*remote sensing*

MDPI

*Article*

# MRFP-Mamba: Multi-Receptive Field Parallel Mamba for Hyperspectral Image Classification

**Xiaofei Yang [1], Lin Li [1], Suihua Xue [1], Sihuan Li [1], Wanjun Yang [1,*], Haojin Tang [1] and Xiaohui Huang [2]**

[1] School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China; xiaofeiyang@gzhu.edu.cn (X.Y.); 2112330037@e.gzhu.edu.cn (L.L.); 1919500073@e.gzhu.edu.cn (S.X.); 2112330082@e.gzhu.edu.cn (S.L.); tanghaojin@gzhu.edu.cn (H.T.)

[2] School of Information Engineering, East China Jiaotong University, Nanchang 330044, China; 2854@ecjtu.edu.cn

* Correspondence: wanjunyang@gzhu.edu.cn

**Abstract**

Deep learning has achieved remarkable success in hyperspectral image (HSI) classification, attributed to its powerful feature extraction capabilities. However, existing methods face several challenges: Convolutional Neural Networks (CNNs) are limited in modeling long-range spectral dependencies because of their limited receptive fields; Transformers are constrained by their quadratic computational complexity; and Mamba-based methods fail to fully exploit spatial–spectral interactions when handling high-dimensional HSI data. To address these limitations, we propose MRFP-Mamba, a novel Multi-Receptive-Field Parallel Mamba architecture that integrates hierarchical spatial feature extraction with efficient modeling of spatial–spectral dependencies. The proposed MRFP-Mamba introduces two key innovation modules: (1) A multi-receptive-field convolutional module employing parallel $1 \times 1$, $3 \times 3$, $5 \times 5$, and $7 \times 7$ kernels to capture fine-to-coarse spatial features, thereby improving discriminability for multi-scale objects; and (2) a parameter-optimized Vision Mamba branch that models global spatial–spectral relationships through structured state space mechanisms. Experimental results demonstrate that the proposed MRFP-Mamba consistently surpasses existing CNN-, Transformer-, and state space model (SSM)-based approaches across four widely used hyperspectral image (HSI) benchmark datasets: PaviaU, Indian Pines, Houston 2013, and WHU-Hi-LongKou. Compared with MambaHSI, our MRFP-Mamba achieves improvements in Overall Accuracy (OA) by 0.69%, 0.30%, 0.40%, and 0.97%, respectively, thereby validating its superior classification capability and robustness.

**Keywords:** hyperspectral image classification; Mambas; deep learning



Academic Editors: Pedro Melo-Pinto and Wen Yang

Received: 21 April 2025
Revised: 9 June 2025
Accepted: 24 June 2025
Published: 26 June 2025

**Citation:** Yang, X.; Li, L.; Xue, S.; Li, S.; Yang, W.; Tang, H.; Huang, X. MRFP-Mamba: Multi-Receptive Field Parallel Mamba for Hyperspectral Image Classification. *Remote Sens.* **2025**, *17*, 2208. https://doi.org/10.3390/rs17132208



## 1. Introduction

Hyperspectral images capture both rich spatial information and continuous spectral data across multiple bands, facilitating the identification of subtle feature differences for applications in precision agriculture [1], environmental monitoring [2], urban planning [3], and geological exploration [4–6]. However, the high data dimensionality, limited training samples, and need for efficient spatial–spectral integration pose challenges in fully leveraging this rich information.

HSI classification aims to divide pixels into categorical labels by jointly leveraging spatial context and spectral information. The traditional methods bifurcate into two

paradigms: spectral-only classifiers (e.g., Support Vector Machines (SVMs) [7,8], Random Forest (RF) [9], and k-Nearest Neighbors (KNNs) [10–13] and shallow hybrid models (e.g., Principal Component Analysis (PCA) [14,15] and Linear Discriminant Analysis (LDA) [16,17]. While these methods achieve baseline performance, they suffer from three critical flaws: (1) neglect of spatial information, leading to misclassification of spectrally similar classes; (2) vulnerability to the "curse of dimensionality" due to redundant spectral bands; and (3) poor generalization under limited training samples.

Recently, deep learning techniques have achieved significant progress in hyperspectral image (HSI) classification. There are many Convolutional Neural Network (CNN)-based methods for HSI classification [18,19]. For example, Yang et al. [20] employed a fusion strategy combining 2D-CNN and 3D-CNN [21] to effectively extract deep features from limited HSI samples. Roy et al. [7] introduced HybridSN, which integrates 3D-CNN for local spatial–spectral feature extraction and further leverages 2D-CNN to model deep hierarchical features, thereby improving classification accuracy. These CNN-based methods advanced the field by extracting joint spatial–spectral features, yet their localized receptive fields hinder long-range dependency capture.

Transformer-based architectures addressed this via self-attention mechaisms to model the long-range dependencies. For example, He et al. [22] applied Transformer structures to HSI classification and proposed the SSF model, which integrates CNNs for local spatial feature extraction while utilizing Transformers for spectral sequence modeling. Graham et al. [23] incorporated convolutional operations before the Transformer module to extract local features before performing global modeling. Moreover, Mei et al. [24] introduced the GAHT architecture, which combines CNNs and Transformers to capture local dependencies between spectral channels and employs a hierarchical Transformer for feature representation. Xu et al. [25] proposed the Dual Selective Fusion Transformer Network (DSFormer), which adaptively selects and fuses spatial–spectral features from multiple receptive fields to enable joint modeling, thereby significantly improving classification accuracy across several benchmark datasets. Cheng et al. [26] developed CACFTNet, which integrates a covariance attention mechanism with cross-layer fusion strategies to enhance feature extraction capability, demonstrating outstanding performance especially in complex land-cover scenarios. Although Transformers excel at capturing long-range dependencies, their computational complexity increases quadratically with the input sequence length. This poses substantial challenges in terms of resource consumption when processing high-dimensional hyperspectral data. The resulting high computational cost limits their scalability in practical applications, making deployment particularly difficult in real-time processing scenarios or environments with constrained hardware resources.

Now, Mamba-based models emerge as efficient alternatives with linear scalability. Mamba integrates convolutional operations with state space modeling (SSM) modules, enabling efficient long-range dependency capture while maintaining significantly lower computational complexity compared to Transformers, making it particularly well-suited for long-sequence modeling tasks. For example, SpectralMamba [27] focuses on spectral dimension modeling, whereas 3DSS-Mamba incorporates both spatial and spectral information, further enhancing the capability of SSM-based approaches in processing hyperspectral data. However, the current Mamba models still face two major challenges in the field of hyperspectral image (HSI) classification.

1. **Spatial–spectral Decoupling in Mamba Variants:** While these methods excel at modeling sequential spectral patterns (e.g., distinguishing subtle reflectance variations between vegetation species), they inadequately integrate spatial hierarchies—the multi-scale geometric and contextual relationships between pixels.

2. **Scale Sensitivity of Single-Receptive-Field Convolutions:** Current HSI classification methods rely on fixed-size convolutional kernels or attention windows, which restrict their ability to capture multi-granular spatial features.

To address the challenges of spatial–spectral decoupling and scale sensitivity in existing Mamba-based hyperspectral image (HSI) classification methods, we propose MRFP-Mamba, a novel Multi-Receptive-Field Parallel Mamba architecture that integrates hierarchical spatial feature extraction with efficient modeling of spatial–spectral dependencies. This architecture combines the strengths of multi-scale convolutional operations and structured state space modeling to achieve adaptive spatial feature extraction and long-range spectral dependency capture, enhancing discriminative power for complex HSI classification tasks. Specifically, the proposed MRFP-Mamba comprises two core innovative modules: (1) a multi-receptive-field convolutional module (MRFCM) to capture fine-grained details (via smaller kernels) and coarse-grained contextual information (via larger kernels), effectively addressing scale sensitivity in single-receptive-field convolutions; (2) a parameter-optimized Vision Mamba branch to model global spatial–spectral dependencies. The contributions are listed as follows:

1. We propose MRFP-Mamba, a hierarchical architecture that combines multi-receptive-field convolutional feature extraction with a parameter-optimized Vision Mamba, enabling adaptive capture of multi-scale spatial features and efficient modeling of global spectral dependencies.
2. We introduce a multi-receptive-field convolutional module to extract hierarchical spatial features, addressing scale sensitivity issues in single-kernel convolutions. This module captures fine-grained details and coarse contextual information simultaneously, improving representation of multi-scale objects and spatial–spectral interactions.
3. We design a parameter-optimized Vision Mamba branch that models long-range spectral dependencies across bands, enabling effective fusion of local spatial hierarchies and global spectral correlations.
4. Extensive experiments on four HSI datasets demonstrate that MRFP-Mamba outperforms state-of-the-art HSI classification methods, achieving significant improvements in Overall Accuracy (OA).

The remainder of this paper is organized as follows. Section 2 reviews related work on hyperspectral image classification based on CNNs, Transformers, and Mamba models. Section 3 presents the proposed MRFP-Mamba model and its key components. Section 4 introduces the four benchmark HSI datasets, describes the experimental setup, and provides the results and analysis. Finally, Section 5 concludes the paper.

## 2. Related Works

HSI classification methods aim to integrate the rich spatial and spectral information to distinguish subtle land cover differences. This section reviews three key paradigms: Convolutional Neural Networks (CNNs), Transformers, and Mamba-based HSI classification models.

### *2.1. Convolution Neural Networks for Hyperspectral Image Classification*

Convolutional Neural Networks (CNNs) have been widely applied in hyperspectral image (HSI) analysis due to their strong capability for local feature extraction [28–31]. Traditional 2D-CNN architectures, such as the design in [28,31,32], process spatial and spectral dimensions separately, achieving moderate classification accuracy but failing to model inter-band correlations effectively. To address this limitation, 3D-CNNs like [7,31], were introduced, using volumetric convolutions to jointly extract spatial–spectral features from 3D data cubes. For example, HybridSN [7], which combines 3D convolutions for

local spatial–spectral feature extraction with 2D convolutions for hierarchical spatial refinement, improves performance on datasets like Indian Pines. However, these methods face two fundamental limitations. First, 2D-CNNs are constrained by limited receptive fields, leading to misclassifications in scenarios where distant spatial contexts are critical due to insufficient modeling of spatial texture differences. Second, 3D-CNNs suffer from excessive computational complexity, as their parameter counts grow cubically with input dimensions, making them impractical for large HSIs with hundreds of spectral bands.

RNNs have been applied to HSI classification for their ability to model spectral continuity in sequential data. Hang et al. [33] proposed a cascaded RNN architecture aimed at reducing redundancy between adjacent spectral bands and improving feature representation capability. Mei et al. [34] combined CNNs and RNNs, constructing a spatial–spectral fusion framework where CNNs extract spatial features while RNNs capture spectral dependencies. Additionally, an increasing number of deep learning architectures, such as Fully Convolutional Networks (FCNs), Generative Adversarial Networks (GANs), and Graph Convolutional Networks (GCNs), have been explored for HSI classification, each enhancing feature modeling capabilities to varying degrees.

Despite these advances, deep learning-based HSI classification still faces several challenges. For instance, RNNs struggle with capturing long-range dependencies, limiting their ability to model distant spectral relationships effectively. Meanwhile, CNNs are constrained by their limited receptive fields, making them less effective at extracting long-range spatial–spectral dependencies. These challenges restrict further improvement of existing methods in HSI classification. Therefore, it remains crucial to explore more efficient architectures that can optimize feature extraction strategies and achieve deeper fusion of spatial and spectral information.

### *2.2. Transformer Networks for Hyperspectral Image Classification*

Transformers have emerged as a transformative approach in HSI classification [22,35–38], utilizing self-attention mechanisms to model global dependencies across spatial and spectral domains. Early works like SSF integrate CNNs for local feature extraction with transformers for spectral sequence modeling, demonstrating improved accuracy on small-scale datasets. More recently, advanced architectures such as SSFTT [39] and GAHT [24] have refined this paradigm: SSFTT decomposes HSI data into spectral and spatial tokens for independent modal refinement before fusion while GAHT employs grouped pixel embedding to constrain self-attention within local spectral contexts, reducing computational redundancy. These models integrate various innovative components that not only enhance the learning of local spectral features but also optimize the alignment between deep semantic features and sample distributions while improving the efficiency of skip connections. Leveraging these techniques, Transformer-based models have demonstrated superior performance in HSI classification compared to conventional Transformers, highlighting their significant application potential.

Although existing Transformer networks can effectively capture long-range dependencies, they often overlook local features. To address this issue, Wu et al. [40] proposed a novel Transformer network, Convolutional Vision Transformer (CvT), which integrates convolutional operations and self-attention mechanisms within its block structure. In this design, convolutional operations are responsible for extracting local features, while the self-attention mechanism captures global representations. However, these Transformer models only employ convolutional operations for local feature extraction and fail to fully exploit the complementary advantages of CNNs and Transformers.

To bridge this gap, Yang et al. [20] incorporated CNNs into Transformers and proposed a novel Transformer-based network, HSI Transformer (HiT). Zhou et al. [41] proposed the Dual-Branch Convolutional Transformer Network (DCTN), which employs a dual-branch

structure to separately process spatial and spectral information. It integrates a spatial–spectral Fusion Projection Module (SFPM) and an Efficient Interactive Self-Attention (EISA) mechanism to enhance feature extraction and fusion, leading to improved classification performance across multiple datasets. Although transformers excel at encoding global contextual information, their scalability limitations in high-dimensional HSI spaces necessitate a shift toward more efficient architectures with linear computational complexity.

### *2.3. Mamba Networks for Hyperspectral Image Classification*

As an emerging structured state space model (SSM), Mamba [42] has gained significant attention in natural language processing and computer vision due to its efficient long-range dependency modeling capabilities. Unlike Transformers, which rely on self-attention mechanisms for feature modeling, Mamba captures long-range dependencies with linear complexity through implicit state representations and selective filtering mechanisms, offering a novel approach for hyperspectral image (HSI) classification.

Recently, several studies have explored the application of the Mamba architecture in HSI tasks [43]. For instance, Yao et al. [27] proposed SpectralMamba, integrating PSS and GSSM modules to enhance sequential learning in the state domain and rectify spectral information, respectively. He et al. [44] proposed a 3D state space model (3D-SSM) to simultaneously model spatial and spectral information, enhancing feature representation while maintaining computational efficiency. However, this approach primarily relies on global information modeling and struggles to effectively capture local spatial features. Additionally, Vision Mamba, an extension of Mamba for computer vision [45], has demonstrated strong capabilities in modeling spatial information in 2D images. Nonetheless, when directly applied to HSI classification, it still faces challenges in adequately leveraging spectral information. Existing Mamba-based HSI classification methods remain in the exploratory stage and lack effective mechanisms for integrating spatial and spectral features in HSI data. Therefore, developing a method that combines Mamba with local spatial feature extraction modules for efficient HSI classification remains an important research direction.

To address this challenge, this study proposes an innovative approach that thoroughly analyzes the impact of channel dimensions on parameter complexity in Vision Mamba and integrates a multi-receptive field convolutional feature extraction module with a parallel Mamba architecture. This design significantly reduces parameter complexity while further enhancing local spatial feature extraction capabilities. Specifically, we introduce a multi-receptive field local feature extraction framework that captures multi-scale spatial information through different receptive field sizes, thereby facilitating the effective integration of spatial and spectral features in HSI data. This approach not only efficiently models the spatial–spectral relationships in hyperspectral images but also reduces the parameter count of Vision Mamba, improving its scalability for large-scale datasets.

## 3. Proposed Methodology

### *3.1. Preliminary*

State space models (SSMs) are widely employed to characterize the dynamic behavior of a system. The fundamental concept involves mapping a one-dimensional input sequence $x(t) \in \mathbb{R}$ to an output sequence $y(t) \in \mathbb{R}$ via an intermediate latent state variable $h(t) \in \mathbb{R}^N$. In classical SSMs, the evolution of the system's state is described by the following ordinary differential equations (ODEs):

$$\frac{d}{dt}h(t) = Ah(t) + Bx(t) \tag{1}$$

$$y(t) = Ch(t) + \lambda x(t) \tag{2}$$

where $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, $C \in \mathbb{R}^{1\times N}$, and $\lambda \in \mathbb{R}$ represent the state transition matrix, the input-to-state mapping matrix, the state-to-output mapping matrix, and the direct input-to-output mapping, respectively.

In order to incorporate SSMs into deep learning frameworks, a discretization of the continuous model is required. By introducing a time step $\Delta$, the system can be discretized as follows:

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t \tag{3}$$

$$y_t = Ch_t + \lambda x_t \tag{4}$$

where the discretized system matrices are computed as

$$\bar{A} = e^{\Delta A} \tag{5}$$

$$\bar{B} = (\Delta A)^{-1}\left(e^{\Delta A} - I\right)\Delta B \tag{6}$$

This discretization process is known as the Selective Scan Mechanism (S6). Unlike conventional Linear Time-Invariant (LTI) state-space models, S6 introduces a dynamic adjustment strategy that modifies system matrices based on both past and present input sequence information. This adaptability effectively overcomes the inherent constraints of LTI SSMs, which struggle to capture long-range dependencies in sequential data.

By allowing system matrices to be updated dynamically, the S6 mechanism enhances the flexibility of spatial and temporal modeling in visual sequence data. This capability is particularly advantageous for deep learning applications as it enables the effective extraction of complex, evolving patterns in image sequences.

### *3.2. Overview of MRFP-Mamba*

Figure 1 illustrates the detailed architecture of the proposed MRFP-Mamba model, which consists of two main components. The first component is the multi-receptive field convolutional feature extraction module, which employs convolutional kernels of different sizes to capture local spatial information, thereby enhancing the hierarchical representation of spatial features. The second component is the parallel Vision Mamba structure, which leverages structured state space models (SSMs) to model long-range dependencies and improve spatial–spectral feature fusion.

After processing through three layers of multi-receptive field convolutional feature extraction and parallel Mamba operations, the features are further refined through global average pooling and a linear classification layer to obtain the final prediction results. The following sections provide a detailed explanation of the multi-receptive field convolutional feature extraction module and the parallel Vision Mamba structure.

### *3.3. Multi-Receptive Field Convolutional Feature Extraction*

Hyperspectral images (HSIs) encapsulate rich spatial contextual details, yet single-scale convolutional layers often fail to comprehensively capture multi-granular local features critical for discriminative representation. To address this limitation, we introduce a multi-receptive-field convolutional feature extraction module, which leverages parallel convolutional branches with diverse kernel sizes to concurrently extract spatial features across multiple scales.

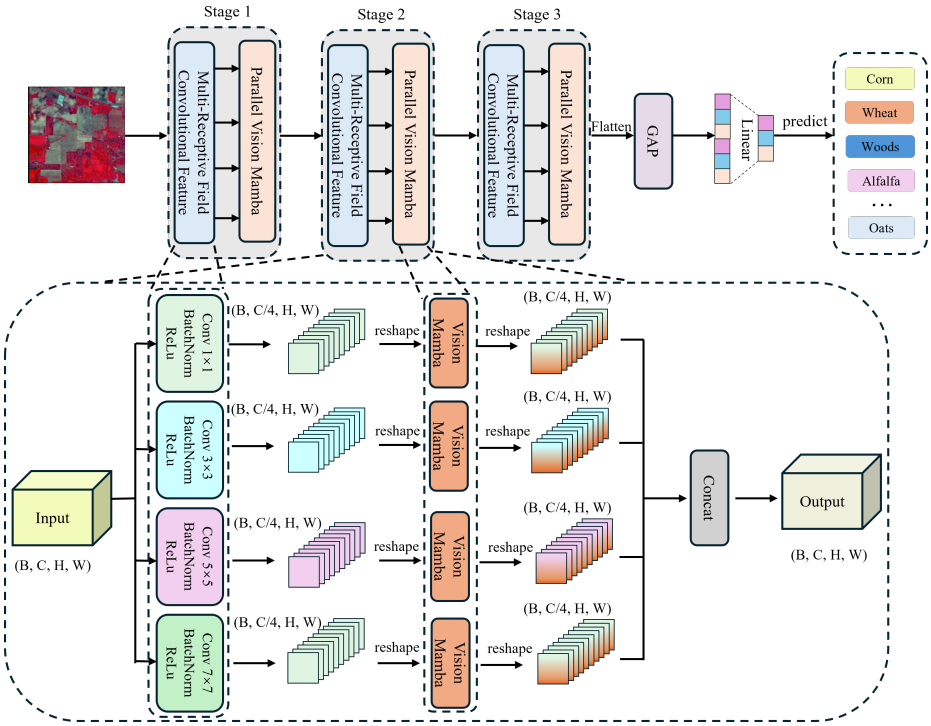


**Figure 1.** The proposed MRFP-Mamba model architecture for HSI classification. The network consists of three structurally identical stages. Each stage comprises a multi-scale receptive field convolution feature extraction module, with convolution kernels of size $1 \times 1$, $3 \times 3$, $5 \times 5$, and $7 \times 7$, and a parallel Vision Mamba module. The multi-scale receptive field convolution module efficiently captures fine-grained and large-scale spatial features through the collaborative effect of different receptive fields. To reduce the number of parameters in the Vision Mamba module, the input channels are equally split into four branches for parallel processing, significantly improving computational efficiency and modeling capability. The complete algorithm flow is detailed in Algorithm 1.

**Algorithm 1** MRFP-Mamba Implementation Process

**Require:** Hyperspectral image $X \in \mathbb{R}^{H \times W \times C}$, label map $Y \in \mathbb{R}^{H \times W}$
Patch size $s = 15$, training ratio $\mu\%$, stages: $C_{out} = [256, 128, 64]$
**Ensure:** Classification map $\hat{Y}$, accuracy metrics: OA, AA, $\kappa$.
Initialize batch size $B = 100$, optimizer Adam (learning rate $10^{-3}$), epochs $E = 100$.
Extract patches $X_{in} \in \mathbb{R}^{B \times C \times s \times s}$ from $X$, split into train/test sets.
**for** epoch $i = 1$ **to** $E$ **do**
  **for** each batch $(X_{batch}, Y_{batch})$ **do**
    Initialize feature map $x = X_{batch} \in \mathbb{R}^{B \times C \times s \times s}$
    **for stage** $= 1$ **to** 3 **do** ▷ 3 stages with decreasing channels
      Multi-scale branching:
      **for** $k \in \{1, 3, 5, 7\}$ **do** ▷ Four branches with kernel sizes $1 \times 1$ to $7 \times 7$
        Branch $k$:
        Conv2D: $x_k = Conv2D_{k \times k}(x)$ ▷ Output channels: $\frac{C_{out}}{4}$
        Vision-Mamba: $\hat{x}_k = Vision - Mamba(x_k)$
      Concatenate outputs: $x_{cat} = Concat(\hat{x}_1, \hat{x}_3, \hat{x}_5, \hat{x}_7)$
    Flatten: $x_{flat} = Flatten(x) \quad (B \times C_{out} \times s^2)$
    Global Average Pooling: $x_{gap} = \frac{1}{s^2} \sum_{i,j} x_{flat}^{(i,j)}$
    Linear layer: $\hat{Y}_{batch} = Softmax(Linear(x_{gap}))$
    Compute loss: $\mathcal{L} = CrossEntropy(\hat{Y}_{batch}, Y_{batch})$
    Backpropagate $\mathcal{L}$ and update weights.
Generate $\hat{Y}$ by aggregating test set predictions.
Calculate OA, AA, $\kappa$ using $Y$.

Specifically, the input features are processed through four parallel 2D convolutional layers with kernel sizes $1 \times 1$, $3 \times 3$, $5 \times 5$, and $7 \times 7$. Each branch independently generates feature maps that are subsequently normalized via Batch Normalization (BatchNorm) and activated by the ReLU function to enhance discriminative power. Mathematically, the operations are defined as follows.

Given the input feature $X \in \mathbb{R}^{H \times W \times C}$, the MRFCFE module employs four different convolution operations:

$$X_1 = \sigma(BN(X * W_{1\times1})) \tag{7}$$

$$X_3 = \sigma(BN(X * W_{3\times3})) \tag{8}$$

$$X_5 = \sigma(BN(X * W_{5\times5})) \tag{9}$$

$$X_7 = \sigma(BN(X * W_{7\times7})) \tag{10}$$

where $W_{k\times k}$ represents the convolution kernel of size $k \times k$, $BN(\cdot)$ denotes the batch normalization operation, and $\sigma(\cdot)$ is the activation function (e.g., ReLU). This parallel architecture enables the simultaneous extraction of fine-grained details (via smaller kernels) and coarse contextual information (via larger kernels), effectively resolving the scale sensitivity issue inherent in single-receptive-field designs and enhancing the model's ability to represent multi-scale spatial structures in HSIs.

### *3.4. Parallel Mamba Structure for Long-Range Spatial–Spectral Dependency Modeling*

#### 3.4.1. Vision Mamba Parameter Impact Analysis

As a state-space model (SSM)-based architercture, Vision Mamba's parameter size is influenced by several key factors, including the number of input channels, the dimensionality of the SSM state, the kernel size of the internal 1D convolution, the expansion ratio of projection layers, and the rank of the stride operation. Among these, the number of input channels plays the most crucial role in determining the model's overall parameter count. To gain deeper insights into how these components contribute to model complexity, this study provides a comprehensive analysis from multiple perspectives.

In Vision Mamba (see Figure 2), $d_i$ represents the expanded projection channels, determined by the input channel size $d_m$ and the projection expansion factor $expand$ (with a default value of 2). The equation for this relationship is

$$d_i = expand * d_m \tag{11}$$

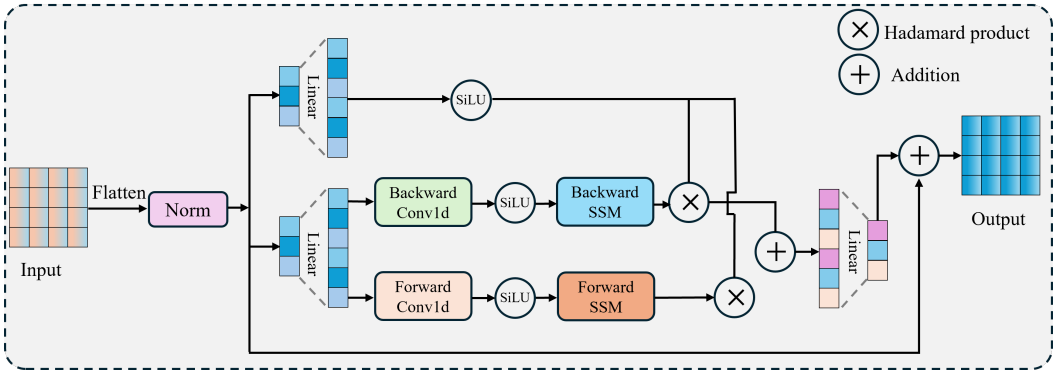


**Figure 2.** Detailed architecture of Vision Mamba.The framework primarily consists of bidirectional (forward–backward) state space models, linear layers, and residual connections, achieving global contextual modeling with linear computational complexity.

This means $d_i$ increases multiplicatively with the increase in input channels, significantly influencing the model's parameter count.

In addition, the input and output projection layers are directly related to $d_m$ and $d_i$. The input projection layer (*in_proj*)maps the input channels $d_m$ to $d_i * 2$, while the output projection layer (*out_proj*) maps $d_i$ back to $d_m$. The specific parameter calculation formulas are as follows:

$$params_{in\_proj} = d_m * d_i * 2 \tag{12}$$

$$params_{out\_proj} = d_i * d_m \tag{13}$$

The parameter count of these projection layers is directly proportional to the product of $d_m$ and $d_i$, meaning that an increase in the number of input channels significantly raises the parameter count of these layers.

Then, the linear projection layers in the state-space model *x_proj* and *dt_proj* are also related to $d_m$ and $d_i$. *x_proj* maps $d_i$ to ($dt\_rank + d\_state * 2$), while *dt_proj* maps *dt_rank* to $d_i$. The parameter count formulas for these layers are as follows:

$$params_{x\_proj} = d_i * (dt_rank + d_state * 2) \tag{14}$$

$$params_{dt\_proj} = d_rank * d_i + d_i \tag{15}$$

where *dt_rank* is the rank of the step, given by ($dt\_rank = d\_model/16$), and *d_state* denotes the size of the state dimension (usually fixed to 16). Therefore, an increase in $d_m$ leads to an increase in *dt_rank*, thereby increasing the parameter count of these projection layers.

Lastly, in Vision Mamba, the convolutional layers usually perform convolution operations on the input using $Conv1d$, where the size of the convolutional kernel is directly related to $d_i$. An increase in $d_i$ leads to an increase in the parameter count of the convolutional layers. In the SSM module, *A_logs* is a parameter matrix with shape ($d_i, d\_state$), so an increase in $d_i$ significantly increases the parameter count of this matrix.

From the above analysis, it is clear that $d_m$ has a big influence on the model's parameter count. Specifically, more $d_m$ leads to more $d_i$, which multiplies the parameter count across layers. Experimental results demonstrate that the parameter size of the Vision Mamba structure is significantly influenced by the number of input channels. When the input channel size is set to $d_m = 256$, the total number of parameters reaches 437,760. However, reducing the input channels to $d_m = 64$ drastically decreases the parameter count to 32,640, representing a 92.54% reduction. In this study, we maintain the total number of channels while equally dividing them into four branches for input. This approach reduces the total parameter count to 130,560, achieving a 70.82% reduction compared to the non-branching case. These findings highlight the critical role of input channel configuration in controlling the parameter size of Vision Mamba and suggest that proper channel partitioning can effectively reduce model parameters while preserving computational capacity.

#### 3.4.2. Parallel Vision Mamba Module

The multi-scale features generated by the MRFCFE module are not directly concatenated but are separately processed through independent Vision Mamba branches:

$$H_1 = VisionMamba(X_1) \tag{16}$$

$$H_3 = VisionMamba(X_3) \tag{17}$$

$$H_5 = VisionMamba(X_5) \tag{18}$$

$$H_7 = VisionMamba(X_7) \tag{19}$$

This design ensures that each Vision Mamba module processes features with a specific receptive field, preserving multi-scale spatial characteristics while leveraging Mamba's

state-space modeling capabilities to enhance long-range spectral dependencies. The Parallel Vision Mamba (PVM) architecture then integrates the processed features through channel-wise concatenation:

$$H_{PVM} = Concat(H_1, H_3, H_5, H_7). \tag{20}$$

Here, $H_i$ represents the output of the Mamba processing for each receptive field branch. The concatenated feature tensor $H_{PVM}$ combines the local spatial information extracted by MRFCFE with the global spectral context modeled by Vision Mamba, constructing a powerful feature representation that captures both multi-scale spatial patterns and long-range dependencies in hyperspectral data.

For classification, the PVM features first undergo global average pooling (GAP):

$$F = GAP(H_{PVM}). \tag{21}$$

The final model output (prediction) is given by a linear layer:

$$Output = Linear(F). \tag{22}$$

Here, $F$ denotes the globally averaged pooled features obtained from the parallel Vision Mamba and MRFCFE feature extraction framework. This linear layer directly maps the features to the classification space, which can then be followed by softmax for probability computation or used for loss calculation.

## 4. Experiments

We selected four hyperspectral image (HSI) datasets, including Indian Pines, Pavia University, Houston 2013, and WHU-Hi-LongKou, to evaluate our proposed method. The experiments included classification result analysis, parameter analysis, and ablation studies.

### *4.1. Datasets*

#### 4.1.1. Indian Pines Dataset

The Indian Pines dataset was collected in 1992 using the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), manufactured by NASA Jet Propulsion Laboratory (JPL), Pasadena, CA, USA. over Northwestern Indiana, specifically covering an area with Indian pine trees. The spectrometer captures wavelengths ranging from 0.4 to 2.5 μm. After removing water absorption bands, the dataset retains 200 spectral bands with a spatial resolution of 20 m per pixel, covering a total area of 145 × 145 pixels. It consists of 10,249 labeled pixels spanning 16 land cover classes, including categories such as Alfalfa and Corn-notill. The specific data partitioning used in our experiments is detailed in Table 1.

#### 4.1.2. Pavia University Dataset

The Pavia University (PU) dataset was collected in 2001 using the ROSIS sensor, capturing 115 spectral bands within the wavelength range of 380 nm to 860 nm. After removing noisy bands, 103 spectral bands were retained for analysis. The dataset consists of an image with a spatial resolution of 610 × 340 pixels and includes 42,776 labeled samples spanning nine land cover categories. For experimentation, 1% of the labeled data is allocated for training, while the remaining 99% is reserved for testing. This partitioning strategy ensures a robust model evaluation and a thorough assessment of classification performance, as detailed in Table 2.

**Table 1.** Number of training and testing samples for the Indian Pines dataset.

| Class No. | Class Name | Training | Testing |
|---|---|---|---|
| 1 | Alfalfa | 5 | 41 |
| 2 | Corn-notill | 143 | 285 |
| 3 | Corn-minill | 83 | 747 |
| 4 | Corn | 24 | 213 |
| 5 | Grass-pasture-mowed | 48 | 435 |
| 6 | Grass-tress | 73 | 657 |
| 7 | Grass-pasture | 3 | 25 |
| 8 | Hay-windrowed | 48 | 430 |
| 9 | Oats | 2 | 18 |
| 10 | Soybean-notill | 97 | 875 |
| 11 | Soybean-mintill | 123 | 1112 |
| 12 | Soybean-clean | 59 | 534 |
| 13 | Wheat | 20 | 185 |
| 14 | Woods | 126 | 1139 |
| 15 | Buildings | 39 | 347 |
| 16 | Stone | 6 | 84 |
| Total | | 1024 | 9225 |

**Table 2.** Number of training and testing samples for the Pavia University dataset.

| Class No. | Class Name | Training | Testing |
|---|---|---|---|
| 1 | Asphalt | 66 | 6565 |
| 2 | Meadows | 186 | 18,463 |
| 3 | Gravel | 20 | 2079 |
| 4 | Trees | 30 | 3034 |
| 5 | Painted metal sheets | 13 | 1332 |
| 6 | Bare Soil | 50 | 4979 |
| 7 | Bitumen | 13 | 1317 |
| 8 | Self-Blocking Bricks | 36 | 3646 |
| 9 | Shadows | 9 | 938 |
| Total | | 423 | 42,353 |

#### 4.1.3. Houston 2013 Dataset

The Houston 2013 dataset is a publicly accessible hyperspectral dataset acquired using an Airborne Laser Mapping (ALM) system equipped with a 2.5 μm wavelength laser. The data were collected in the summer of 2013 over Houston, Texas, USA and were initially introduced as part of the 2013 IEEE GRSS Data Fusion Competition. The dataset consists of an image with a spatial resolution of 949 × 1905 pixels, captured from an aircraft flying at an altitude of 500 m between 12:30 p.m. and 4:30 p.m. on 18 June 2013. It includes 15 distinct land cover categories with a total of 15,029 labeled samples. For experimental purposes, 10% of the labeled data was allocated for training, while the remaining 90% was designated for testing, as detailed in Table 3.

#### 4.1.4. WHU-Hi-LongKou Dataset

The WHL dataset was captured using a Headwall Nano-Hyperspec imaging sensor with an 8 mm focal length, mounted on a DJI Matrice 600 Pro UAV (manufactured by DJI Technology Co., Ltd., Shenzhen, China). The UAV operated at an altitude of 500 m, producing hyperspectral images with a spatial resolution of 0.463 m per pixel. The dataset consists of a 550 × 400 pixel image spanning 270 spectral bands within the 400–1000 nm wavelength range. It includes a total of 203,523 labeled samples categorized into 9 distinct

land cover types. For evaluation, we allocated 0.5% of the labeled data for training and reserved the remaining 99.5% for testing, as summarized in Table 4.

**Table 3.** Number of training and testing samples for the Houston 2013 dataset.

| Class No. | Class Name | Training | Testing |
|---|---|---|---|
| 1 | Healthy Grass | 125 | 126 |
| 2 | Stressed Grass | 125 | 129 |
| 3 | Synthetic Grass | 70 | 627 |
| 4 | Trees | 124 | 120 |
| 5 | Soil | 124 | 118 |
| 6 | Water | 33 | 292 |
| 7 | Residential | 127 | 1141 |
| 8 | Commercial | 124 | 120 |
| 9 | Road | 125 | 1127 |
| 10 | Highway | 123 | 1104 |
| 11 | Railway | 123 | 1112 |
| 12 | Parking Lot 1 | 123 | 123 |
| 13 | Parking Lot 2 | 47 | 422 |
| 14 | Tennise Court | 43 | 385 |
| 15 | Running Track | 66 | 594 |
| Total | | 1502 | 13,527 |

**Table 4.** Number of training and testing samples for the WHU-Hi-LongKou dataset.

| Class No. | Class Name | Training | Testing |
|---|---|---|---|
| 1 | Corn | 172 | 34,339 |
| 2 | Cotton | 41 | 8333 |
| 3 | Sesame | 15 | 3016 |
| 4 | Broad-leaf soybean | 316 | 62,896 |
| 5 | Narrow-leaf soybean | 20 | 4131 |
| 6 | Rice | 59 | 1795 |
| 7 | Water | 335 | 66,721 |
| 8 | Roads and houses | 35 | 7089 |
| 9 | Mixed weed | 26 | 5203 |
| Total | | 1019 | 203,523 |

### *4.2. Experimental Setup*

#### 4.2.1. Implementation Details

To facilitate a fair and efficient performance assessment, the proposed MRFP-Mamba architecture was implemented within the PyTorch 2.1.1 framework and deployed on an NVIDIA GeForce RTX 4090 GPU (manufactured by NVIDIA Corporation, Santa Clara, CA, USA). For training, 100 non-overlapping ($15 \times 15$) pixel patches were randomly extracted from each dataset to construct the input feature space, ensuring representative sampling of spatial–spectral contexts. The training procedure spanned 100 epochs, leveraging the Adam optimizer with an initial learning rate of ($1 \times 10^{-3}$) and a consistent mini-batch size of 100 across all experimental setups. Model performance was quantified using three standard evaluation metrics: Overall Accuracy (OA) for global classification correctness, Average Accuracy (AA) to assess class-wise consistency, and the Kappa coefficient ($\kappa$) to measure agreement beyond random chance. This configuration balances computational efficiency with rigorous validation, ensuring reliable comparison against state-of-the-art baselines.

#### 4.2.2. Comparison with State-of-the-Art Backbone Methods

To validate the proposed method, we benchmark against a diverse set of state-of-the-art classification networks spanning CNN, Transformer, and Mamba architectures: 2D-CNN [46], 3D-CNN [47], ViT [48], Deep-ViT [49], HiT [20], SSFTT [39], GAHT [24], DCTN [41], MambaHSI [50] and 3DSS-Mamba [44].

Conventional CNN-based approaches serve as foundational baselines: 2D-CNNs employ standard convolutional layers with batch normalization and ReLU activation to extract spatial features while 3D-CNNs extend this to volumetric operations, jointly modeling spectral and spatial dimensions through 3D kernels. In contrast, ViT introduces a paradigm shift by leveraging linear projection and transformer encoders to model global contextual dependencies across spectral bands, though at the cost of quadratic complexity.

Transformer variants refine this framework with specialized designs: HiT integrates a Spectral-Adaptive 3D Convolution Projection (SACP) module and Convolutional Permutator to enhance spatial–spectral interactions, while SSFTT decomposes HSI data into spectral and spatial tokens for independent refinement, improving local structure learning and semantic alignment. GAHT mitigates global attention's computational and representational challenges through a Grouped Pixel Embedding strategy, constraining Multi-Head Self-Attention (MHSA) within localized spectral contexts to reduce feature dispersion.

Mamba-based baselines, such as 3DSS-Mamba, adopt structured state space models (SSMs) for efficient long-range modeling, combining 3D state representations with linear-complexity operations across spectral and spatial axes.

The proposed MRFP-Mamba distinguishes itself by introducing a parallel architecture that fuses a multi-receptive-field convolutional module with a parameter-optimized Vision Mamba branch. This design jointly enhances local spatial feature extraction and global spectral dependency modeling, addressing key limitations of both CNN and Transformer paradigms.

### *4.3. Results and Analysis*

We conduct experiments on four widely used hyperspectral image (HSI) datasets and evaluate the results using three metrics. The outcomes are summarized in Tables 5–8, with the best results highlighted in bold.

From Tables 5–8, it can be seen that the proposed MRFP-Mamba achieves notable improvements in classification performance across four widely used hyperspectral image (HSI) datasets, thereby validating the effectiveness of the proposed multi-receptive field parallel Mamba architecture in HSI classification tasks. Compared with conventional 2D convolutional neural networks (2D-CNNs), MRFP-Mamba integrates a multi-receptive field convolutional feature extraction module which enables the simultaneous capture of fine-grained local features and large-scale contextual information. This design effectively addresses the limitations of traditional convolutional models, which struggle to model long-range dependencies due to their restricted receptive fields. On the four benchmark datasets, MRFP-Mamba outperforms 2D-CNNs in Overall Accuracy (OA) by 0.18%, 9.05%, 3.79%, and 0.68%, respectively.

Building on this analysis, we develop a parallel Vision Mamba module through a comprehensive examination of the Vision Mamba architecture's parameter composition. This design drastically reduces the parameter count while concurrently enhancing the model's capability to model long-range dependencies in both spatial and spectral domains. Experimental comparisons with 3DSS-Mamba demonstrate that MRFP-Mamba outperforms the baseline across all datasets, achieving improvements in Overall Accuracy (OA) of 0.22%, 0.52%, 0.16%, and 0.42%, respectively.

**Table 5.** Classification results of the Indian Pines dataset with 10% training samples. All bolded values indicate the best results. This applies to all the following tables.

| Class No. | CNNs | | Transformers | | | | | | Mambas | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2D-CNN | 3D-CNN | ViT | Deep-ViT | HiT | SSFTT | GAHT | DCTN | MambaHSI | 3DSS-Mamba | Ours |
| 1 | 95.10 ± 3.52 | 11.46 ± 16.22 | 78.78 ± 16.65 | 9.27 ± 15.03 | 4.88 ± 8.09 | 89.62 ± 6.89 | 88.29 ± 7.30 | 87.07 ± 11.75 | 71.71 ± 17.49 | 86.59 ± 9.27 | 82.44 ± 13.22 |
| 2 | 91.56 ± 1.98 | 89.75 ± 2.74 | 94.21 ± 3.36 | 67.27 ± 12.07 | 90.67 ± 2.73 | 94.13 ± 1.09 | 92.67 ± 1.99 | 93.58 ± 3.20 | 92.45 ± 3.44 | 91.61 ± 2.80 | 93.25 ± 3.00 |
| 3 | 92.08 ± 1.56 | 85.98 ± 4.70 | 95.34 ± 3.90 | 50.04 ± 12.45 | 80.17 ± 7.87 | 90.10 ± 2.66 | 94.39 ± 4.27 | 96.14 ± 3.08 | 92.90 ± 3.03 | 95.77 ± 3.18 | 96.87 ± 3.19 |
| 4 | 97.94 ± 1.46 | 68.08 ± 10.32 | 96.48 ± 2.90 | 76.38 ± 17.90 | 87.37 ± 7.39 | 94.93 ± 3.45 | 94.27 ± 5.55 | 96.95 ± 3.37 | 85.07 ± 3.62 | 95.49 ± 2.84 | 94.74 ± 3.60 |
| 5 | 93.09 ± 3.34 | 86.07 ± 5.50 | 94.00 ± 2.93 | 40.00 ± 11.97 | 78.46 ± 7.07 | 93.09 ± 2.48 | 93.20 ± 3.54 | 94.83 ± 1.40 | 93.47 ± 1.37 | 95.13 ± 2.62 | 89.56 ± 4.38 |
| 6 | 95.67 ± 2.98 | 93.14 ± 1.47 | 97.17 ± 0.94 | 90.12 ± 4.25 | 93.93 ± 1.20 | 95.98 ± 1.25 | 96.21 ± 0.86 | 96.16 ± 1.02 | 95.25 ± 2.26 | 97.23 ± 1.21 | 95.80 ± 2.20 |
| 7 | 7.93 ± 19.30 | 0.00 ± 0.00 | 62.80 ± 25.43 | 0.00 ± 0.00 | 0.00 ± 0.00 | 54.65 ± 35.88 | 27.20 ± 28.78 | 46.40 ± 40.37 | 81.60 ± 22.57 | 50.00 ± 30.79 | 92.00 ± 8.39 |
| 8 | 99.69 ± 0.45 | 99.67 ± 0.68 | 100 ± 0.00 | 97.07 ± 1.33 | 99.74 ± 0.69 | 98.74 ± 1.40 | 100 ± 0.00 | 100.00 ± 0.00 | 98.56 ± 0.77 | 100.00 ± 0.00 | 97.26 ± 2.23 |
| 9 | 73.35 ± 29.10 | 0.00 ± 0.00 | 11.11 ± 11.65 | 0.00 ± 0.00 | 0.00 ± 0.00 | 14.35 ± 32.06 | 10.56 ± 19.95 | 3.33 ± 5.67 | 45.56 ± 12.86 | 73.89 ± 19.25 | 53.33 ± 20.67 |
| 10 | 87.75 ± 1.58 | 76.34 ± 1.67 | 79.65 ± 9.39 | 62.19 ± 10.05 | 75.43 ± 3.53 | 87.12 ± 1.76 | 81.13 ± 2.27 | 82.86 ± 2.39 | 94.74 ± 1.04 | 84.83 ± 5.98 | 92.57 ± 4.84 |
| 11 | 96.23 ± 1.26 | 95.69 ± 2.03 | 89.18 ± 19.02 | 89.67 ± 4.14 | 95.86 ± 1.70 | 97.68 ± 0.83 | 98.11 ± 1.03 | 98.16 ± 1.32 | 96.90 ± 1.08 | 97.61 ± 1.67 | 97.04 ± 1.90 |
| 12 | 91.75 ± 2.23 | 88.91 ± 6.05 | 90.11 ± 7.70 | 59.31 ± 16.66 | 88.97 ± 3.96 | 89.52 ± 3.33 | 93.75 ± 3.15 | 89.53 ± 8.68 | 88.20 ± 3.45 | 92.30 ± 2.85 | 86.03 ± 4.20 |
| 13 | 98.12 ± 1.25 | 80.11 ± 12.35 | 96.00 ± 2.18 | 79.03 ± 0.07 | 93.03 ± 3.61 | 95.02 ± 3.61 | 87.03 ± 9.35 | 94.49 ± 4.04 | 90.70 ± 6.48 | 90.92 ± 6.69 | 91.89 ± 6.84 |
| 14 | 98.25 ± 2.46 | 98.93 ± 0.76 | 99.75 ± 0.19 | 96.71 ± 3.23 | 99.44 ± 0.60 | 98.67 ± 0.64 | 99.46 ± 0.55 | 99.60 ± 0.28 | 98.00 ± 1.37 | 99.28 ± 0.52 | 98.02 ± 1.09 |
| 15 | 97.82 ± 1.45 | 77.84 ± 6.81 | 91.67 ± 4.58 | 74.47 ± 15.57 | 86.22 ± 12.60 | 96.10 ± 2.79 | 92.77 ± 5.91 | 92.51 ± 8.54 | 89.80 ± 3.54 | 93.63 ± 4.09 | 90.32 ± 3.56 |
| 16 | 51.99 ± 22.01 | 8.81 ± 16.07 | 30.36 ± 20.06 | 11.55 ± 18.72 | 8.81 ± 17.82 | 39.23 ± 32.33 | 31.90 ± 16.30 | 38.93 ± 25.76 | 85.71 ± 2.71 | 44.40 ± 20.75 | 86.90 ± 6.16 |
| OA (%) | 94.28 ± 1.41 | 88.57 ± 0.78 | 91.73 ± 5.38 | 75.30 ± 5.40 | 88.94 ± 1.42 | 94.09 ± 0.97 | 93.56 ± 0.72 | 94.14 ± 0.63 | 94.16 ± 1.59 | 94.24 ± 0.59 | **94.46 ± 1.33** |
| AA (%) | 85.52 ± 4.26 | 66.30 ± 2.22 | 81.66 ± 3.03 | 56.44 ± 5.88 | 67.69 ± 1.95 | 83.06 ± 2.15 | 80.06 ± 2.46 | 81.91 ± 3.57 | 87.54 ± 4.11 | 86.79 ± 3.16 | **89.88 ± 2.54** |
| $\kappa$ (%) | 93.70 ± 1.65 | 86.88 ± 0.90 | 90.61 ± 6.02 | 71.34 ± 6.42 | 87.32 ± 1.64 | 93.26 ± 0.12 | 92.64 ± 0.82 | 93.31 ± 0.73 | 93.34 ± 1.81 | 93.42 ± 0.68 | **93.67 ± 1.53** |

**Table 6.** Classification results of the PaviaU dataset with 1% training samples.

| Class No. | CNNs | | Transformers | | | | | | Mambas | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2D-CNN | 3D-CNN | ViT | Deep-ViT | HiT | SSFTT | GAHT | DCTN | MambaHSI | 3DSS-Mamba | Ours |
| 1 | 93.04 ± 2.35 | 91.80 ± 4.64 | 73.71 ± 4.03 | 91.86 ± 5.58 | 97.13 ± 0.93 | 96.31 ± 2.85 | 96.47 ± 1.68 | 96.05 ± 1.55 | 96.24 ± 2.31 | 97.23 ± 1.50 | 97.19 ± 1.52 |
| 2 | 99.21 ± 0.45 | 97.73 ± 1.50 | 95.26 ± 2.42 | 44.13 ± 9.52 | 98.95 ± 0.53 | 99.80 ± 0.18 | 99.48 ± 0.43 | 99.58 ± 0.70 | 99.96 ± 0.04 | 99.76 ± 0.19 | 99.83 ± 0.19 |
| 3 | 37.26 ± 17.15 | 35.90 ± 16.58 | 27.47 ± 19.09 | 72.48 ± 8.67 | 87.94 ± 3.03 | 80.55 ± 20.82 | 89.03 ± 6.12 | 88.02 ± 6.48 | 90.37 ± 5.17 | 88.16 ± 5.40 | 91.58 ± 6.04 |
| 4 | 88.83 ± 3.44 | 74.78 ± 11.96 | 22.70 ± 10.64 | 98.77 ± 1.00 | 87.13 ± 2.74 | 88.73 ± 3.29 | 87.26 ± 3.89 | 86.47 ± 11.25 | 88.69 ± 2.53 | 88.23 ± 2.91 | 91.51 ± 2.23 |
| 5 | 99.86 ± 0.34 | 92.30 ± 12.75 | 97.18 ± 2.15 | 70.22 ± 19.06 | 100 ± 0.00 | 99.73 ± 0.53 | 99.50 ± 0.73 | 99.91 ± 0.18 | 99.98 ± 0.05 | 98.41 ± 2.59 | 99.88 ± 0.20 |
| 6 | 84.09 ± 14.11 | 79.47 ± 5.72 | 26.28 ± 9.46 | 31.47 ± 13.30 | 96.29 ± 1.07 | 100 ± 0.00 | 99.76 ± 0.23 | 99.42 ± 0.86 | 99.60 ± 0.32 | 99.95 ± 0.15 | 99.93 ± 0.18 |
| 7 | 39.07 ± 13.26 | 42.41 ± 7.49 | 19.65 ± 12.40 | 51.84 ± 28.80 | 97.44 ± 2.63 | 91.09 ± 10.69 | 93.02 ± 8.96 | 90.73 ± 10.92 | 90.10 ± 5.33 | 97.98 ± 2.72 | 98.90 ± 1.32 |
| 8 | 84.41 ± 15.87 | 50.72 ± 32.72 | 73.58 ± 3.53 | 42.04 ± 15.61 | 96.53 ± 1.54 | 97.77 ± 1.95 | 98.12 ± 1.46 | 97.06 ± 2.05 | 97.53 ± 1.38 | 98.05 ± 1.74 | 98.35 ± 1.12 |
| 9 | 56.89 ± 17.69 | 27.75 ± 19.52 | 4.79 ± 7.80 | 78.44 ± 2.24 | 73.97 ± 2.80 | 69.24 ± 8.30 | 59.47 ± 9.66 | 72.86 ± 5.87 | 74.65 ± 3.24 | 73.06 ± 6.06 | 72.76 ± 8.43 |
| OA (%) | 88.63 ± 3.41 | 82.50 ± 3.62 | 69.13 ± 1.62 | 65.40 ± 2.36 | 96.19 ± 0.30 | 96.42 ± 1.52 | 96.46 ± 0.73 | 96.44 ± 1.70 | 96.99 ± 0.52 | 97.16 ± 0.37 | **97.68 ± 0.48** |
| AA (%) | 75.85 ± 6.01 | 65.87 ± 5.40 | 48.96 ± 3.96 | 71.03 ± 2.70 | 92.82 ± 0.53 | 91.47 ± 4.02 | 91.35 ± 1.59 | 92.23 ± 3.02 | 93.01 ± 1.00 | 93.43 ± 0.94 | **94.44 ± 1.54** |
| $\kappa$ (%) | 84.67 ± 4.84 | 76.30 ± 4.94 | 56.53 ± 2.70 | 91.86 ± 5.58 | 94.94 ± 0.41 | 95.25 ± 2.03 | 95.30 ± 0.98 | 95.26 ± 2.29 | 96.00 ± 0.69 | 96.23 ± 0.50 | **96.92 ± 0.64** |

**Table 7.** Classification results of the Houston 2013 dataset with 10% training samples.

| Class No. | CNNs | | Transformers | | | | | | Mambas | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2D-CNN | 3D-CNN | ViT | Deep-ViT | HiT | SSFTT | GAHT | DCTN | MambaHSI | 3DSS-Mamba | Ours |
| 1 | 95.20 ± 4.66 | 91.29 ± 5.38 | 87.20 ± 5.85 | 85.37 ± 5.00 | 89.44 ± 2.43 | 98.06 ± 1.15 | 96.94 ± 2.55 | 98.06 ± 1.15 | 96.77 ± 1.94 | 98.14 ± 1.55 | 98.08 ± 1.71 |
| 2 | 98.29 ± 1.26 | 92.12 ± 6.99 | 85.67 ± 5.45 | 89.79 ± 6.56 | 95.18 ± 1.02 | 96.18 ± 1.73 | 95.03 ± 1.75 | 96.18 ± 1.73 | 96.38 ± 1.92 | 95.97 ± 2.14 | 99.13 ± 0.86 |
| 3 | 99.74 ± 0.46 | 94.68 ± 3.26 | 98.41 ± 0.77 | 98.56 ± 0.95 | 97.75 ± 0.43 | 97.11 ± 1.07 | 97.02 ± 1.00 | 97.11 ± 1.07 | 98.79 ± 0.45 | 97.18 ± 0.51 | 99.49 ± 0.27 |
| 4 | 97.26 ± 2.47 | 93.31 ± 9.21 | 65.51 ± 6.81 | 70.55 ± 7.48 | 83.57 ± 4.52 | 96.71 ± 1.51 | 96.01 ± 2.01 | 96.71 ± 1.51 | 95.23 ± 2.03 | 97.51 ± 1.56 | 97.73 ± 0.92 |
| 5 | 99.53 ± 0.74 | 98.17 ± 1.72 | 98.80 ± 0.64 | 99.55 ± 0.50 | 99.75 ± 0.19 | 99.97 ± 0.04 | 99.94 ± 0.11 | 99.97 ± 0.04 | 99.94 ± 0.11 | 99.95 ± 0.11 | 99.61 ± 0.31 |
| 6 | 97.46 ± 2.34 | 87.21 ± 8.12 | 66.58 ± 6.82 | 78.22 ± 6.49 | 76.61 ± 2.78 | 89.38 ± 5.34 | 89.55 ± 1.52 | 89.38 ± 5.34 | 90.31 ± 3.44 | 94.32 ± 2.82 | 99.73 ± 0.26 |
| 7 | 93.04 ± 2.26 | 91.46 ± 3.55 | 70.21 ± 4.24 | 85.85 ± 4.60 | 77.59 ± 5.33 | 97.91 ± 2.15 | 98.36 ± 0.85 | 97.91 ± 2.15 | 98.31 ± 1.02 | 97.94 ± 1.80 | 95.21 ± 1.74 |
| 8 | 81.36 ± 3.99 | 84.41 ± 7.90 | 61.80 ± 4.02 | 79.27 ± 5.74 | 84.96 ± 2.55 | 96.18 ± 2.24 | 98.01 ± 1.43 | 96.18 ± 2.24 | 98.51 ± 0.85 | 96.75 ± 2.35 | 96.71 ± 1.33 |
| 9 | 90.02 ± 3.41 | 87.75 ± 3.82 | 65.19 ± 8.48 | 75.18 ± 3.30 | 79.47 ± 2.40 | 98.76 ± 0.89 | 98.00 ± 1.20 | 98.76 ± 0.89 | 98.19 ± 0.89 | 97.91 ± 1.75 | 97.18 ± 1.88 |
| 10 | 96.52 ± 1.51 | 82.10 ± 5.10 | 80.54 ± 6.08 | 96.29 ± 1.06 | 97.22 ± 2.33 | 100.00 ± 0.00 | 100.00 ± 0.00 | 100.00 ± 0.00 | 99.92 ± 0.24 | 99.97 ± 0.08 | 99.86 ± 0.25 |
| 11 | 92.77 ± 2.16 | 85.91 ± 4.62 | 57.72 ± 8.01 | 77.74 ± 10.58 | 89.87 ± 3.52 | 99.69 ± 0.83 | 99.72 ± 0.84 | 99.69 ± 0.83 | 99.89 ± 0.24 | 99.29 ± 0.88 | 98.36 ± 1.12 |
| 12 | 91.19 ± 3.69 | 85.33 ± 5.86 | 65.55 ± 2.87 | 91.98 ± 2.55 | 96.10 ± 0.94 | 97.20 ± 1.23 | 98.39 ± 0.62 | 97.20 ± 1.23 | 98.93 ± 0.60 | 97.80 ± 1.54 | 98.90 ± 0.48 |
| 13 | 97.67 ± 2.03 | 86.81 ± 6.28 | 45.31 ± 20.56 | 78.72 ± 10.51 | 84.91 ± 5.79 | 96.16 ± 3.36 | 95.57 ± 4.28 | 96.16 ± 3.36 | 97.42 ± 2.08 | 97.42 ± 2.12 | 97.16 ± 1.04 |
| 14 | 100.00 ± 0.00 | 91.65 ± 6.35 | 89.40 ± 3.56 | 98.52 ± 1.12 | 99.95 ± 0.10 | 100.00 ± 0.00 | 99.97 ± 0.08 | 100.00 ± 0.00 | 100.00 ± 0.00 | 100.00 ± 0.00 | 99.38 ± 0.53 |
| 15 | 100.00 ± 0.00 | 96.18 ± 2.67 | 74.95 ± 2.22 | 95.15 ± 3.47 | 99.68 ± 0.37 | 99.95 ± 0.15 | 100.00 ± 0.00 | 99.95 ± 0.15 | 100.00 ± 0.00 | 99.95 ± 0.15 | 99.80 ± 0.27 |
| OA (%) | 94.47 ± 0.82 | 89.13 ± 2.12 | 74.41 ± 1.75 | 86.23 ± 1.39 | 90.02 ± 1.15 | 97.74 ± 0.46 | 97.87 ± 0.46 | 97.91 ± 0.42 | 97.86 ± 0.34 | 98.10 ± 0.41 | **98.26 ± 0.26** |
| AA (%) | 95.34 ± 0.77 | 89.89 ± 1.74 | 74.19 ± 2.19 | 86.72 ± 1.06 | 90.14 ± 1.11 | 97.91 ± 0.42 | 97.50 ± 0.51 | 97.55 ± 0.54 | 97.90 ± 0.39 | 98.01 ± 0.36 | **98.42 ± 0.23** |
| $\kappa$ (%) | 94.23 ± 1.92 | 88.24 ± 2.29 | 72.31 ± 1.91 | 85.11 ± 1.50 | 89.20 ± 1.24 | 97.55 ± 0.54 | 97.69 ± 0.50 | 97.74 ± 0.46 | 98.01 ± 0.37 | 97.95 ± 0.44 | **98.11 ± 0.28** |

**Table 8.** Classification results of the WHU-Hi-LongKou dataset with 0.5% training samples.

| Class No. | CNNs | | Transformers | | | | | | Mambas | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2D-CNN | 3D-CNN | ViT | Deep-ViT | HiT | SSFTT | GAHT | DCTN | MambaHSI | 3DSS-Mamba | Ours |
| 1 | 99.88 ± 0.02 | 98.37 ± 0.37 | 98.00 ± 0.50 | 98.71 ± 0.61 | 99.37 ± 0.36 | 99.77 ± 0.18 | 99.71 ± 0.10 | 99.90 ± 0.06 | 99.56 ± 0.28 | 99.86 ± 0.05 | 99.88 ± 0.04 |
| 2 | 99.73 ± 0.09 | 90.40 ± 5.37 | 91.54 ± 4.61 | 85.55 ± 7.66 | 91.48 ± 4.68 | 97.72 ± 0.83 | 97.82 ± 1.16 | 99.00 ± 0.78 | 96.65 ± 2.01 | 98.54 ± 1.56 | 99.48 ± 0.23 |
| 3 | 94.96 ± 1.26 | 1.00 ± 2.84 | 0.13 ± 0.38 | 65.74 ± 20.15 | 74.22 ± 12.89 | 96.41 ± 2.95 | 89.61 ± 3.43 | 99.05 ± 0.57 | 95.38 ± 2.80 | 95.43 ± 4.01 | 95.86 ± 1.45 |
| 4 | 99.12 ± 0.07 | 99.69 ± 0.19 | 98.21 ± 0.81 | 99.37 ± 0.22 | 99.59 ± 0.17 | 99.55 ± 0.18 | 99.83 ± 0.07 | 99.82 ± 0.11 | 99.77 ± 0.20 | 99.76 ± 0.11 | 99.49 ± 0.11 |
| 5 | 95.43 ± 0.72 | 68.09 ± 5.45 | 22.02 ± 6.25 | 58.82 ± 19.46 | 68.03 ± 12.74 | 89.49 ± 4.05 | 85.29 ± 2.93 | 97.14 ± 1.84 | 88.41 ± 3.99 | 91.08 ± 4.06 | 98.06 ± 0.92 |
| 6 | 98.56 ± 0.20 | 96.57 ± 2.16 | 91.60 ± 2.96 | 97.02 ± 2.81 | 98.11 ± 1.89 | 98.51 ± 1.01 | 98.60 ± 0.52 | 98.39 ± 0.66 | 97.47 ± 1.71 | 99.25 ± 0.24 | 98.09 ± 0.56 |
| 7 | 99.74 ± 0.05 | 99.93 ± 0.08 | 99.96 ± 0.03 | 99.84 ± 0.24 | 99.99 ± 0.00 | 99.85 ± 0.10 | 99.94 ± 0.05 | 99.94 ± 0.08 | 99.87 ± 0.10 | 99.98 ± 0.01 | 99.78 ± 0.10 |
| 8 | 85.56 ± 0.79 | 66.97 ± 8.10 | 71.64 ± 6.45 | 69.03 ± 11.48 | 81.13 ± 7.50 | 83.46 ± 5.58 | 91.46 ± 3.47 | 89.59 ± 5.46 | 83.93 ± 5.25 | 88.65 ± 3.06 | 92.83 ± 3.88 |
| 9 | 78.58 ± 2.05 | 67.57 ± 8.36 | 68.68 ± 7.22 | 75.84 ± 9.50 | 71.19 ± 4.74 | 80.82 ± 8.07 | 84.14 ± 5.38 | 77.41 ± 6.49 | 76.49 ± 9.90 | 78.68 ± 8.07 | 90.77 ± 2.17 |
| OA (%) | 98.35 ± 0.09 | 94.92 ± 0.54 | 93.41 ± 0.37 | 95.73 ± 0.67 | 96.88 ± 0.49 | 98.26 ± 0.32 | 98.55 ± 0.20 | 98.76 ± 0.21 | 98.06 ± 0.31 | 98.61 ± 0.25 | **99.03 ± 0.22** |
| AA (%) | 94.61 ± 0.51 | 76.51 ± 2.10 | 71.31 ± 1.09 | 83.32 ± 3.43 | 87.01 ± 2.57 | 93.95 ± 1.40 | 94.04 ± 0.69 | 95.58 ± 0.70 | 93.06 ± 1.40 | 94.58 ± 1.03 | **97.14 ± 0.90** |
| $\kappa$ (%) | 97.81 ± 0.12 | 93.22 ± 0.73 | 91.25 ± 0.49 | 94.35 ± 0.89 | 95.88 ± 0.65 | 97.71 ± 0.42 | 98.09 ± 0.26 | 98.37 ± 0.27 | 97.45 ± 0.42 | 98.15 ± 0.34 | **98.73 ± 0.28** |

While 3DSS-Mamba employs 3D convolutions to extract joint spatial–spectral features, its spatial modeling capacity is constrained by single-scale convolutional kernels and the absence of an effective multi-scale fusion mechanism. This limitation hinders its ability to capture hierarchical spatial details, leading to inferior classification performance compared to MRFP-Mamba's parallel architecture—where multi-receptive-field convolutions and structured state space modeling work synergistically to integrate multi-scale spatial contexts with global spectral dependencies.

In addition, we compared MRFP-Mamba with several Transformer-based state-of-the-art models, including ViT, DeepViT, HiT, SSFTT, and GAHT. The experimental results show that MRFP-Mamba consistently outperforms these methods across various evaluation metrics, further highlighting its competitiveness in HSI classification tasks. In the visualized classification maps shown in Figures 3–6, MRFP-Mamba produces results with clearer boundaries, finer detail restoration, and stronger spatial consistency compared to the other models. These improvements can be attributed to the synergy between its multi-receptive field convolution module, which captures multi-scale spatial features, and the Vision Mamba, which excels at modeling contextual dependencies.

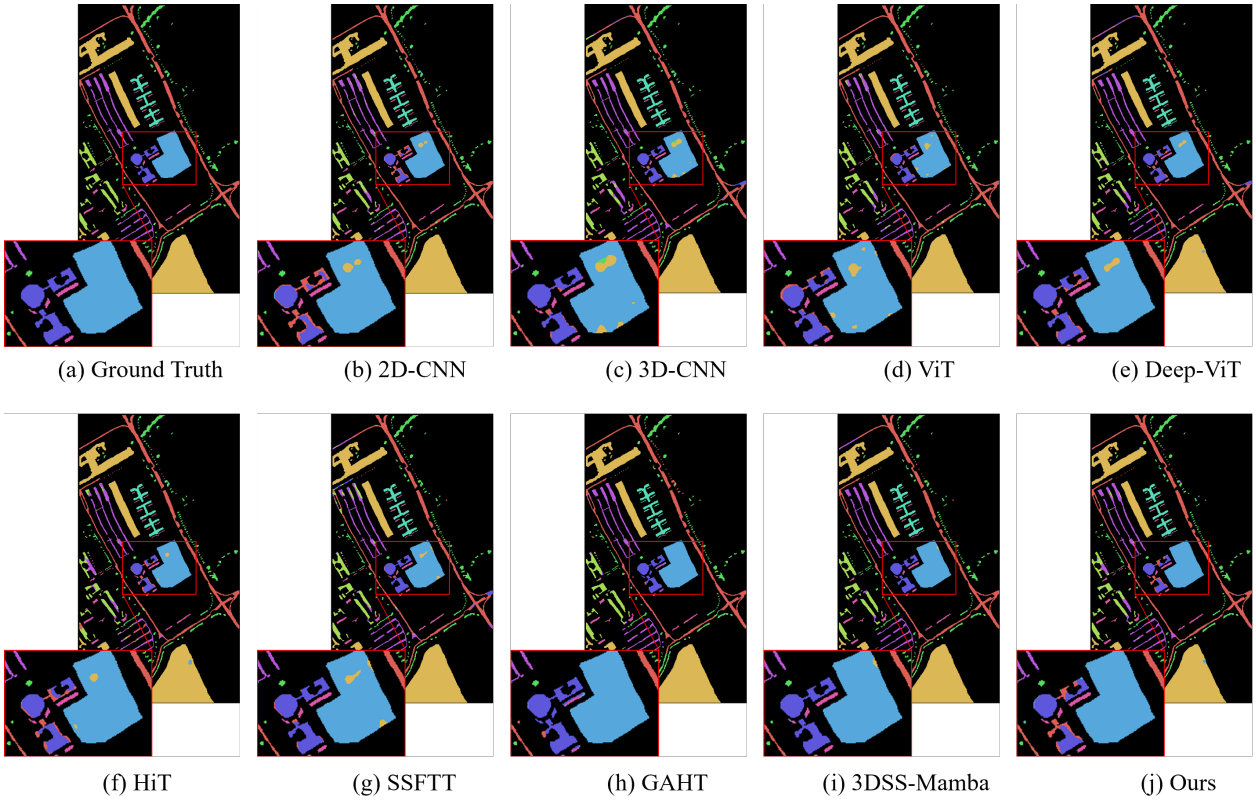


**Figure 3.** Classification maps obtained using different methrods on the PaviaU dataset (with 1% training samples).

To further evaluate the model's discriminative capability, we conducted a t-SNE visualization on the PaviaU dataset (as illustrated in Figure 7), comparing MRFP-Mamba with 2D-CNN, ViT, DeepViT, HiT, SSFTT, GAHT, DCTN, MambaHSI, and 3DSS-Mamba. The results show that MRFP-Mamba produces more compact and better separated class clusters in the feature space, significantly reducing inter-class confusion while enhancing intra-class cohesion. In contrast, 2D-CNN and 3DSS-Mamba display more dispersed and overlapping class distributions, indicating suboptimal performance in extracting either local or global features. Benefiting from the collaborative design of multi-receptive field convolutions and the parallel Vision Mamba, MRFP-Mamba achieves effective multi-scale spatial modeling

and efficient long-range dependency capture, demonstrating superior accuracy and visual representation. Collectively, these findings underscore that developing a model capable of effectively integrating local and global features through a compact and efficient architecture is key to advancing the state of the art in hyperspectral image classification.

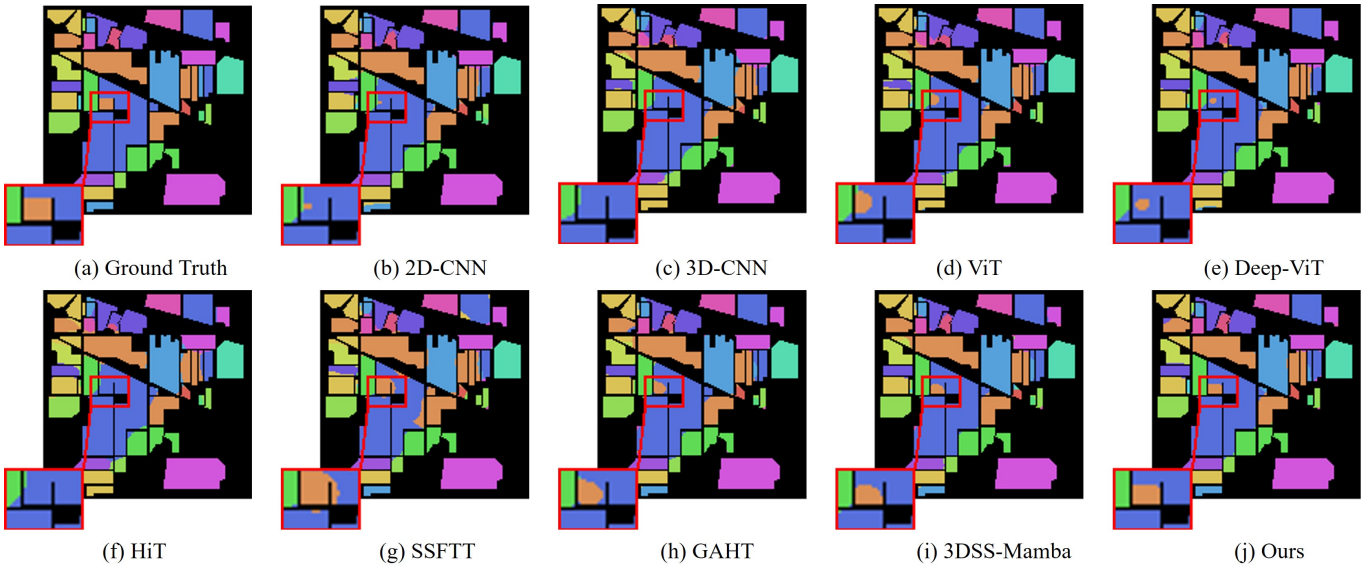


**Figure 4.** Classification maps obtained using different methods on the Indian Pines dataset (with 10% training samples).

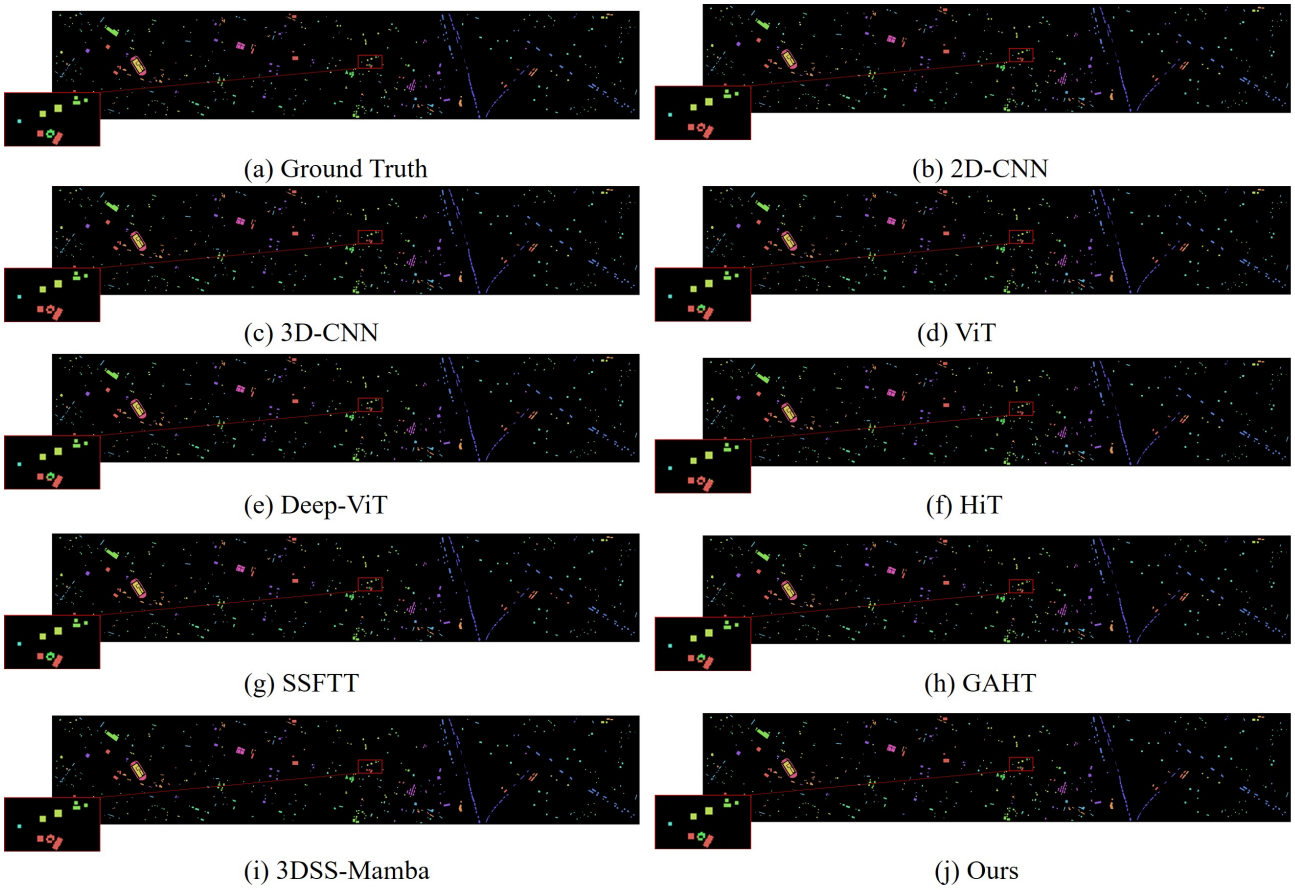


**Figure 5.** Classification maps obtained using different methods on the Houston 2013 dataset (with 10% training samples).

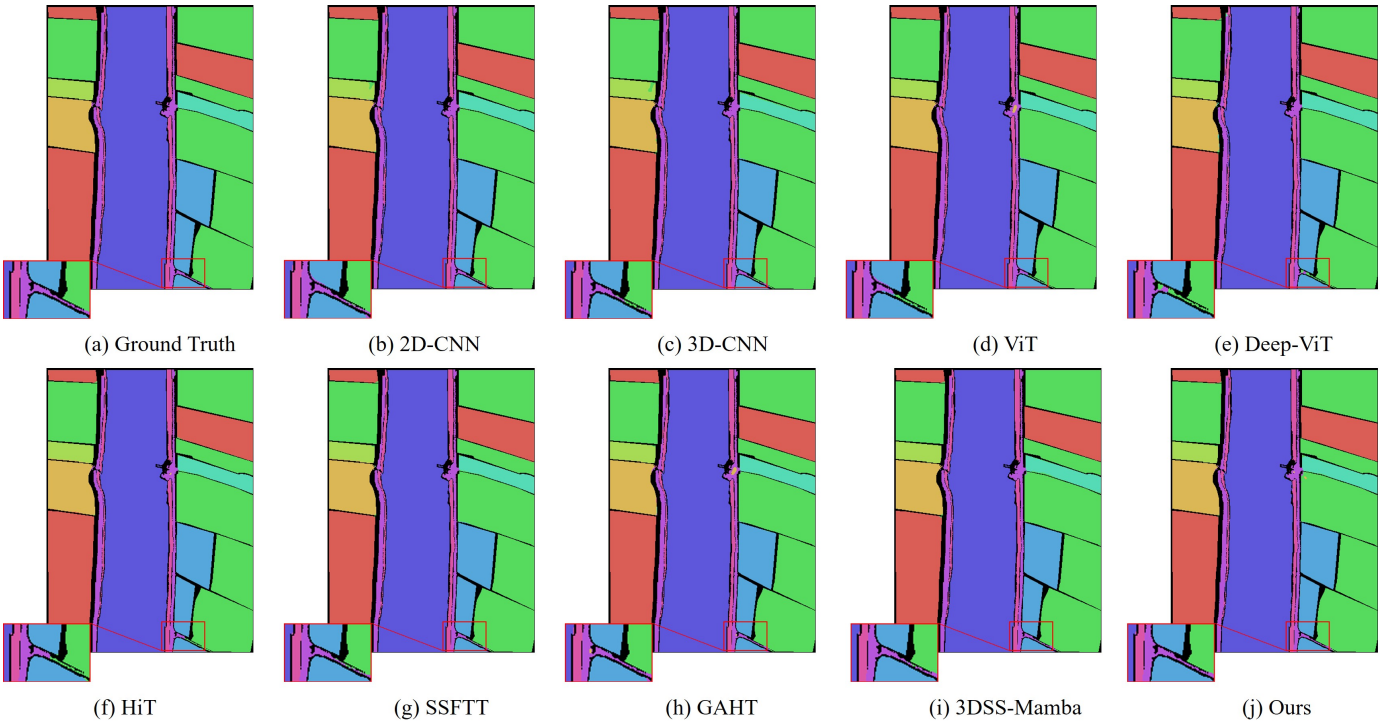


**Figure 6.** Classification maps obtained using different methods on the WHU-Hi-LongKou dataset (with 0.5% training samples).

Table 9 presents the classification performance of various methods on the Indian Pines and Houston 2013 datasets using only 5% of the training samples. On the Indian Pines dataset, the proposed method achieves the best results across all three metrics: Overall Accuracy (OA) at 92.18%, Average Accuracy (AA) at 77.61%, and Kappa coefficient at 91.07%, significantly outperforming other compared methods. For instance, compared to 3DSS-Mamba (which attains an OA of 89.23%), our method improves overall accuracy by nearly three percentage points, while also demonstrating notable advantages in AA and Kappa scores. Similarly, on the Houston 2013 dataset, the proposed method exhibits superior performance, achieving an OA of 96.55%, AA of 96.20%, and Kappa of 95.93%, surpassing all baseline methods. Notably, compared to DCTN, it attains an approximate one percentage point increase in OA, highlighting its strong classification capability and generalization performance. In summary, the proposed method demonstrates robust competitiveness in hyperspectral image classification across two distinct scenarios, confirming its effectiveness and robustness under low-sample training conditions.

We evaluated the model complexity of various methods on the Indian Pines dataset by measuring floating-point operations (FLOPs), number of parameters (Param), training time, and testing time, as summarized in Table 10. Notably, our proposed MRFP-Mamba method achieves strong performance across these metrics. From the perspective of computational complexity, traditional 2D-CNN models exhibit relatively low FLOPs and parameter counts, measuring 0.06 G and 1.67 MB respectively, accompanied by short training and testing times. In contrast, Transformer-based architectures such as the ViT model significantly increase computational demands, with FLOPs reaching 0.34 G and parameters totaling 6.54 MB, resulting in notably longer training and inference durations. Methods like GAHT and SSFTT demonstrate computational complexity that lies between conventional CNNs and ViT, reflecting moderate improvements in efficiency. Although the DCTN model achieves strong classification accuracy, it incurs the highest computational cost, with FLOPs of 2.95 G, 53.94 MB parameters, training time exceeding 441 s, and testing time of 7.93 s, indicating substantial resource consumption. Our proposed approach maintains moderate

FLOPs and parameter levels (0.73 G FLOPs, 1.64 MB parameters), with training and testing times of 80.36 s and 3.86 s, respectively, substantially lower than DCTN but higher than traditional 2D-CNN- and certain Transformer-based methods. In terms of classification performance, our method attains the highest Overall Accuracy (94.46%), Average Accuracy (89.88%), and Kappa coefficient (93.67%), demonstrating superior comprehensive effectiveness. Furthermore, MambaHSI and 3DSS-Mamba exhibit extremely low FLOPs and parameter counts (both at 0.01 G FLOPs and approximately 0.42 MB parameters), yet their training times are relatively long, around 360 s. Compared to these, our method achieves a favorable balance between high classification accuracy and computational efficiency, highlighting its practical applicability and efficiency advantages in hyperspectral image classification tasks.

**Table 9.** Classification results of the Indian Pines and Houston 2013 datasets with 5% training samples.

| Methods | Indian Pines | | | Methods | Houston 2013 | | |
|---|---|---|---|---|---|---|---|
| | OA | AA | $\kappa$ | | OA | AA | $\kappa$ |
| SSFTT | 90.11 ± 1.31 | 67.88 ± 1.79 | 88.65 ± 1.51 | SSFTT | 93.97 ± 0.57 | 93.53 ± 0.57 | 93.47 ± 0.62 |
| DCTN | 90.04 ± 1.05 | 68.93 ± 2.38 | 88.60 ± 1.21 | DCTN | 95.53 ± 0.54 | 95.10 ± 0.55 | 95.17 ± 0.58 |
| MambaHSI | 88.36 ± 1.25 | 72.69 ± 1.56 | 86.45 ± 1.47 | MambaHSI | 95.06 ± 0.35 | 94.86 ± 0.89 | 94.26 ± 0.45 |
| 3DSS-Mamba | 89.23 ± 1.07 | 75.58 ± 1.72 | 87.72 ± 1.21 | 3DSS-Mamba | 95.65 ± 0.40 | 95.59 ± 0.48 | 95.30 ± 0.43 |
| Ours | **92.18 ± 0.93** | **77.61 ± 2.16** | **91.07 ± 1.06** | Ours | **96.55 ± 0.45** | **96.20 ± 0.38** | **95.93 ± 0.49** |

**Table 10.** Comparison of computational complexity.

| Methods | FLOPs (G) | Param (MB) | Training Time (s) | Testing Time (s) | OA (%) | AA (%) | $\kappa$ (%) |
|---|---|---|---|---|---|---|---|
| 2D-CNN | 0.06 | 1.67 | 34.69 | 1.78 | 94.28 ± 1.41 | 85.52 ± 4.26 | 93.70 ± 1.65 |
| ViT | 0.34 | 6.54 | 52.23 | 2.78 | 91.73 ± 5.38 | 81.66 ± 3.03 | 90.61 ± 6.02 |
| GAHT | 0.31 | 0.68 | 46.79 | 2.84 | 93.56 ± 0.72 | 80.06 ± 2.46 | 92.64 ± 0.82 |
| SSFTT | 0.24 | 0.94 | 45.12 | 2.46 | 94.09 ± 0.97 | 83.06 ± 2.15 | 93.26 ± 0.12 |
| DCTN | 2.95 | 53.94 | 441.56 | 7.93 | 94.14 ± 0.63 | 81.91 ± 3.57 | 93.31 ± 0.73 |
| MambaHSI | 0.01 | 0.42 | 360.26 | 6.75 | 94.16 ± 1.59 | 87.54 ± 4.11 | 93.34 ± 1.81 |
| 3DSS-Mamba | 0.01 | 0.43 | 356.45 | 7.68 | 94.24 ± 0.59 | 86.79 ± 3.16 | 93.42 ± 0.68 |
| Ours | 0.73 | 1.64 | 80.36 | 3.86 | **94.46 ± 1.33** | **89.88 ± 2.54** | **93.67 ± 1.53** |

### *4.4. Ablation Studies*

#### 4.4.1. Ablation Study of the Input Size

As shown in Table 11, we investigate the impact of different input patch sizes on classification performance across three hyperspectral datasets. The input sizes range from $7 \times 7$ to $19 \times 19$. The results demonstrate that classification accuracy varies with the change in input size. In most cases, smaller input patches tend to achieve higher Overall Accuracy (OA) and Kappa coefficient ($\kappa$). This may be attributed to the fact that smaller patches can better focus on local spatial details, thus extracting more representative fine-grained features for land cover discrimination. In contrast, larger input regions might introduce redundant or even noisy information, which could weaken feature discriminability and negatively impact classification. These observations further validate the robustness and adaptability of the proposed MRFP-Mamba under different input configurations.

**Table 11.** Ablation study of the input size.

| Sizes | Indian Pines | | PaviaU | | Houston 2013 | | WHL | |
|---|---|---|---|---|---|---|---|---|
| | OA | $\kappa$ | OA | $\kappa$ | OA | $\kappa$ | OA | $\kappa$ |
| $7 \times 7$ | **97.99 ± 0.46** | **97.71 ± 0.53** | **98.97 ± 1.31** | **98.81 ± 1.41** | **98.41 ± 0.54** | **97.76 ± 0.71** | **99.63 ± 0.08** | **99.51 ± 0.11** |
| $9 \times 9$ | 97.50 ± 0.95 | 97.15 ± 1.08 | 98.78 ± 0.25 | 98.69 ± 0.27 | 98.25 ± 0.96 | 97.55 ± 1.27 | 99.58 ± 0.07 | 99.44 ± 0.09 |
| $11 \times 11$ | 97.26 ± 0.38 | 96.87 ± 0.43 | 98.46 ± 0.21 | 98.40 ± 0.35 | 97.89 ± 1.70 | 97.44 ± 2.27 | 99.35 ± 0.03 | 99.21 ± 0.15 |
| $13 \times 13$ | 95.71 ± 0.38 | 95.11 ± 0.43 | 98.39 ± 0.34 | 98.21 ± 0.37 | 97.80 ± 1.11 | 97.02 ± 1.47 | 99.19 ± 0.13 | 98.93 ± 0.17 |
| $15 \times 15$ | 94.46 ± 1.33 | 93.67 ± 1.53 | 98.26 ± 0.26 | 98.11 ± 0.28 | 97.68 ± 0.48 | 96.92 ± 0.64 | 99.03 ± 0.22 | 98.73 ± 0.28 |
| $17 \times 17$ | 92.59 ± 1.07 | 91.53 ± 1.23 | 95.51 ± 0.21 | 97.31 ± 0.29 | 96.82 ± 1.05 | 95.78 ± 1.39 | 98.71 ± 0.19 | 98.30 ± 0.25 |

#### 4.4.2. Ablation Study of Different Modules

As shown in Table 12, we conducted ablation experiments on the multi-receptive-field convolutional module. In the MRFP-Mamba model, there are three such modules composed of multi-receptive-field convolution and parallel Vision Mamba, with output channels of 256, 128, and 64, respectively. Accordingly, we designed ablation settings with pairwise combinations of convolution kernels of Sizes 1, 3, 5, and 7. The experimental results indicate that the combination of larger kernels (5 and 7) yields relatively poor performance (OA = 97.22 $\pm$ 0.26), which may be attributed to the fact that large kernels tend to introduce redundant information during feature extraction, thereby reducing the discriminability of local details. In contrast, the combination of all four kernel sizes (1, 3, 5, and 7) achieves the best result ($OA = 99.03 \pm 0.22$), demonstrating that the full fusion of multi-scale information effectively captures spatial features at different granularities. These results further validate the importance of the proposed multi-receptive-field convolutional module in enhancing the representational capacity of the model.

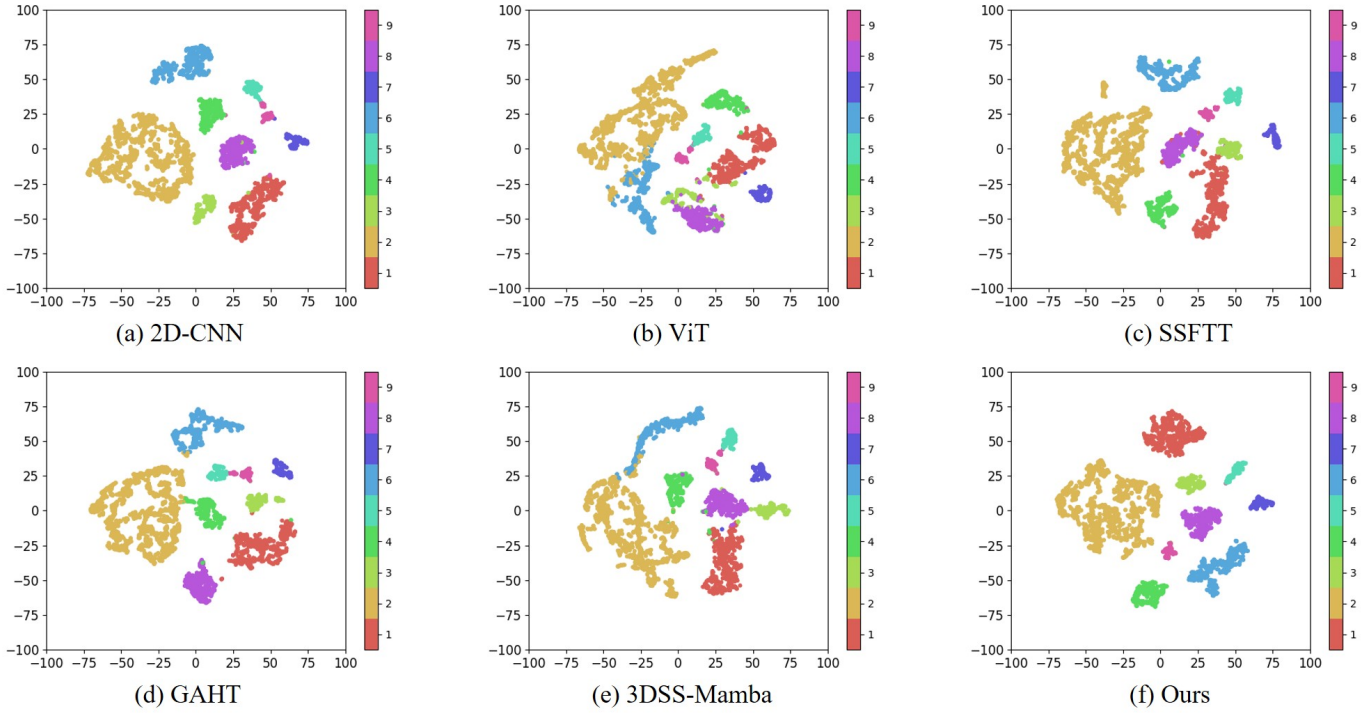


**Figure 7.** Visualization of t-SNE data analysis on the PaviaU dataset.

#### 4.4.3. Ablation Study of the Numbers of the Training Samples

We conducted a systematic series of experiments on four representative hyperspectral image (HSI) datasets, examining the impact of varying training sample sizes. As illustrated in Figure 8, a clear upward trend in Overall Accuracy (OA) is observed as the number of training samples increases. This trend indicates that larger training sets significantly enhance the model's discriminative capacity, particularly in distinguishing between different classes and reducing random misclassifications. These findings highlight the positive influence of expanded training data on the classification accuracy and robustness of the proposed method and further underscore the critical role of data scale in optimizing the performance of deep learning models.

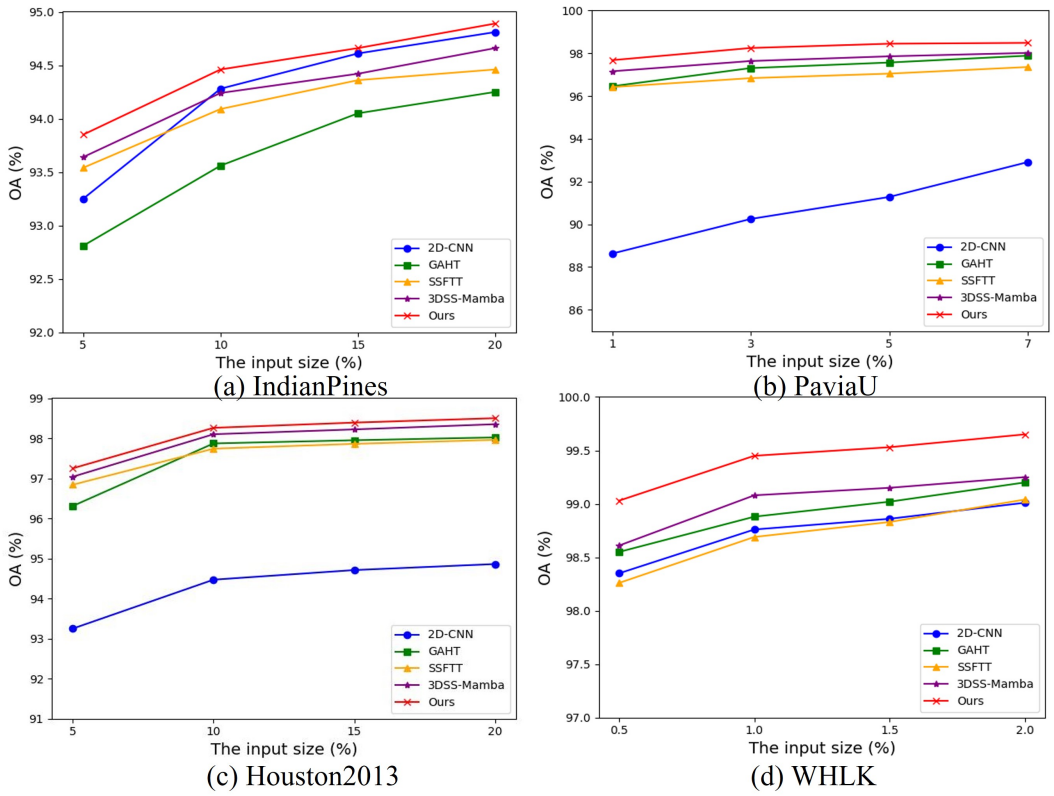


**Figure 8.** OA of different models with different percentages training samples on four datasets.

**Table 12.** Ablation study of the Different Modules on WHU-Hi-LongKou.

| Conv 1 × 1 | Conv 3 × 3 | Conv 5 × 5 | Conv 7 × 7 | OA (%) | AA (%) | $\kappa$ (%) |
|---|---|---|---|---|---|---|
| ✓ | ✓ | × | × | 97.91 ± 0.41 | 96.04 ± 0.89 | 97.86 ± 0.21 |
| ✓ | × | ✓ | × | 97.83 ± 1.21 | 95.87 ± 0.68 | 97.05 ± 0.62 |
| ✓ | × | × | ✓ | 98.56 ± 0.35 | 96.56 ± 1.20 | 98.03 ± 0.32 |
| × | ✓ | ✓ | × | 98.43 ± 0.33 | 96.24 ± 0.45 | 97.56 ± 0.63 |
| × | ✓ | × | ✓ | 98.32 ± 0.56 | 96.23 ± 0.89 | 97.89 ± 0.56 |
| × | × | ✓ | ✓ | 97.22 ± 0.26 | 95.46 ± 0.74 | 96.56 ± 0.87 |
| ✓ | ✓ | ✓ | ✓ | **99.03 ± 0.22** | **97.14 ± 0.90** | **98.73 ± 0.28** |

### *4.5. Discussion*

Through extensive experiments, we identified the strengths of MRFP-Mamba in its efficient spatial–spectral feature extraction and long-range dependency modeling. By leveraging the multi-receptive field convolutional feature extraction module, MRFP-Mamba is able to simultaneously capture fine-grained local features and broader contextual information, which enables it to perform excellently when handling the complexities of hyperspectral image (HSI) data. The parallel Vision Mamba architecture further enhances the model's ability to capture long-range dependencies while maintaining linear complexity, thus improving its performance in HSI classification.

Moreover, the integration of these components allows MRFP-Mamba to effectively combine spatial and spectral information. The multi-receptive field convolution captures detailed spatial features, while the Vision Mamba structure efficiently models the relationships between these features, leading to better spatial–spectral fusion and improved classification accuracy. Extensive experimental results validate its effectiveness and highlight its potential for practical applications in HSI analysis.

## 5. Conclusions and Fulture Work

In this paper, we propose a novel Multi-Receptive-Field Parallel Mamba (MRFP-Mamba) framework for hyperspectral image (HSI) classification. This method integrates a multi-receptive-field convolutional feature extractor with a parallel Vision Mamba module, achieving a seamless fusion of local feature extraction and long-range spectral dependency modeling. Specifically, the multi-receptive-field convolutional structure is employed to capture spatial features at different scales, while Vision Mamba efficiently models spectral dependencies, thereby enhancing feature representation capability. Extensive experiments on multiple benchmark HSI datasets demonstrate that MRFP-Mamba outperforms existing CNN-, Transformer-, and SSM-based methods in classification accuracy. In addition to enhancing classification performance, our method also demonstrates superior computational efficiency. Compared to 3DSS-Mamba, it reduces training time by 77.46% and testing time by 49.74%, making it more suitable for large-scale hyperspectral applications. Furthermore, the introduction of the Mamba architecture provides a new paradigm for hyperspectral classification, addressing the computational complexity issues of traditional self-attention mechanisms and enhancing model scalability.

In future work, we plan to further explore adaptive receptive field learning to enhance spatial feature extraction and investigate more efficient spectral modeling methods to reduce computational costs while maintaining high classification accuracy. Moreover, extending MRFP-Mamba to self-supervised and semi-supervised learning holds great potential for improving its applicability in real-world remote sensing tasks.

**Author Contributions:** Conceptualization, X.Y.; methodology, L.L.; validation, L.L., S.X. and S.L.; formal analysis, X.Y., L.L. and H.T.; data curation, L.L. and X.Y.; writing—original draft, X.Y. and L.L.; funding acquisition, W.Y. and X.H.; writing—review and editing, X.Y., S.X., S.L. and L.L.; visualization, L.L.; supervision, W.Y. and X.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by National Natural Science Foundation of China (NSFC) Fund under Grant 62301174 and Guangzhou basic and applied basic research topics under Grant 2024A04J2081 and Grant 2025A04J3375. This work also was supported in part by the National Natural Science Foundation of China under Grant No. 62462031, in part by the Natural Science Foundation of Jiangxi Province under Grant 20242BAB26023.

**Data Availability Statement:** The codes and parameters of our model are publicly available at https://github.com/Li-gzhu/MRFP-Mamba (accessed on 20 April 2025).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Gevaert, C.M.; Suomalainen, J.; Tang, J.; Kooistra, L. Generation of Spectral—Temporal Response Surfaces by Combining Multispectral Satellite and Hyperspectral UAV Imagery for Precision Agriculture Applications. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3140–3146. [CrossRef]
2. Hestir, E.L.; Brando, V.E.; Bresciani, M.; Giardino, C.; Matta, E.; Villa, P.; Dekker, A.G. Measuring freshwater aquatic ecosystems: The need for a hyperspectral global mapping satellite mission. *Remote Sens. Environ.* **2015**, *167*, 181–195. [CrossRef]
3. Chen, F.; Wang, K.; Voorde, T.V.D.; Tang, T.F. Mapping urban land cover from high spatial resolution hyperspectral data: An approach based on simultaneously unmixing similar pixels with jointly sparse spectral mixture analysis. *Remote Sens. Environ.* **2017**, *196*, 324–342. [CrossRef]
4. Sun, L.; Wu, F.; Zhan, T.; Liu, W.; Wang, J.; Jeon, B. Weighted Nonlocal Low-Rank Tensor Decomposition Method for Sparse Unmixing of Hyperspectral Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1174–1188. [CrossRef]
5. Wang, J.; Zhang, L.; Tong, Q.; Sun, X. The Spectral Crust project—Research on new mineral exploration technology. In Proceedings of the 2012 4th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Shanghai, China, 4–7 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1–4.

6. Zhang, Y.; Yan, S.; Jiang, X.; Zhang, L.; Cai, Z.; Li, J. Dual Graph Learning Affinity Propagation for Multimodal Remote Sensing Image Clustering. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5521713. [CrossRef]
7. Roy, S.K.; Krishna, G.; Dubey, S.R.; Chaudhuri, B.B. HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 277–281. [CrossRef]
8. Pasolli, E.; Melgani, F.; Tuia, D.; Pacifici, F.; Emery, W.J. SVM active learning approach for image classification using spatial information. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 2217–2233. [CrossRef]
9. Ham, J.; Chen, Y.; Crawford, M.M.; Ghosh, J. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 492–501. [CrossRef]
10. Ma, L.; Crawford, M.M.; Tian, J. Local Manifold Learning-Based *k* -Nearest-Neighbor for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 4099–4109. [CrossRef]
11. Song, W.; Li, S.; Kang, X.; Huang, K. Hyperspectral image classification based on KNN sparse representation. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2411–2414.
12. Zhang, Y.; Wang, X.; Jiang, X.; Zhang, L.; Du, B. Elastic Graph Fusion Subspace Clustering for Large Hyperspectral Image. *IEEE Trans. Circuits Syst. Video Technol.* 2025, *early access*.
13. Zhang, Y.; Jiang, G.; Cai, Z.; Zhou, Y. Bipartite Graph-based Projected Clustering with Local Region Guidance for Hyperspectral Imagery. *IEEE Trans. Multimed.* **2024**, *26*, 9551–9563. [CrossRef]
14. Sheykhmousa, M.; MahdianPari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308–6325. [CrossRef]
15. Villa, A.; Benediktsson, J.A.; Chanussot, J.; Jutten, C. Hyperspectral Image Classification with Independent Component Discriminant Analysis. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4865–4876. [CrossRef]
16. Liao, W.; Pizurica, A.; Scheunders, P.; Philips, W.; Pi, Y. Semisupervised Local Discriminant Analysis for Feature Extraction in Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 184–198. [CrossRef]
17. Wang, Q.; Meng, Z.; Li, X. Locality Adaptive Discriminant Analysis for Spectral-Spatial Classification of Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2077–2081. [CrossRef]
18. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE Computer Society: Piscataway, NJ, USA, 2017; pp. 764–773.
19. Zhao, C.; Zhu, W.; Feng, S. Superpixel Guided Deformable Convolution Network for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2022**, *31*, 3838–3851. [CrossRef] [PubMed]
20. Yang, X.; Cao, W.; Lu, Y.; Zhou, Y. Hyperspectral Image Transformer Classification Networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5528715. [CrossRef]
21. Ahmad, M.; Khan, A.M.; Mazzara, M.; Distefano, S.; Ali, M.; Sarfraz, M.S. A Fast and Compact 3-D CNN for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5502205. [CrossRef]
22. He, X.; Chen, Y.; Lin, Z. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]
23. Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; Douze, M. Levit: A vision transformer in convnet's clothing for faster inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12259–12269.
24. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5539014. [CrossRef]
25. Xu, Y.; Wang, D.; Zhang, L.; Zhang, L. Dual selective fusion transformer network for hyperspectral image classification. *Neural Networks* **2025**, *187*, 107311. [CrossRef]
26. Cheng, S.; Chan, R.; Du, A. CACFTNet: A Hybrid Cov-Attention and Cross-Layer Fusion Transformer Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [CrossRef]
27. Yao, J.; Hong, D.; Li, C.; Chanussot, J. SpectralMamba: Efficient Mamba for Hyperspectral Image Classification. *arXiv* **2024**, arXiv:2404.08489.
28. Lee, H.; Kwon, H. Going Deeper with Contextual CNN for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2017**, *26*, 4843–4855. [CrossRef] [PubMed]
29. Cao, X.; Zhou, F.; Xu, L.; Meng, D.; Xu, Z.; Paisley, J.W. Hyperspectral Image Classification with Markov Random Fields and a Convolutional Neural Network. *IEEE Trans. Image Process.* **2018**, *27*, 2354–2367. [CrossRef]
30. Wang, Z.; Chen, B.; Lu, R.; Zhang, H.; Liu, H.; Varshney, P.K. FusionNet: An Unsupervised Convolutional Variational Network for Hyperspectral and Multispectral Image Fusion. *IEEE Trans. Image Process.* **2020**, *29*, 7565–7577. [CrossRef]
31. Li, Y.; Zhang, H.; Shen, Q. Spectral-Spatial Classification of Hyperspectral Imagery with 3D Convolutional Neural Network. *Remote Sens.* **2017**, *9*, 67. [CrossRef]

32. Zhang, Y.; Yan, S.; Zhang, L.; Du, B. Fast Projected Fuzzy Clustering with Anchor Guidance for Multimodal Remote Sensing Imagery. *IEEE Trans. Image Process.* **2024**, *33*, 4640–4653. [CrossRef]
33. Hang, R.; Liu, Q.; Hong, D.; Ghamisi, P. Cascaded recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5384–5394. [CrossRef]
34. Mei, X.; Pan, E.; Ma, Y.; Dai, X.; Huang, J.; Fan, F.; Du, Q.; Zheng, H.; Ma, J. Spectral-spatial attention networks for hyperspectral image classification. *Remote Sens.* **2019**, *11*, 963. [CrossRef]
35. Yang, X.; Cao, W.; Tang, D.; Zhou, Y.; Lu, Y. ACTN: Adaptive Coupling Transformer Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5503115. [CrossRef]
36. Xu, Y.; Du, B.; Zhang, L. Self-Attention Context Network: Addressing the Threat of Adversarial Attacks for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2021**, *30*, 8671–8685. [CrossRef]
37. 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.* **2022**, *60*, 5518615. [CrossRef]
38. Ouyang, E.; Li, B.; Hu, W.; Zhang, G.; Zhao, L.; Wu, J. When Multigranularity Meets Spatial-Spectral Attention: A Hybrid Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–18. [CrossRef]
39. Xu, Y.; Xie, Y.; Li, B.; Xie, C.; Zhang, Y.; Wang, A.; Zhu, L. Spatial-Spectral 1DSwin Transformer with Groupwise Feature Tokenization for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5516616. [CrossRef]
40. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. CvT: Introducing Convolutions to Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 22–31.
41. Zhou, Y.; Huang, X.; Yang, X.; Peng, J.; Ban, Y. DCTN: Dual-Branch Convolutional Transformer Network with Efficient Interactive Self-Attention for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5508616. [CrossRef]
42. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752.
43. Lu, S.; Zhang, M.; Huo, Y.; Wang, C.; Wang, J.; Gao, C. SSUM: Spatial—Spectral Unified Mamba for Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 4653. [CrossRef]
44. 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]
45. Wu, R.; Liu, Y.; Liang, P.; Chang, Q. UltraLight VM-UNet: Parallel Vision Mamba Significantly Reduces Parameters for Skin Lesion Segmentation. *arXiv* **2024**, arXiv:2403.20035.
46. Liu, B.; Yu, X.; Zhang, P.; Tan, X.; Yu, A.; Xue, Z. A semi-supervised convolutional neural network for hyperspectral image classification. *Remote Sens. Lett.* **2017**, *8*, 839–848. [CrossRef]
47. Sharma, V.; Diba, A.; Tuytelaars, T.; Van Gool, L. *Hyperspectral CNN for Image Classification & Band Selection, with Application to Face Recognition*; Technical Report KUL/ESAT/PSI/1604; KU Leuven, ESAT: Leuven, Belgium, 2016.
48. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929. [CrossRef]
49. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S.J. Rethinking spatial dimensions of vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11936–11945.
50. 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]