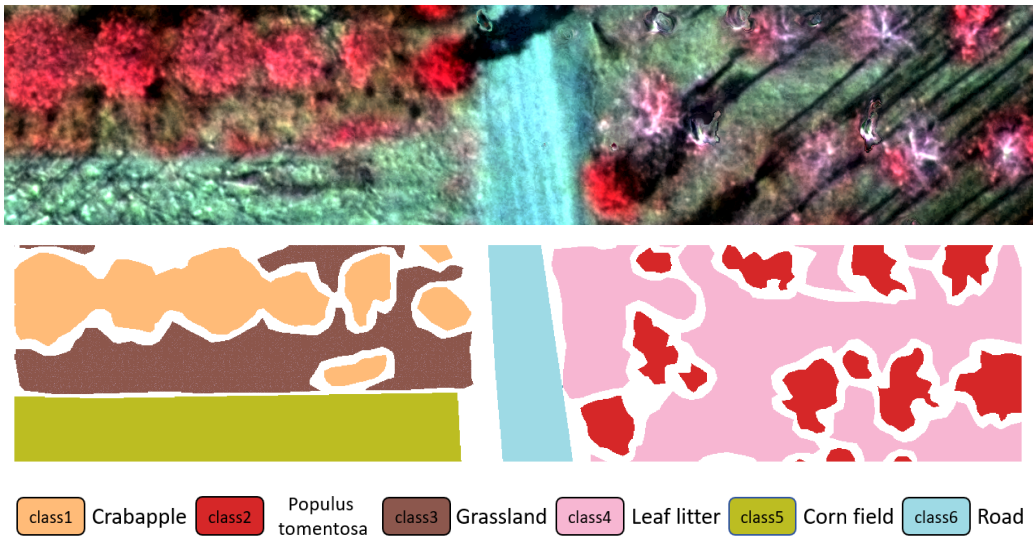


Figure 7. *HL_2* dataset. RGB image. Ground truth map and color legends.

4.2.5. Spectral Characteristic Analysis

In the *HL_1* scene, various land-cover categories exhibit distinctive spectral signatures. The prominent "red edge" effect and near-infrared (NIR) plateau features provide a solid physical foundation for the fine-grained identification of different tree species. As illustrated in Figure 8, the reflectance intensities of Chinese pine (C2) and Chinese scholar tree (C7) within the NIR plateau region are significantly higher than those of Crabapple (C1) and Mono maple (C4). This spectral heterogeneity, driven by the intrinsic physiological structures of the vegetation, serves as an advantageous condition for enabling accurate classification on this dataset.

In contrast, the *HL_2* scene presents a substantially greater classification challenge. As shown in Figure 9, the core difficulty of this dataset primarily stems from the severe "same spectrum, different objects" phenomenon and cross-band curve interleaving:

Extreme Overlapping in the NIR Region: The mean spectral profiles of Grassland (C3) and Corn field (C5) almost completely overlap within the NIR plateau interval from bands 100 to 140, constituting the primary technical bottleneck for classification.

Intertwining in the Visible and Transition Zones: Although the reflectance of Leaf litter (C4) remains low in the NIR region, its curve cross-cuts frequently with those of Crabapple (C1) and Populus tomentosa (C2) in the visible light range (bands 1–80). Meanwhile, a drastic slope crossover occurs between Crabapple (C1) and categories C2 and C3 within the red edge transition zone (bands 80–90).

Such extreme similarity in localized spectral regions, coupled with non-linear interleaving across the entire spectrum, constitutes the core challenge for fine-grained tree species classification in the *HL_2* dataset.

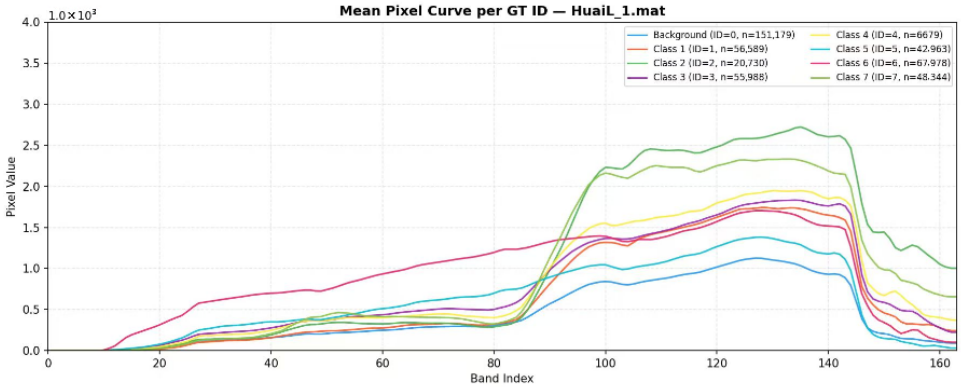


Figure 8. Spectral curves of *HL_1*.

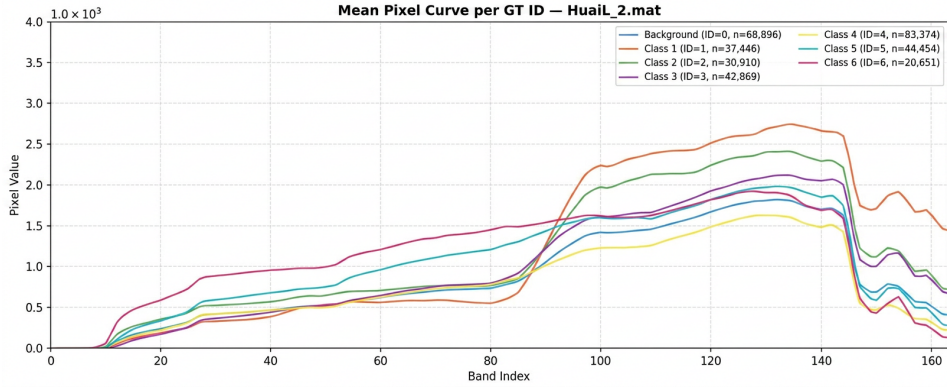


Figure 9. Spectral curves of *HL_2*.

4.3. Other Public Datasets

4.3.1. Pavia Centre

The Pavia Centre dataset was acquired by the Reflective Optics System Imaging Spectrometer sensor over the central urban area of Pavia, Italy. This HSI cube comprises 1096 × 715 pixels with 102 spectral bands spanning the wavelength range of 430–860 nm, capturing 9 distinct land-cover categories at a spatial resolution of 1.3 m. Table 4 shows the details of the samples and classes in this dataset. The ground truth and RGB scene of the dataset are shown in Figure 10.

Table 4. Land cover classes and sample statistics for each type in the Pavia Centre dataset.

No.	Class Name	Train	Test	Total
C1	Water	660	65,311	65,971
C2	Trees	76	7522	7598
C3	Asphalt	31	3059	3090
C4	Self-Blocking Bricks	27	2658	2685
C5	Bitumen	66	6518	6584
C6	Tiles	92	9156	9248
C7	Shadows	73	7214	7287
C8	Meadows	428	42,398	42,826
C9	Bare Soil	29	2834	2863
Total	/	1025	146,670	148,152

This dataset is adopted for comparative experiments and ablation studies. Unless otherwise stated, 1% of all labeled samples are randomly selected with a fixed random seed for model training, and the rest are used for testing.

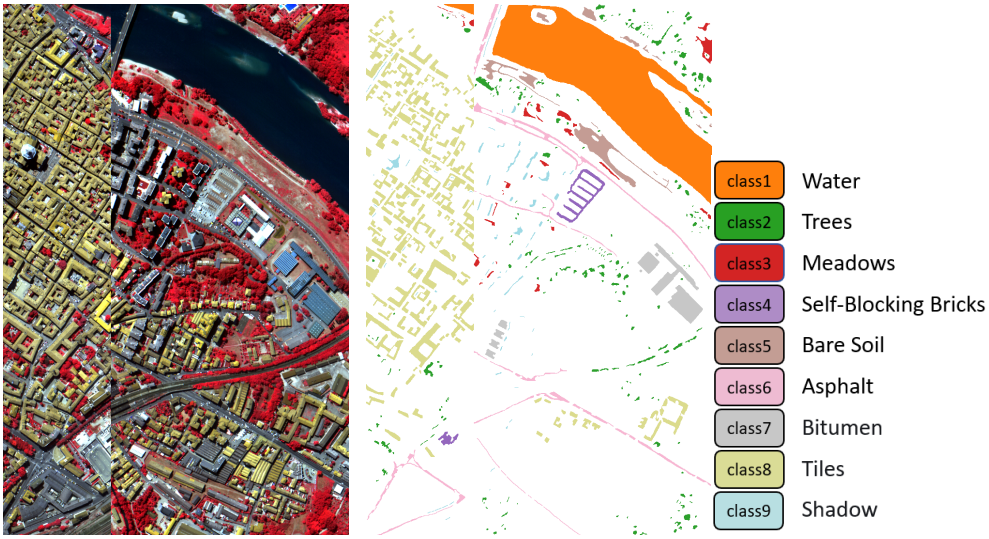


Figure 10. Pavia dataset: (**left**) RGB image, (**right**) Ground truth map and color legends.

4.3.2. Houston

The Houston dataset focuses on the University of Houston campus and its neighboring urban areas. This dataset was acquired by the National Center. It features image dimensions of 601 × 2384 pixels with a spatial resolution of 1 m, and includes 144 spectral bands spanning a wavelength range from 380 to 1050 nm. The dataset consists of 53,200 labeled samples categorized into 7 distinct land-cover classes. The detailed class-wise sample distribution is provided in Table 5. Both the RGB image and the corresponding ground truth map of the dataset are illustrated in Figure 11. This dataset is adopted to validate the model's generalization performance, where 10% of the pixels are used for training and the rest for testing.

Table 5. Land cover classes and sample statistics for each type in the Houston dataset.

No.	Class Name	Train	Test	Total
C1	Grass-healthy	135	1218	1353
C2	Grass-stressed	489	4399	4888
C3	Trees	277	2489	2766
C4	Water	2	20	22
C5	Residential buildings	535	4812	5347
C6	Non-residential buildings	3246	29,213	32,459
C7	Road	636	5729	6365
Total	/	5320	50,280	53,200

4.3.3. WHU-Hi-LongKou

The WHU-Hi-LongKou dataset was acquired over Longkou Town, Hubei Province, China, which covers a typical fine-grained agricultural scene with diverse and fragmented farmland landscapes. This dataset was collected using a Headwall Nano-Hyperspec imaging sensor mounted on a DJI Matrice 600 Pro unmanned aerial vehicle platform. It features image dimensions of 550 × 400 pixels, comprising 270 spectral bands spanning a wavelength range from 400 to 1000 nm, with a spatial resolution of approximately 0.463 m. The dataset consists of 9 distinct land-cover classes, and the detailed class-wise sample

distribution is provided in Table 6. Both the RGB image and the corresponding ground truth map of the dataset are illustrated in Figure 12. This dataset is also adopted to validate the model's generalization performance, where 1% of the pixels are used for training and the rest for testing.

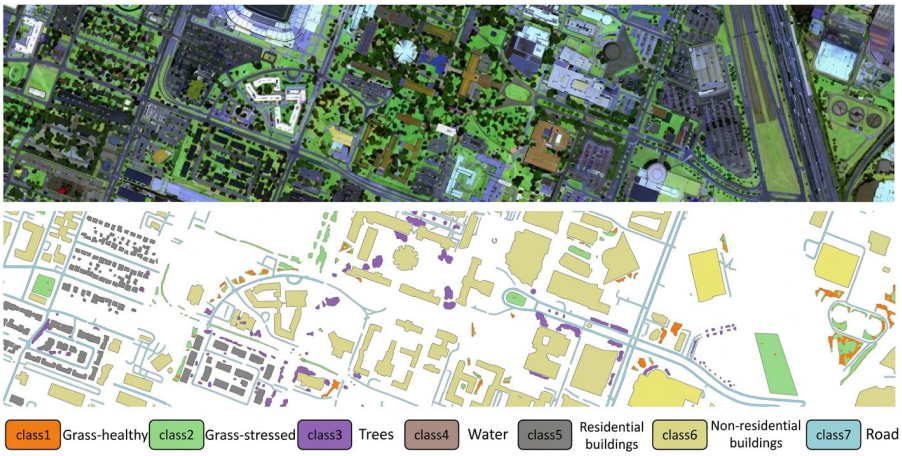


Figure 11. Houston dataset: (**top**) RGB image, (**bottom**) Ground truth map and color legends.

Table 6. Land cover classes and sample statistics for each type in the WHU-Hi-LongKou dataset.

No.	Class Name	Train	Test	Total
C1	Corn	345	34,166	34,511
C2	Cotton	84	8290	8374
C3	Sesame	30	3001	3031
C4	Broad-leaf soybean	632	62,580	63,212
C5	Narrow-leaf soybean	42	4109	4151
C6	Rice	119	11,735	11,854
C7	Water	671	66,385	67,056
C8	Roads and houses	71	7053	7124
C9	Mixed weed	52	5177	5229
Total	/	2046	202,496	204,542

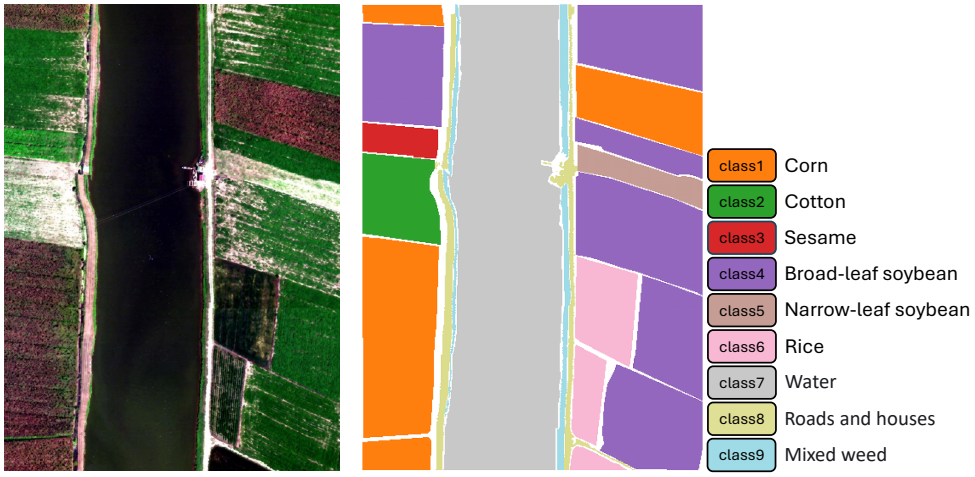


Figure 12. WHU-Hi-LongKou dataset: (**left**) False-color composite, (**right**) Ground truth map and color legends.

4.4. Experiment

4.4.1. Experimental Settings

1. **Evaluation metrics:** Following existing HSIC studies, overall accuracy (OA), average accuracy (AA), and the Kappa coefficient were employed for quantitative accuracy assessment. Meanwhile, parameter size, floating point operations (FLOPs), peak GPU memory usage, training time, and testing time were adopted to comprehensively evaluate model computational performance.
2. **Compared methods:** To comprehensively validate the superiority and effectiveness of the proposed network, a series of representative HSIC methods were selected for qualitative and quantitative comparison. The comparison methods cover multiple mainstream technical frameworks, including traditional machine learning, CNN-based, Transformer-based, and Mamba-based models. Specifically, the compared approaches involve the support vector machine (SVM) [12], two CNN-based methods (Tri-CNN [35] and HybridSN [38]), one Transformer-based method (SSFTT) [21], and three advanced Mamba-based methods (MambaHSI [25], 3DSS-Mamba [23], and SS-Mamba [39]).
    - **(a)** **SVM:** As a classic machine learning algorithm for HSIC, SVM follows the structural risk minimization principle and employs kernel functions to process high-dimensional nonlinear hyperspectral data, achieving efficient spectral feature classification. The detailed hyperparameter settings of SVM are as follows: SVM with RBF kernel, C = 10.0, gamma = "scale", TEST_RATIO = 0.01, Standard Scaler normalization.
    - **(b)** **Tri-CNN:** Tri-CNN adopts a multi-branch feature fusion architecture, which is capable of capturing multi-scale and multimodal spectral–spatial features from HSIs. The detailed hyperparameter settings of Tri-CNN are as follows: epochs = 1, batch size = 16, learning rate = 1×10^{-3} (Adam), window size = 9, PCA components = 45, Conv3D filters = 64 with kernel sizes of (1,1,3), (3,3,1), and (3,3,3) for the three branches, dense layers of 512 and 256 units, dropout rate = 0.3, early stopping patience = 10.
    - **(c)** **HybridSN:** HybridSN integrates 2D and 3D convolutions to collaboratively model spatial and spectral information, thereby enabling the extraction of deep joint spectral–spatial features for HSIC tasks. The detailed hyperparameter settings of our HybridSN model are as follows: patch size = 15, PCA components = 30, batch size = 256, epochs = 40, learning rate = 0.001, weight decay = 1×10^{-4}, dropout = 0.4.
    - **(d)** **SSFTT:** The SSFTT model converts 3D spectral–spatial cubes into feature tokens and leverages a dual-branch encoder to implement end-to-end global joint feature learning. The detailed hyperparameter settings of our SSFTT model are as follows: epochs = 200, batch size = 2, learning rate = 0.0005 (with StepLR decay at step 30, gamma 0.5), weight decay = 1×10^{-5}, num tokens = 4, embedding dimension dim = 64, depth = 1, attention heads = 8, mlp dimension = 8, dropout = 0.1, embedding dropout = 0.1, patch size = 9, PCA components = 40.
    - **(e)** **MambaHSI:** MambaHSI alternately stacks spatial and spectral Mamba modules to realize decoupled modeling and interactive fusion of long-range spectral–spatial dependencies in hyperspectral data. The detailed hyperparameter settings of MambaHSI are as follows: epochs = 200, learning rate = 0.0003, training ratio = 0.005, and hidden dimension = 64.
    - **(f)** **3DSS-Mamba:** 3DSS-Mamba reorganizes 3D hyperspectral data into long sequence forms and utilizes bidirectional Mamba blocks to achieve unified global spectral–spatial modeling. The detailed hyperparameter settings of

3DSS-Mamba are as follows: training epochs = 100, batch size = 64, learning rate = 0.001 (Adam optimizer), weight decay = 5×10^{-4}, patch size = 11, PCA components = 30, embed dim = 32, depth = 1, d_state = 16, dt_rank = 2, k_group = 4 with a 3D convolution frontend (channels = 32, kernel = (3,5,5)).

(g) **SS-Mamba:** SS-Mamba constructs a Mamba-based spectral–spatial learning framework, which extracts refined hyperspectral features via a token generation mechanism and a dedicated feature enhancement module. The detailed hyperparameter settings of SS-Mamba model are as follows: model_id = 2 (Spectral-Spatial fusion), epoch = 190, batch_size = 512, learning_rate = 5×10^{-4} with MultiStepLR decay (milestones at [80, 140, 170], gamma = 0.5), embed_dim = 64, depth = 4, hid_chans = 64, spa_patch_size = 3, spe_patch_size = 2, with bidirectional SSM (use_bi = True), global pooling (use_global = True), CLS token (use_cls = True), and center augmentation fusion (use_fu = True).

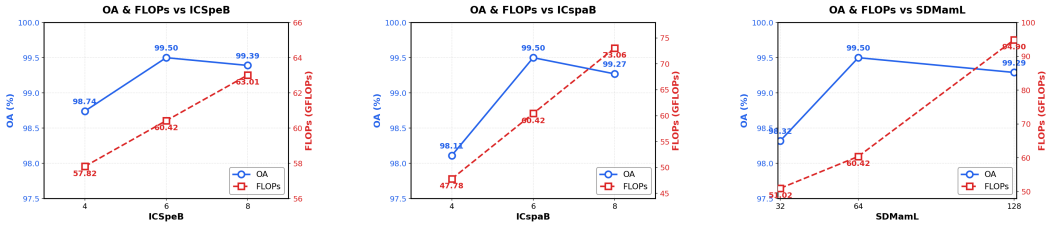
3. **Implementation details:** All experiments were implemented under the PyTorch (version 2.10.0) and Python (version 3.12.13) framework, except for the comparative experiment of Tri-CNN (TensorFlow). The hardware platform was equipped with an NVIDIA A100 GPU (40 GB) (NVIDIA Corporation, Santa Clara, CA, USA), a 16-core Intel CPU (Intel Corporation, Santa Clara, CA, USA), and 96 GB RAM. A series of experiments were conducted, including parameter analysis, comparison experiments, ablation studies and generalization tests, to fully assess model performance. To ensure fair and reproducible comparisons, we use a fixed random seed to split training and test sets for reproducibility, with partition details illustrated earlier. To reduce randomness, each experiment runs ten times with ten distinct random seeds for model initialization. The same seed settings are applied to our model and all competing approaches, and averaged results are presented. Meanwhile, all comparative algorithms were run using the official open-source codes and optimal hyperparameters provided by their authors. The proposed MambaHSINet and all competing methods adopted the same data preprocessing pipeline. Only the hardware and software environments differed during model execution. Given that the proposed dual-branch bidirectional Mamba architecture takes the entire HSI as input, the batch size was set to 1. The detailed hyperparameter settings of our model are as follows: epochs = 100, batch size = 1, and learning rate = 0.005.

4.4.2. Parameter Analysis

This section analyzes three key hyperparameters that govern the multi-scale feature extraction and global dependency modeling capabilities of MambaHSINet, including the Intermediate Channels of the Spectral Branch (ICSpeB), Intermediate Channels of the Spatial Branch (ICSpaB), and State Dimension of the Mamba Layer (SDMamL). ICSpeB regulates the channel capacity for multi-scale spectral learning, where three 1D convolutions with kernel sizes of 3, 5, and 7 map pixel-wise spectral vectors to intermediate spectral features with a fixed channel dimension. Hyperspectral spectral curves exhibit continuous and locally correlated properties. We empirically adopt these kernel sizes to capture correlations among adjacent spectral bands at different local ranges. This combination effectively extracts short-, medium- and long-range local spectral variations across adjacent bands, helping to identify subtle spectral discrepancies between similar categories. Similarly, ICSpaB controls the channel dimensionality of multi-scale spatial features, which are derived using three 2D convolutions (3×3, 5×5, and 7×7) for spatial texture modeling. As a core hyperparameter of the Mamba module, SDMamL defines the hidden state dimension and

directly determines the model's ability to capture long-range spatial–spectral dependencies within the entire HSI.

To explore the parameter sensitivity and determine the optimal configuration, ablation experiments were implemented on the HL_1 dataset using only 1% training samples. These three hyperparameters present a typical trade-off between model representation capacity and computational cost. Small values restrict feature learning and reduce classification accuracy, whereas excessively large values bring marginal accuracy gains but significantly increase parameter volume, memory usage, and training overhead. Accordingly, we set the search ranges as ICSpeB $\in [4, 6, 8]$, ICSpaB $\in [4, 6, 8]$, and SDMamL $\in [32, 64, 128]$, with quantitative results in terms of Overall Accuracy (OA) and FLOPs reported in Figure 13.



(**a**) (**b**) (**c**)

Figure 13. Parameter analysis (**a**) ICSpeB vs. OA, (**b**) ICspaB vs. OA, (**c**) SDMamL vs. OA.

The experimental results show distinct performance variations across hyperparameter settings. For ICSpeB, the value of 4 achieves the lowest OA (98.74%) with low computation (57.82 GFLOPs), while ICSpeB = 8 slightly increases OA to 99.39% but raises FLOPs to 63.01 GFLOPs, and only ICSpeB = 6 obtains the highest OA of 99.50% with moderate computational overhead (60.42 GFLOPs). For ICSpaB, the value of 4 causes low OA (98.11%) due to inadequate spatial feature extraction; although ICSpaB = 8 yields acceptable accuracy (99.27%), it brings excessive computational overhead (73.06 GFLOPs), whereas ICSpaB = 6 reaches the peak OA of 99.50% without redundant computation. For SDMamL, the setting of 32 limits global spectral-spatial modeling and leads to low OA (98.32%), and SDMamL = 128 delivers only suboptimal accuracy (99.29%) but drastically increases FLOPs to 94.90 GFLOPs, while SDMamL = 64 achieves the maximum OA of 99.50% with controlled computational complexity.

Accordingly, the optimal hyperparameter combination for MambaHSINet is ICSpeB = 6, ICSpaB = 6 and SDMamL = 64. All three hyperparameters achieve identical top classification accuracy with affordable computational costs. This combination balances classification accuracy and inference efficiency, which is applicable to general practical UAV hyperspectral classification tasks.

4.4.3. Ablation Study

To quantitatively verify the efficacy of the core components of MambaHSINet, systematic ablation experiments are conducted on three representative hyperspectral datasets (*HL_1*, *HL_2*, and *Pavia Centre*) covering forest and urban scenarios. Three standard metrics, including overall accuracy (OA), average accuracy (AA), and Kappa coefficient, are employed for evaluation. The ablation results with 1% training samples are presented in Table 7.

First, we analyze the contributions of the spatial branch, spectral branch, and Mamba module (no Mamba, unidirectional Mamba, and bidirectional Mamba) under the 1% training sample setting. As shown in Table 7, the bidirectional Mamba module is critical for performance improvement. Removing the Mamba module causes severe accuracy drops:

OA decreases from 99.50% to 95.68% on *HL_1*, from 98.39% to 96.03% on *HL_2*, and from 99.05% to 97.61% on *Pavia Centre*. In addition, bidirectional Mamba outperforms unidirectional Mamba with OA gains of 0.45% (*HL_1*), 0.49% (*HL_2*), and 0.38% (*Pavia Centre*). These results demonstrate that bidirectional Mamba effectively captures long-range global dependencies and compensates for the limited receptive field of local convolutions.

Table 7. Ablation results of core network components with 1% training samples on three datasets (%). ✓: module adopted; ×: module excluded.

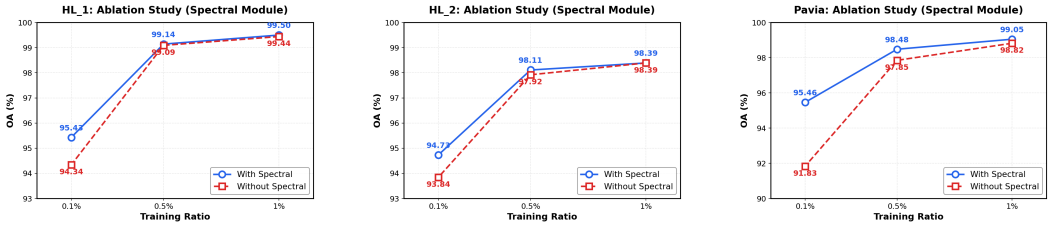
Components			HL_1 (%)			HL_2 (%)			Pavia (%)		
Spatial	**Spectral**	**Mamba**	**OA**	**AA**	**K × 100**	**OA**	**AA**	**K × 100**	**OA**	**AA**	**K × 100**
×	✓	bidirectional	97.16 ± 0.67	97.60 ± 0.53	96.56 ± 0.81	94.61 ± 0.57	95.36 ± 0.53	93.28 ± 0.71	98.40 ± 0.31	96.13 ± 0.76	97.74 ± 0.44
✓	×	bidirectional	99.44 ± 0.14	99.52 ± 0.15	99.32 ± 0.17	98.39 ± 0.14	98.58 ± 0.13	97.98 ± 0.17	98.82 ± 0.26	96.82 ± 0.75	98.33 ± 0.37
✓	✓	no	95.68 ± 0.67	95.81 ± 0.73	94.78 ± 0.80	96.03 ± 0.32	96.48 ± 0.28	95.05 ± 0.40	97.61 ± 0.36	92.93 ± 0.94	96.61 ± 0.50
✓	✓	bidirectional	**99.50 ± 0.20**	**99.62 ± 0.17**	**99.39 ± 0.24**	**98.39 ± 0.25**	**98.58 ± 0.15**	**97.98 ± 0.31**	**99.05 ± 0.17**	**97.16 ± 0.66**	**98.65 ± 0.24**
✓	✓	unidirectional	99.05 ± 0.21	99.12 ± 0.22	98.86 ± 0.25	97.90 ± 0.27	98.09 ± 0.17	97.38 ± 0.33	98.67 ± 0.24	95.98 ± 0.81	98.12 ± 0.34

Apart from the Mamba module, the spatial branch plays a dominant role in ensuring stable classification performance. In contrast to the marginal contribution of the spectral branch at the 1% training ratio, removing the spatial branch leads to a significant accuracy decline. Specifically, the model without the spatial branch yields OA values of 97.16% (*HL_1*), 94.61% (*HL_2*), and 98.40% (*Pavia Centre*), which are 2.34%, 3.78%, and 0.65% lower than those of the complete dual-branch model, respectively. Such obvious performance degradation indicates that spatial structural features are the primary guarantee for accurate fine-grained classification under standard sample quantities. Nevertheless, the spectral branch contributes marginally at the 1% training ratio. The spatial-only model achieves comparable performance to the full dual-branch model, with OA differences of only 0.06% on *HL_1*, 0.00% on *HL_2*, and 0.23% on *Pavia Centre*, along with negligible AA and Kappa gaps. This indicates that spectral features provide limited incremental information when spatial features are fully available and training samples are sufficient. To further validate the practical value of spectral features under sparse samples, we conduct supplementary ablation experiments under 0.5% and 0.1% training ratios, as shown in Figure 14.

Multi-ratio experimental results verify the indispensable complementary effect of spectral features under insufficient labeled samples. As the training ratio decreases, the performance advantage of the dual-branch structure gradually increases. At the 0.5% ratio, the dual-branch model achieves OA improvements of 0.05%, 0.19%, and 0.63% on *HL_1*, *HL_2*, and *Pavia Centre*, respectively. Under the extreme 0.1% ratio, the performance gains are substantially amplified, with OA increases of 1.09%, 0.89%, and 3.63% on the three datasets, accompanied by consistent AA and Kappa promotions. These results confirm that spatial features provide stable structural prior information, while spectral features supply unique spectral signatures. Integrating with bidirectional Mamba's global modeling capability, MambaHSINet achieves excellent accuracy and strong robustness under both adequate and limited training samples.

Collectively, the ablation results verify the cooperative and complementary effects of the model's core components. The spatial branch dominates stable fine-grained classification, the spectral branch supplements discriminative features under sparse sample conditions, and the bidirectional Mamba module enhances global feature representation to compensate for the drawbacks of local convolutions. This collaborative design effectively balances local detail extraction and global semantic modeling, improving the classification

performance and sample robustness of the proposed model under both sufficient and limited labeled samples.



(**a**) (**b**) (**c**)

Figure 14. Ablation experiments on the spectral branch with different proportions of training samples (**a**) HL_1 dataset, (**b**) HL_2 dataset, (**c**) Pavia Centre dataset.

4.5. Comparison with Methods

To evaluate MambaHSINet, we compared it with traditional (SVM), CNN-based (Tri-CNN, HybridSN), Transformer-based (SSFTT), and SSM-based (MambaHSI, 3DSS-Mamba, SS-Mamba) methods. Quantitative results for the *HL_1*, *HL_2*, and *Pavia Centre* datasets are summarized in Tables 8–10.

Table 8. Classification performance (%) of SVM, Tri-CNN, HybridSN, SSFTT, MambaHSI, 3DSS-Mamba, SS-Mamba, and the proposed MambaHSINet on the HL_1 dataset.

Class	SVM	Tri-CNN	HybridSN	SSFTT	MambaHSI	3DSS-Mamba	SS-Mamba	Proposed
C1	91.14 ± 1.02	98.69 ± 0.97	98.59 ± 0.86	98.06 ± 0.96	**100.00 ± 0.00**	87.09 ± 73.70	98.72 ± 0.40	99.93 ± 0.14
C2	79.77 ± 1.23	96.70 ± 0.99	95.40 ± 1.44	97.27 ± 1.27	99.39 ± 0.34	80.90 ± 5.01	95.56 ± 1.18	**99.70 ± 0.25**
C3	92.76 ± 0.72	96.35 ± 0.85	96.37 ± 2.54	92.57 ± 1.91	**99.64 ± 0.20**	87.17 ± 37.77	97.20 ± 0.39	99.26 ± 0.32
C4	36.53 ± 3.60	85.46 ± 10.36	84.04 ± 11.33	84.50 ± 3.73	98.06 ± 1.12	58.06 ± 32.81	84.82 ± 3.16	**99.32 ± 1.13**
C5	96.67 ± 0.97	98.24 ± 0.79	98.28 ± 0.59	96.81 ± 1.56	**99.69 ± 0.16**	94.49 ± 4.55	98.71 ± 0.50	99.51 ± 0.41
C6	96.21 ± 0.32	99.11 ± 0.44	98.17 ± 0.84	97.69 ± 1.05	**99.76 ± 0.15**	97.00 ± 1.44	98.94 ± 0.33	98.94 ± 0.49
C7	86.52 ± 1.31	98.06 ± 0.46	98.02 ± 0.98	97.81 ± 0.96	**99.97 ± 0.04**	85.09 ± 22.09	98.01 ± 0.31	99.95 ± 0.08
OA	90.64 ± 0.27	97.75 ± 0.13	97.40 ± 0.60	96.37 ± 0.55	**99.74 ± 0.07**	89.02 ± 1.89	97.84 ± 0.16	99.50 ± 0.20
AA	82.80 ± 0.60	96.09 ± 1.27	95.55 ± 1.81	94.96 ± 0.76	99.50 ± 0.18	84.26 ± 2.23	96.00 ± 0.54	**99.62 ± 0.17**
Kappa	88.61 ± 0.33	97.27 ± 0.15	96.85 ± 0.01	95.61 ± 0.67	**99.69 ± 0.09**	86.68 ± 2.79	97.38 ± 0.19	99.39 ± 0.24

Table 9. Classification performance (%) of SVM, Tri-CNN, HybridSN, SSFTT, MambaHSI, 3DSS-Mamba, SS-Mamba, and the proposed MambaHSINet on the HL_2 dataset.

Class	SVM	Tri-CNN	HybridSN	SSFTT	MambaHSI	3DSS-Mamba	SS-Mamba	Proposed
C1	96.14 ± 0.81	98.71 ± 0.32	97.82 ± 1.04	96.22 ± 1.24	98.80 ± 0.42	95.66 ± 1.32	97.20 ± 1.19	**98.94 ± 0.23**
C2	85.26 ± 1.87	94.16 ± 0.74	95.65 ± 0.87	91.17 ± 1.82	95.76 ± 0.75	83.67 ± 16.60	91.85 ± 1.62	**96.85 ± 0.76**
C3	89.92 ± 0.72	95.49 ± 1.26	96.83 ± 1.32	91.77 ± 2.52	**98.97 ± 0.35**	86.66 ± 9.88	94.53 ± 1.18	98.64 ± 0.51
C4	95.97 ± 0.34	97.58 ± 0.59	95.15 ± 2.08	92.80 ± 1.65	**98.92 ± 0.23**	93.10 ± 4.56	96.28 ± 1.04	97.44 ± 0.78
C5	95.36 ± 0.28	99.01 ± 0.33	98.05 ± 1.00	96.05 ± 1.14	98.91 ± 0.69	93.43 ± 7.08	97.85 ± 1.19	**99.92 ± 0.08**
C6	96.40 ± 0.55	99.39 ± 0.67	97.90 ± 0.57	99.06 ± 0.55	99.67 ± 0.31	95.50 ± 14.93	97.87 ± 1.02	**99.71 ± 0.39**
OA	93.65 ± 0.17	97.38 ± 0.17	96.59 ± 0.98	93.98 ± 0.62	**98.59 ± 0.17**	91.53 ± 0.71	95.99 ± 0.46	98.39 ± 0.25
AA	93.18 ± 0.31	97.39 ± 0.11	96.90 ± 0.72	94.51 ± 0.52	98.51 ± 0.19	91.34 ± 1.55	95.93 ± 0.55	**98.58 ± 0.15**
Kappa	92.02 ± 0.23	96.72 ± 0.21	95.74 ± 0.01	92.49 ± 0.77	**98.24 ± 0.21**	89.37 ± 1.55	94.98 ± 0.57	97.98 ± 0.31

Table 10. Classification performance (%) of SVM, Tri-CNN, HybridSN, SSFTT, MambaHSI, 3DSS-Mamba, SS-Mamba, and the proposed MambaHSINet on the Pavia Centre dataset.

Class	SVM	Tri-CNN	HybridSN	SSFTT	MambaHSI	3DSS-Mamba	SS-Mamba	Proposed
C1	99.96 ± 0.08	**100.00 ± 0.00**	99.99 ± 0.02	99.33 ± 1.37	99.99 ± 0.01	99.81 ± 0.03	99.99 ± 0.01	99.98 ± 0.02
C2	92.73 ± 2.47	95.56 ± 0.90	96.68 ± 1.05	94.96 ± 1.92	96.61 ± 1.06	90.95 ± 38.23	95.05 ± 1.51	95.51 ± 1.72
C3	84.17 ± 4.51	93.91 ± 5.78	87.06 ± 20.62	88.71 ± 5.96	82.76 ± 1.42	85.51 ± 39.86	**90.30 ± 3.90**	88.66 ± 5.43
C4	65.72 ± 3.94	88.37 ± 9.94	96.39 ± 2.03	90.00 ± 5.89	96.93 ± 1.76	94.91 ± 8.95	85.81 ± 4.58	**98.12 ± 1.63**
C5	95.46 ± 1.51	97.43 ± 2.65	96.18 ± 1.61	90.51 ± 7.15	95.99 ± 1.58	93.23 ± 24.14	96.86 ± 2.38	**99.19 ± 0.37**
C6	95.35 ± 0.88	**99.30 ± 0.80**	98.28 ± 1.61	98.59 ± 0.67	99.03 ± 0.39	95.07 ± 3.59	97.94 ± 0.26	96.10 ± 1.05
C7	88.92 ± 1.71	96.53 ± 0.71	97.17 ± 1.54	93.54 ± 3.45	95.72 ± 2.58	93.12 ± 8.15	94.35 ± 0.75	**97.38 ± 1.40**
C8	99.37 ± 0.14	**99.92 ± 0.07**	99.67 ± 0.19	99.29 ± 0.16	99.92 ± 0.04	99.10 ± 0.03	99.86 ± 0.04	99.91 ± 0.07
C9	**99.95 ± 0.02**	97.22 ± 2.28	84.74 ± 7.40	95.63 ± 1.49	95.74 ± 1.94	86.62 ± 87.84	99.94 ± 0.08	99.57 ± 0.30
OA	97.44 ± 0.07	99.03 ± 0.09	98.68 ± 0.60	97.91 ± 0.90	98.85 ± 0.19	97.59 ± 0.56	98.70 ± 0.18	**99.05 ± 0.17**
AA	91.29 ± 0.44	96.47 ± 0.63	95.13 ± 3.03	94.51 ± 0.68	95.86 ± 0.45	93.15 ± 3.98	95.57 ± 0.71	**97.16 ± 0.66**
Kappa	96.37 ± 0.11	98.62 ± 0.13	98.13 ± 0.01	97.05 ± 1.24	98.38 ± 0.26	96.58 ± 1.13	98.16 ± 0.25	**98.65 ± 0.24**

(1) Quantitative Analysis

The traditional machine learning method, SVM, shows the weakest classification capability across all datasets. For example, on *HL_1*, its OA, AA, and Kappa are only 90.64%, 82.80%, and 88.61%, respectively. Deep learning models, including Tri-CNN, HybridSN, and SSFTT, achieve substantial performance gains over the traditional SVM across all three datasets. Within this group, Tri-CNN achieves outstanding overall performance, with OA of 97.75% on *HL_1*, 97.38% on *HL_2*, and 99.03% on *Pavia Centre*. HybridSN also obtains competitive classification results, while SSFTT yields relatively lower accuracy compared with the two CNN-based methods. Even though these deep learning methods deliver promising results, they are universally outperformed by MambaHSI and the proposed MambaHSINet.

Significant performance gaps exist among different SSM-based Mamba variants. 3DSS-Mamba obtains the lowest OA and AA across all datasets: its OA and AA are only 89.02% and 84.26% on *HL_1*, 91.53% and 91.34% on *HL_2*, and 97.59% and 93.15% on *Pavia Centre*. More notably, most of its per-class accuracies exhibit extremely large standard deviations, which indicates severe result fluctuations and poor classification stability and robustness. SS-Mamba achieves stable and competitive classification performance. On the *HL_1* dataset, its OA reaches 97.84%, slightly higher than Tri-CNN (97.75%) and HybridSN (97.40%). On *HL_2* and *Pavia Centre*, its OA is 95.99% and 98.70%, respectively, which is lower than Tri-CNN but superior to SSFTT. Nevertheless, the comprehensive metrics of SS-Mamba are consistently inferior to MambaHSI and the proposed MambaHSINet on all three test datasets.

Both MambaHSI and the proposed MambaHSINet adopt the SSM architecture and rank in the optimal tier among all competing methods, with comparable overall classification results. Detailed comparisons between the two algorithms on the three datasets are elaborated as follows. On the *HL_1* dataset, MambaHSI yields an OA of 99.74% and a Kappa coefficient of 99.69%, slightly exceeding those of the proposed algorithm. Moreover, MambaHSI achieves superior classification accuracy for categories C3 and C6, indicating its strong fitting capability for certain samples. In contrast, the proposed MambaHSINet achieves an AA of 99.62%, surpassing MambaHSI (99.50%). Noticeable accuracy improvements are observed on C2 (99.70% vs. 99.39%) and C4 (99.32% vs. 98.06%). Both algorithms maintain low standard deviations for per-class prediction results, demonstrating reliable classification stability. On the *HL_2* dataset, MambaHSINet achieves competitive comprehensive performance. Its AA reaches 98.58%, slightly higher than the 98.51% of MambaHSI. Nevertheless, its OA is 98.39%, and the Kappa coefficient is 97.98%, both marginally inferior to MambaHSI (OA = 98.59%, Kappa = 98.24%). In terms of per-class classification results, the proposed algorithm achieves accuracy improvements for most categories, including

C2, C5, and C6, which verifies its stronger feature extraction and category discrimination capacity. On the *Pavia Centre* dataset, the superiority of MambaHSINet is further amplified. Its three core metrics, namely OA (99.05%), AA (97.16%), and Kappa coefficient (98.65%), rank first among all comparison methods and comprehensively outperform MambaHSI. Specifically, it achieves substantially higher classification accuracy on categories such as C4, C5, C7, and C9, fully validating its outstanding generalization ability and overall classification performance.

Taken together, quantitative comparisons on three datasets illustrate that the proposed MambaHSINet achieves better balanced classification performance and higher per-class accuracy for most categories. Although it lags slightly behind MambaHSI on OA and Kappa metrics for the *HL_1* and *HL_2* datasets, our method achieves significantly superior core evaluation indices on the *Pavia Centre* dataset and outperforms all competing models, including traditional machine learning, CNN-based methods, and other SSM variants. Admittedly, MambaHSI also achieves competitive accuracy and can be regarded as a reliable, strong baseline for hyperspectral classification tasks.

(2) Qualitative Analysis

Figures 15–17 show the visual classification results of all methods on three hyperspectral datasets. As can be seen, the proposed MambaHSINet achieves the optimal visual performance. Its classification maps feature complete spatial regions for each land-cover class, with minimal misclassification and in-map noise. The interior areas of individual categories are cleaner compared with SVM, CNN, Transformer, and other Mamba-based models.

All algorithms encounter classification errors and random noise at the boundaries of different ground objects. Such errors are primarily caused by spectral aliasing between adjacent land covers and the inherent performance limitations of classification models. No large-area continuous misclassification appears in any result. Compared with other approaches, MambaHSINet remarkably reduces boundary misclassifications and produces clearer, more precise category boundaries. The above visual observations are consistent with the quantitative evaluation, further demonstrating the superiority of our method.

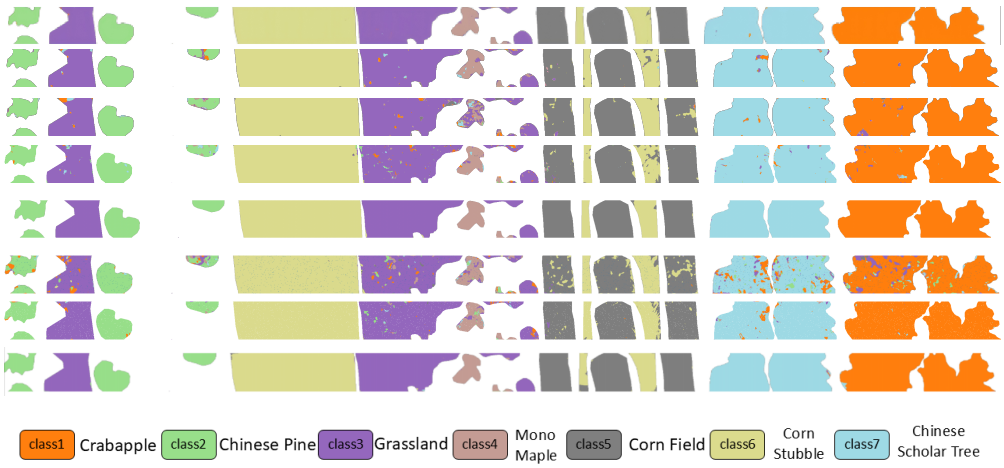


Figure 15. Overview of annotated images and classification performance of SVM, Tri-CNN, HybridSN, SSFTT, MambaHSI, 3DSS-Mamba, SS-Mamba, and the proposed MambaHSINet on the HL_1 dataset.

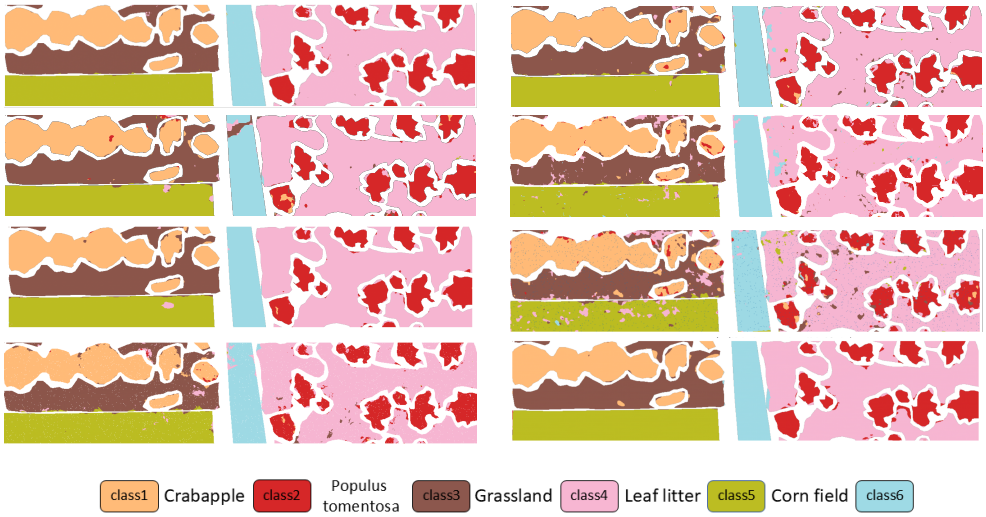


Figure 16. Overview of annotated images and classification performance of SVM, Tri-CNN, HybridSN, SSFTT, MambaHSI, 3DSS-Mamba, SS-Mamba, and the proposed MambaHSINet on the HL_2 dataset.

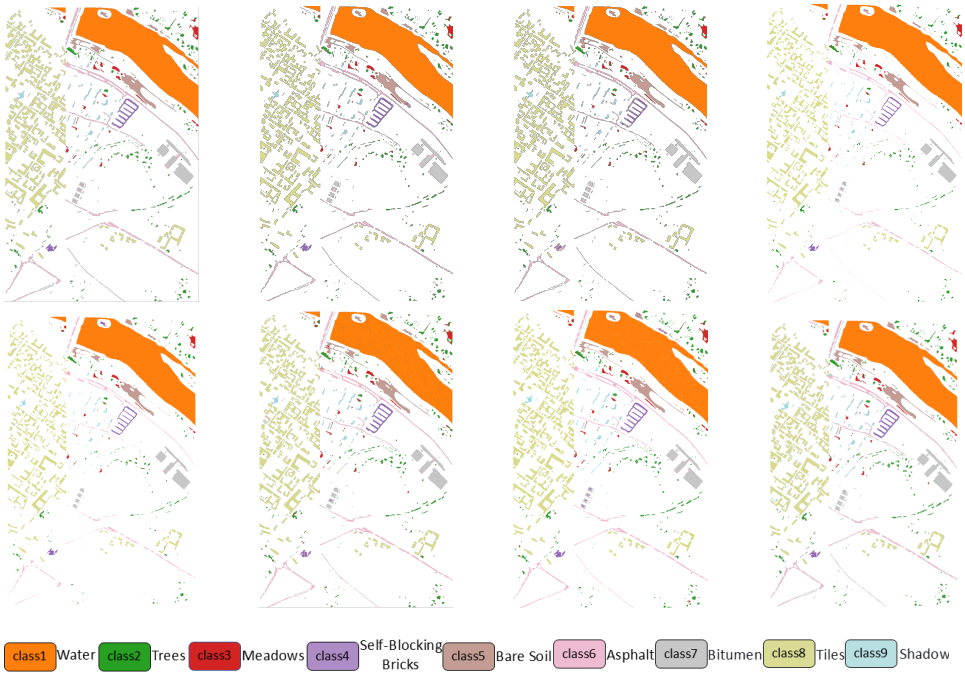


Figure 17. Overview of annotated images and classification performance of SVM, Tri-CNN, HybridSN, SSFTT, MambaHSI, 3DSS-Mamba, SS-Mamba, and the proposed MambaHSINet on the Pavia Centre dataset.

4.6. Computational Performance Comparison

To verify the computational efficiency of the proposed full-image inference strategy, we quantitatively compare the floating point operations (FLOPs), parameter quantity, peak GPU memory usage, training time and testing time among the proposed model and seven comparison methods (SVM, HybridSN, SSFTT, MambaHSI, 3DSS-Mamba, SS-Mamba) on three public hyperspectral datasets, including *HL_1*, *HL_2*, and *Pavia Centre*. For consistent experimental conditions, we compare PyTorch-implemented methods only. Given its different framework, Tri-CNN is excluded from the present evaluation. The detailed results are presented in Table 11.

Table 11. Computational performance comparison across models and datasets.

Dataset	Metric	Model						
		SVM	**HybridSN**	**SSFTT**	**MambaHSI**	**3DSS-Mamba**	**SS-Mamba**	**Proposed**
HL_1	FLOPs (G)	0.04	0.05	<0.01	4.99	0.01	20.93	60.42
	Param (M)	0.09	1.19	0.19	0.02	0.01	0.31	0.12
	Training Time (s)	0.18 ± 0.00	10.37 ± 0.51	2156.89 ± 62.36	217.26 ± 0.80	216.93 ± 0.13	63.77 ± 0.10	944.81 ± 0.92
	Testing Time (s)	49.73 ± 1.66	5.91 ± 0.05	313.50 ± 2.46	0.82 ± 0.01	24.64 ± 0.02	12.76 ± 0.08	0.44 ± 0.02
	Training GPU (GB)	0.38 ± 0.00	0.28 ± 0.00	3.73 ± 0.00	4.19 ± 0.00	0.82 ± 0.00	2.08 ± 0.00	19.85 ± 0.00
	Testing GPU (GB)	0.38 ± 0.00	0.21 ± 0.00	3.73 ± 0.00	4.19 ± 0.00	0.53 ± 0.00	3.02 ± 0.00	0.88 ± 0.00
HL_2	FLOPs (G)	0.03	0.05	<0.01	3.64	0.01	15.18	44.03
	Param (M)	0.09	1.19	0.19	0.02	0.01	0.31	0.12
	Training Time (s)	0.11 ± 0.00	8.94 ± 0.05	1849.85 ± 34.84	162.56 ± 0.24	188.72 ± 0.12	60.63 ± 0.16	674.58 ± 0.00
	Testing Time (s)	26.85 ± 0.73	5.19 ± 0.02	266.96 ± 1.35	2.34 ± 0.02	21.32 ± 0.00	11.19 ± 0.06	0.30 ± 0.01
	Training GPU (GB)	0.33 ± 0.00	0.28 ± 0.00	3.24 ± 0.00	3.06 ± 0.00	0.82 ± 0.00	2.00 ± 0.00	14.49 ± 0.00
	Testing GPU (GB)	0.33 ± 0.00	0.21 ± 0.00	3.24 ± 0.00	3.06 ± 0.00	0.53 ± 0.00	2.94 ± 0.00	0.65 ± 0.00
Pavia	FLOPs (G)	0.05	0.05	<0.01	5.59	0.01	32.87	75.96
	Param (M)	0.06	1.19	0.19	0.02	0.01	0.30	0.09
	Training Time (s)	0.01 ± 0.00	5.10 ± 0.07	1005.37 ± 4.58	363.49 ± 0.35	82.37 ± 7.10	28.26 ± 0.14	1215.89 ± 2.30
	Testing Time (s)	2.43 ± 0.16	2.94 ± 0.05	143.59 ± 0.61	3.93 ± 0.03	14.20 ± 0.17	5.77 ± 0.04	0.50 ± 0.01
	Training GPU (GB)	0.12 ± 0.07	0.28 ± 0.05	3.24 ± 0.00	7.10 ± 0.00	0.82 ± 0.00	1.53 ± 0.00	23.78 ± 0.00
	Testing GPU (GB)	0.12 ± 0.07	0.21 ± 0.05	3.24 ± 0.00	7.10 ± 0.00	0.53 ± 0.00	2.16 ± 0.00	0.97 ± 0.00

In terms of inference efficiency, the proposed method delivers the lowest inference latency on all three datasets. Specifically, the testing time of the proposed model is 0.44 s, 0.30 s, and 0.50 s on *HL_1*, *HL_2*, and *Pavia Centre*, respectively, surpassing all patch-driven Mamba variants, Transformer-based SSFTT, and traditional machine learning SVM. The significant inference speed gain stems from the full-image end-to-end inference pipeline. Unlike patch-based competitors that require sliding window traversal and redundant overlapping region computation, our paradigm eliminates repeated feature extraction for overlapping spatial patches, which fundamentally reduces inference latency. In terms of model lightweightness, the proposed model maintains a tiny parameter scale ranging from 0.09 M to 0.12 M on all datasets. Its parameter volume is only approximately 10% of HybridSN (1.19 M) and 30% of SS-Mamba (∼0.30 M), demonstrating prominent lightweight characteristics. For inference-side GPU memory usage, the proposed model occupies merely 0.65–0.97 GB across all datasets. Notably, it outperforms SVM and MambaHSI in testing memory overhead, and exhibits competitive memory efficiency compared with other lightweight Mamba-based backbones.

Nevertheless, the proposed method suffers from higher computational overhead during the training phase, which constitutes its primary limitation. First, it has the largest training FLOPs among all compared methods: 60.42 G for *HL_1*, 44.03 G for *HL_2*, and 75.96 G for *Pavia Centre*. The excessive FLOPs directly lead to prolonged training durations, which reach 944.81 s, 674.58 s, and 1215.89 s for the three datasets. Second, the proposed model requires considerable peak GPU memory during training, with recorded values of 19.85 GB, 14.49 GB and 23.78 GB for the three datasets. This high hardware requirement restricts deployment on low-end edge computing devices with limited GPU memory. Two intrinsic structural factors cause this training overhead: the bidirectional state space modeling module involves dual-directional sequence iteration, and full-image batch training requires loading complete hyperspectral cubes into GPU memory without patch segmentation.

Overall, benefiting from the full-image inference framework and bidirectional Mamba-based state space modeling, the proposed model achieves ultra-low inference latency, small parameter footprint and limited inference GPU memory consumption. The elevated training computational complexity and memory usage are acceptable trade-offs for remote sensing offline training scenarios, where models are trained once and deployed repeatedly.

4.7. Generalization Experiments

We further test the generalization ability of MambaHSINet on two public hyperspectral datasets, Houston and WHU-Hi-LongKou. A 10% training ratio is adopted for the *Houston* dataset and 1% for *WHU-Hi-LongKou*. The quantitative results and visual classification maps are presented in Table 12 and Figures 18 and 19.

Table 12. Generalization experiments on Houston and WHU-Hi-LongKou datasets.

Dataset	Houston	WHU-Hi-LongKou
OA	94.25 ± 1.08	98.83 ± 0.43
AA	94.88 ± 1.16	98.34 ± 0.35
Kappa	90.55 ± 1.67	98.47 ± 0.56

The model achieves excellent performance on *WHU-Hi-LongKou*, with OA, AA, and Kappa all exceeding 98%. It also obtains competitive results on the *Houston* dataset despite its complex surface features and obvious spectral aliasing.

Visually, each land-cover class forms complete, well-defined regions. No large-area misclassification is observed, and only a few sporadic misclassified pixels appear in complex local areas. These results verify that the proposed network possesses good generalization

performance. MambaHSINet can adapt effectively to diverse hyperspectral data and is applicable to a wide range of hyperspectral analysis tasks beyond tree species classification.

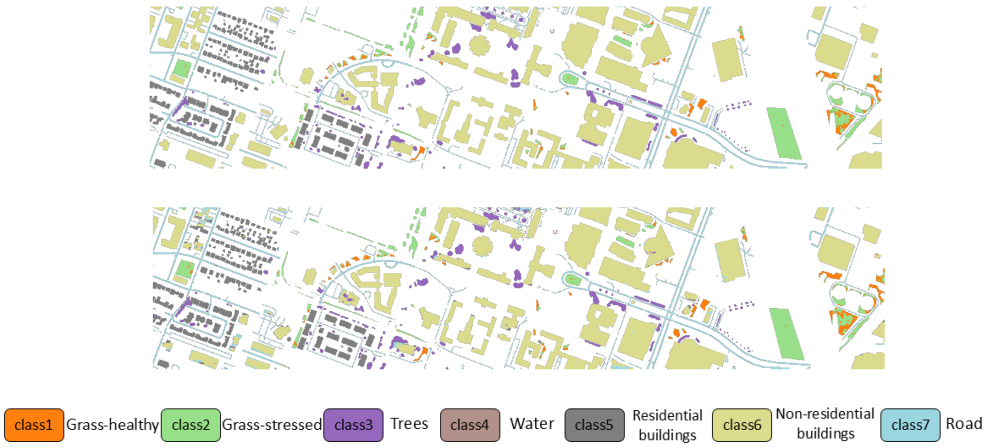


Figure 18. Overview of annotated images and classification performance of MambaHSINet on the Houston dataset.



Figure 19. Overview of annotated images and classification performance of MambaHSINet on the WHU-Hi-LongKou dataset.

5. Discussion

5.1. Core Innovations and Technical Superiority

This section systematically clarifies the methodological novelties and technical advantages of the proposed MambaHSINet, by conducting a detailed comparative analysis against three representative Mamba-based hyperspectral classification methods: MambaHSI, 3DSS-Mamba and SS-Mamba. As summarized in Table 13, existing Mamba variants for HSI classification generally adopt patch-based input and sliding-window inference, combined with mixed spectral-spatial feature extraction or unidirectional state space modeling (SSM). Distinct from these pipelines, MambaHSINet integrates three mutually reinforced core innovations, namely full-image end-to-end inference, decoupled multi-scale spectral-spatial dual-branch architecture, and bidirectional S6 global modeling. These targeted designs not only differentiate our method from existing Mamba-based works, but also make it specially tailored for the challenging task of fine-grained tree species classification, as well as general HSIC tasks.

First, the full-image input and end-to-end inference paradigm is a fundamental distinction from all compared Mamba methods. MambaHSI, 3DSS-Mamba and SS-Mamba all follow the conventional patch-based pipeline: they crop small local patches from raw HSI data for training and adopt sliding-window inference in the testing phase. This paradigm inevitably produces massive overlapping calculations and fails to utilize the global contextual information of the entire scene. In contrast, MambaHSINet directly takes the complete hyperspectral cube as input and outputs full-resolution pixel-wise classification results in a single forward pass. It thoroughly eliminates redundant computation caused by patch cropping and sliding traversal, and maximizes the utilization of full-scene global context. This design is particularly critical for UAV-borne hyperspectral tree species classification, where forest and farmland present large-scale continuous distribution and long-range spatial correlations across the whole landscape.

Second, we propose a decoupled multi-scale spectral-spatial dual-branch structure with residual connections, which differs essentially from the mixed feature extraction schemes of existing Mamba methods. Specifically, SS-Mamba adopts a mixed spectral-spatial feature learning manner, which fuses spectral and spatial information in an entangled way and causes mutual interference between the two types of features.MambaHSI mainly focuses on multi-scale spatial modeling and lacks an independent, dedicated branch for spectral feature mining. 3DSS-Mamba leverages 3D convolution to jointly extract spectral-spatial features, which is prone to overfitting when facing spectrally similar tree species.Different from the above designs, our model splits spectral and spatial feature learning into two independent branches. The spectral branch uses multi-scale 1D convolutions combined with residual paths to excavate subtle spectral differences between similar tree species, while the spatial branch employs multi-scale 2D convolutions and residual connections to capture complex local textures and boundary features of forest scenes. The decoupled design avoids feature interference, and the multi-scale convolution plus residual structure preserves original spectral signatures and spatial details. This addresses the core difficulty of fine-grained tree species classification: highly overlapping spectral curves and similar spatial structures among adjacent tree species.

Table 13. Architectural comparison between MambaHSINet and SOTA Mamba methods.

Model	Input Paradigm	Feature Extraction	Context Scope	Computational Complexity	Inference Mode	Context Modeling Module
SS-Mamba	Patch-based	Mixed Spectral-Spatial	Patch-limited	Linear ($O(N)$)	Sliding window	Unidirectional SSM
MambaHSI	Patch-based	Multi-scale Spatial	Local-scale	Linear ($O(N)$)	Sliding window	Unidirectional SSM
3DSS-Mamba	Patch-based	3D Conv + SSM	Local-scale	Linear ($O(N)$)	Sliding window	Unidirectional SSM
MambaHSINet	Full-image	Decoupled Dual-branch	Full-image Global	Linear ($O(N)$)	End-to-end full-image	Bidirectional SSM

Third, the bidirectional S6 scanning global modeling module is another key technical improvement over existing unidirectional Mamba architectures. MambaHSI, 3DSS-Mamba and SS-Mamba all adopt unidirectional SSM for sequence modeling. Such causal sequence scanning can only capture one-way contextual dependencies, and cannot fully characterize the non-causal, bidirectional spatial-spectral correlations inherent in hyperspectral data. In this work, we adopt two independent S6 modules to implement forward and reverse sequence scanning respectively. The bidirectional mechanism complements missing contextual information from single-direction modeling and achieves comprehensive full-image long-range dependency learning under linear $O(N)$ computational complexity. Compared with unidirectional SSM, it better models the bidirectional spatial adjacency of tree groups and the continuous variation of spectral bands, further boosting the discriminative power for fine-grained tree categories.

Collectively, the above three core modules form an integrated solution rather than simple combination of existing components. The full-image paradigm breaks the limitations of patch-based pipelines in global context utilization and inference efficiency. The decoupled dual-branch multi-scale structure is customized to separate spectral and spatial features, targeting the spectral similarity problem of tree species. The bidirectional S6 module enhances global long-range modeling capability while retaining the linear complexity advantage of Mamba.

A series of experimental results verify the superiority of the proposed method. On multiple datasets covering forest, urban and farmland scenarios, MambaHSINet achieves competitive classification accuracy comparable to MambaHSI, while delivering faster inference speed. It also outperforms 3DSS-Mamba and SS-Mamba in overall accuracy, average accuracy and classification stability. Beyond competitive performance on general HSI classification tasks and strong generalization ability across conventional hyperspectral classification scenarios, the proposed framework shows excellent adaptability to fine-grained tree species classification. It effectively addresses two core challenges in this task: subtle spectral differences between tree species and highly similar spatial textures. Meanwhile, its lightweight parameter design and ultra-fast full-image inference enable great practical value for UAV forest monitoring and tree species mapping, which stands out as a prominent application advantage over existing patch-based Mamba methods.

5.2. Discussion on Experimental Fairness Between Full-Image and Patch-Based Paradigms

Different from mainstream patch-based hyperspectral classification methods such as HybridSN and SSFTT, the proposed MambaHSINet adopts a full-image end-to-end training pipeline. During model training, the complete hyperspectral cube is fed into the network as the sole input, which contains training pixels, test pixels and background pixels. A binary mask is adopted in the training phase: only labeled training pixels are used for weighted cross-entropy loss calculation and network parameter backpropagation, while test pixels do not participate in parameter updating. Nevertheless, in the forward propagation process, all pixels in the entire image (including test pixels) pass through convolution layers and the bidirectional Mamba module, and their spectral-spatial contextual information is incorporated into global feature modeling.The inherent differences between the two input paradigms lead to inconsistent experimental conditions and affect the fairness of comparison, which are elaborated as follows:

Differences in the utilization of contextual information: Patch-based methods strictly follow the inductive learning paradigm. During training, they only crop independent local patches from training areas, and data from test areas are completely invisible to the model. In contrast, the full-image scheme in this work enables the model to perceive the global spectral-spatial distribution of the entire image. The bidirectional Mamba module further explores long-range dependencies across all pixels, allowing the model to acquire global prior information that patch-based methods cannot access. Hence, the proposed method exhibits characteristics of transductive learning.

Differences in receptive field scope: Restricted by fixed patch sizes, patch-based methods only have effective receptive fields limited to local regions. By contrast, full-image input naturally endows the model with full-scene global perception capability, which is an inherent advantage of the proposed framework for intra-scene classification tasks.

To minimize the unfairness caused by paradigm differences, we unified all experimental variables in this study. All methods adopt identical dataset partitioning rules, fixed random seeds, stratified sampling ratios, data preprocessing pipelines and evaluation metrics. All comparative algorithms use official open-source codes and optimal hyperparameters, with consistent software and hardware experimental environments.

The comparative experiments in this work are not merely conducted for fair benchmarking under standard inductive learning constraints. Instead, they focus on evaluating the comprehensive performance of different frameworks in practical engineering applications. The core goal of the full-image design is to eliminate redundant calculation caused by patch-based sliding-window inference and accelerate inference speed, which is the core innovation of this paper. In future research, we will adopt cross-scene training and regional masking strategies to eliminate transductive learning bias and realize fair comparison under standard inductive learning settings.

5.3. Limitations and Future Work

Despite the promising performance of MambaHSINet, two limitations remain to be addressed. Firstly, the full-image training mode and bidirectional SSM lead to high computational overhead. The model presents large training FLOPs, long training duration, and high peak GPU memory usage. Such high hardware requirements restrict its application on low-end devices and bring challenges for training on ultra-large-scale HSIs. Secondly, the full-image input scheme may introduce transductive learning bias. Different from conventional patch-based inductive methods, MambaHSINet feeds the entire HSI into the network during training. Although loss calculation is only performed on labeled pixels, convolution and Mamba sequence scanning inevitably exploit spatial and spectral context from unlabeled test pixels. This characteristic makes the model deviate from standard inductive inference and tends toward transductive learning, which may affect its pure generalization ability in fully unseen scenes.

For future research, we will optimize the network structure and training strategies to reduce training computation and GPU memory occupancy, so as to adapt to large-scale hyperspectral data and edge computing devices. Moreover, we will improve the full-image input mechanism to weaken transductive learning bias and enhance the pure inductive generalization capability of the model. To achieve this, we will also draw inspiration from and compare with other advanced architectures, including efficient attention mechanisms [40–42], generalized state space models [43,44], classical morphological classifiers [45], and the latest tokenized networks [46].

6. Conclusions

This paper proposes MambaHSINet, a bidirectional SSM-based dual-branch network for full-image hyperspectral classification. With full-image end-to-end inference, decoupled spectral–spatial feature extraction, and linear-complexity global modeling, our approach overcomes redundant computation and weak long-range modeling of traditional patch-based Mamba methods. Evaluated on diverse hyperspectral datasets, the model achieves high classification accuracy, strong stability, lightweight parameters, fast inference, and low inference memory usage, and generalizes well across different scenes. We also release two well-annotated subsets of the *ZJK* UAV-borne hyperspectral tree species dataset. Covering typical forest-farmland transition areas, these subsets enrich labeled data for fine-grained forest classification and act as useful benchmarks for UAV forest monitoring.

Despite its competitive performance, the model still has two major limitations. Full-image training leads to heavy computation and high GPU memory occupancy, restricting its use on low-end devices and large-scale data. Meanwhile, full-image input may cause transductive learning bias. In future work, we will optimize training strategies to reduce hardware consumption and mitigate such bias.

Author Contributions: Conceptualization, Y.Z. and X.L. (Xinying Liu); methodology, X.L. (Xinwei Li); software, Y.Z. and T.C.; validation, Y.Z., T.C. and J.W.; formal analysis, Y.Z.; investigation, Y.L. and X.W.; resources, X.L. (Xinwei Li); data curation, X.L. (Xinying Liu); writing—original draft

preparation, X.L. (Xinying Liu); writing—review and editing, X.L. (Xinwei Li), Y.Z. and X.L. (Xinying Liu); visualization, X.L. (Xinying Liu); supervision, X.L. (Xinwei Li); project administration, X.L. (Xinwei Li), Y.Z. and X.L. (Xinying Liu). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Innovation and Entrepreneurship Training Program for Undergraduates of Beijing Forestry University, grant number X202510022381.

Data Availability Statement: The source code of this paper will be made available at https://github.com/YFENG-123/mambaHSI (accessed on 14 June 2026) after the publication of the paper.

Acknowledgments: We sincerely thank Azure Aviation Technology Ltd. for providing UAV flight services and completing raw data stitching. We also appreciate Wenqi Dong, Huilin Feng, and Junxin Zhao for their dedicated work on field surveys and image acquisition. The *ZJK HL_1* and *HL_2* datasets are available on the cloud disk. Interested researchers may contact the corresponding author to obtain access to the data.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zhong, L.; Dai, Z.; Fang, P.; Cao, Y.; Wang, L. A Review: Tree Species Classification Based on Remote Sensing Data and Classic Deep Learning-Based Methods. *Forests* **2024**, *15*, 852. [CrossRef]
2. Ni-Meister, W.; Albanese, A.; Lingo, F. Assessing Data Preparation and Machine Learning for Tree Species Classification Using Hyperspectral Imagery. *Remote Sens.* **2024**, *16*, 3313. [CrossRef]
3. Zhao, X.; Qi, J.; Huang, H. Sensitivity Testing Analysis of Airborne Hyperspectral Lidar Signals for Monitoring Insects and Diseases Based on 3D Radiative Transfer Model. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Pasadena, CA, USA, 16–21 July 2023; pp. 4613–4616.
4. Goetz, A. Three Decades of Hyperspectral Remote Sensing of the Earth: A Personal View. *Remote Sens. Environ.* **2009**, *113*, S5–S16. [CrossRef]
5. Dalponte, M.; Ørka, H.; Gobakken, T.; Gianelle, D.; Næsset, E. Tree Species Classification in Boreal Forests with Hyperspectral Data. *IEEE Trans. Geosci. Remote Sens.* **2012**, *51*, 2632–2645. [CrossRef]
6. Yang, R.; Kan, J. Classification of Tree Species in Different Seasons and Regions Based on Leaf Hyperspectral Images. *Remote Sens.* **2022**, *14*, 1524. [CrossRef]
7. Hou, C.; Liu, Z.; Chen, Y.; Wang, S.; Liu, A. Tree Species Classification from Airborne Hyperspectral Images Using Spatial–Spectral Network. *Remote Sens.* **2023**, *15*, 5679. [CrossRef]
8. Clark, M.L.; Roberts, D.A.; Clark, D.B. Hyperspectral Discrimination of Tropical Rain Forest Tree Species at Leaf to Crown Scales. *Remote Sens. Environ.* **2005**, *96*, 375–398. [CrossRef]
9. Dalponte, M.; Bruzzone, L.; Gianelle, D. Tree Species Classification in the Southern Alps Based on the Fusion of Very High Geometrical Resolution Multispectral/Hyperspectral Images and LiDAR Data. *Remote Sens. Environ.* **2012**, *123*, 258–270. [CrossRef]
10. Immitzer, M.; Atzberger, C.; Koukal, T. Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data. *Remote Sens.* **2012**, *4*, 2661–2693. [CrossRef]
11. Yel, S.G.; Tunc Gormus, E. Exploiting Hyperspectral and Multispectral Images in the Detection of Tree Species: A Review. *Front. Remote Sens.* **2023**, *4*, 1136289. [CrossRef]
12. Melgani, F.; Bruzzone, L. Classification of Hyperspectral Remote Sensing Images with Support Vector Machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
13. Mountrakis, G.; Im, J.; Ogole, C. Support Vector Machines in Remote Sensing: A Review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]
14. Tong, F.; Zhang, Y. Spectral–Spatial and Cascaded Multilayer Random Forests for Tree Species Classification in Airborne Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4411711. [CrossRef]
15. Chen, Y.; Nasrabadi, N.M.; Tran, T.D. Sparse Representation for Target Detection in Hyperspectral Imagery. *IEEE J. Sel. Top. Signal Process.* **2011**, *5*, 629–640. [CrossRef]
16. Hu, W.; Huang, Y.; Wei, L.; Zhang, F.; Li, H. Deep Convolutional Neural Networks for Hyperspectral Image Classification. *J. Sens.* **2015**, *2015*, 258619. [CrossRef]

17. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6232–6251. [CrossRef]
18. Li, W.; Wu, G.; Zhang, F.; Du, Q. Hyperspectral Image Classification Using Deep Pixel-Pair Features. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 844–853. [CrossRef]
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
20. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5518615. [CrossRef]
21. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral–Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5522214. [CrossRef]
22. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In Proceedings of the First Conference on Language Modeling, Pennsylvania, PA, USA, 7–6 October 2024.
23. He, Y.; Tu, B.; Liu, B.; Li, J.; Plaza, A. 3DSS-Mamba: 3D-Spectral-Spatial Mamba for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5534216. [CrossRef]
24. Liang, L.; Zhang, J.; Duan, P.; Kang, X.; Wu, T.X.; Li, J.; Plaza, A. LKMA: Learnable Kernel and Mamba with Spatial-Spectral Attention Fusion for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5530914. [CrossRef]
25. Li, Y.; Luo, Y.; Zhang, L.; Wang, Z.; Du, B. MambaHSI: Spatial–Spectral Mamba for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5524216. [CrossRef]
26. Zhou, Z.; Guo, X.; Xiong, Y.; Xia, C. Kalman-SSM: Modeling Long-Term Time Series with Kalman Filter Structured State Spaces. *IEEE Signal Process. Lett.* **2024**, *31*, 2470–2474. [CrossRef]
27. Huang, L.; Chen, Y.; He, X. Spectral–Spatial Mamba for Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 2449. [CrossRef]
28. Liao, J.; Wang, L. HyperspectralMamba: A Novel State Space Model Architecture for Hyperspectral Image Classification. *Remote Sens.* **2025**, *17*, 2577. [CrossRef]
29. Liu, Q.; Yue, J.; Fang, Y.; Xia, S.; Fang, L. HyperMamba: A Spectral-Spatial Adaptive Mamba for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5536514. [CrossRef]
30. Peng, H.; Lin, K.; Liu, H. HS-Mamba: Full-Field Interaction Multi-Groups Mamba for Hyperspectral Image Classification. *arXiv* **2025**, arXiv:2504.15612 .
31. Ahmad, M.; Usama, M.; Mazzara, M.; Distefano, S. WaveMamba: Spatial-Spectral Wavelet Mamba for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *22*, 5500505.
32. Wang, G.; Zhang, X.; Peng, Z.; Zhang, T.; Jiao, L. S^2 Mamba: A Spatial-Spectral State Space Model for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5511413. [CrossRef]
33. Pan, Z.; Li, C.; Plaza, A.; Chanussot, J.; Hong, D. Hyperspectral Image Classification with Mamba. *IEEE Trans. Geosci. Remote Sens.* **2024**, *63*, 5602814. [CrossRef]
34. Liang, L.; Xie, P.; Zhang, Y.; Li, J.; Zhang, Z.; Li, J.; Plaza, A. DBMLLA: Double-Branch Mamba-Like Linear Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5524315. [CrossRef]
35. Alkhatib, M.; Al-Saad, M.; Aburaed, N.; Almansoori, S.; Zabalza, J.; Marshall, S.; Al-Ahmad, H. Tri-CNN: A Three Branch Model for Hyperspectral Image Classification. *Remote Sens.* **2023**, *15*, 316. [CrossRef]
36. Jiang, X.; Han, C.; Mesgarani, N. Dual-Path Mamba: Short and Long-Term Bidirectional Selective Structured State Space Models for Speech Separation. In Proceedings of the ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 6–11 April 2025; pp. 1–5.
37. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [CrossRef]
38. Roy, S.; Krishna, G.; Dubey, S.; Chaudhuri, B. HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 277–281.
39. Li, J.; Li, C.; Wang, S.; Yan, J.; Fu, H.; Gao, L.; Liao, W. SS-Mamba: A Spectral-Spatial Mamba-Based Learning Framework for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14 . [CrossRef]
40. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A Survey. *ACM Comput. Surv.* **2022**, *55*, 1–28. [CrossRef]
41. Roy, S.; Deria, A.; Shah, C.; Haut, J.; Du, Q.; Plaza, A. Spectral–Spatial Morphological Attention Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5503615. [CrossRef]
42. Sun, H.; Zheng, X.; Lu, X.; Wu, S. Spectral–Spatial Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3232–3245. [CrossRef]
43. Wang, J.; Zhu, W.; Wang, P.; Yu, X.; Liu, L.; Omar, M.; Hamid, R. Selective Structured State-Spaces for Long-Form Video Understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 6387–6397.

44. Dao, T.; Gu, A. Transformers Are SSMs: Generalized Models and Efficient Algorithms through Structured State Space Duality. *arXiv* **2024**, arXiv:2405.21060.
45. Fauvel, M.; Benediktsson, J.; Chanussot, J.; Sveinsson, J. Spectral and Spatial Classification of Hyperspectral Data Using SVMs and Morphological Profiles. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 3804–3814. [CrossRef]
46. Zhou, W.; Kamata, S.; Wang, H.; Wong, M.; Hou, H. Mamba-in-Mamba: Centralized Mamba-Cross-Scan in Tokenized Mamba Model for Hyperspectral Image Classification. *Neurocomputing* **2025**, *613*, 128751. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.