MambaMSF: A Mamba-based Multi-scale Feature Fusion Method for Hyperspectral Image Classification

Junyan Xuan
Wuyi University
Jiangmen, China
2112424089@wyu.edu.cn

Zhenzhen Ren
Wuyi University
Jiangmen, China
2112424005@wyu.edu.cn

Bin Deng*
Wuyi University
Jiangmen, China
*Corresponding author: bindeng@wyu.edu.cn

Yikui Zhai
Wuyi University
Jiangmen, China
yikuizhai@163.com

***Abstract*—Hyperspectral image (HSI) classification faces challenges due to complex spectral-spatial characteristics and limited labeled samples. While Mamba-based models have shown significant progress in HSI classification by effectively handling long-sequence data and modeling spatial-spectral relationships, existing approaches primarily focus on single-scale feature extraction, neglecting multi-scale information. To address this, we propose a Mamba-based multi-scale feature fusion (MambaMSF) method. MambaMSF integrates multi-scale pixel patches to capture both local and global features, followed by an adaptive fusion mechanism that optimizes interactions between pixel-level and patch-level representations, enhancing classification accuracy. Experiments on three HSI benchmarks demonstrated that MambaMSF exceeds the comparison models.**

***Index Terms*—Hyperspectral image classification, Multi-scale feature fusion, Adaptive fusion mechanism, Mamba**

I. Introduction

The growing availability of remote sensing data and advances in sensing technologies have improved Earth surface analysis [1]. Hyperspectral imaging (HSI), with its high spectral resolution, continuous coverage, and strong material discriminability, enables the detection of subtle spectral signatures for resource management and ecological monitoring [2]. However, HSI classification is challenged by high dimensionality, spectral redundancy, inter-band correlations, intra-class variability, and limited labeled samples, complicating feature extraction and deep learning applications [3]. These challenges underscore the need for efficient network architectures that optimize both accuracy and computational cost.

To address these challenges, various deep network models have been proposed for HSI classification. CNN-based models excel in spatial feature extraction but are limited by local receptive fields, restricting their ability to capture long-range dependencies [4]. RNN-based models, though effective for sequential data, suffer from vanishing gradient issues when processing long spectral sequences [5]. Recently, transformer-based models have gained attention for their ability to model long-range dependencies and global contextual patterns in HSI [6]. However, their quadratic computational complexity with respect to input size imposes a high computational burden and hinders fine-grained spatial feature extraction at the pixel level.

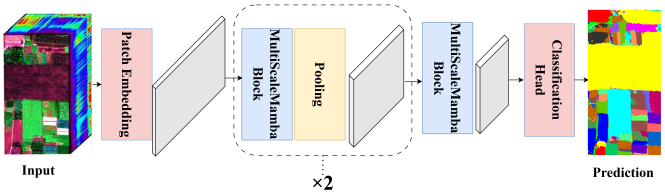

Figure 1. Framework overview of MambaMSF.

Recently, the Mamba model [7] has demonstrated significant potential for HSI classification, leveraging its strong long-range modeling capabilities and linear computational complexity. Notably, MambaHSI [8], a Mamba-based HSI classification model, has surpassed previous state-of-the-art models in accuracy. However, while numerous Mamba-based HSI models have emerged [9], they primarily focus on pixel-level sequence, often overlooking multi-scale spatial contextual relationships, which limits their performance in complex scenes. Exploring multi-scale analysis methods within the Mamba framework could enhance feature utilization across different scales, particularly in scenarios with limited labeled data, thereby improving model effectiveness.

In this paper, we propose MambaMSF, a Mamba-based multi-scale pixel patch feature fusion framework for HSI classification, designed to leverage multi-scale spatial-spectral information while addressing computational complexity in long-sequence data processing. The framework incorporates three key components: (1) a Mamba-based architecture with linear-time complexity and global dependency modeling, (2) a multi-scale extraction mechanism using a parallel multi-branch structure to overcome single-scale limitations, and (3) a dynamic adaptive fusion strategy for integrating multi-scale spatial-spectral features. These components synergistically process information across diverse receptive fields, enhancing discriminative power and practical applicability.

We conduct extensive experiments on three diverse real-world HSI benchmarks to evaluate the effectiveness of MambaMSF. The results demonstrate that MambaMSF outperforms state-of-the-art methods in both classification accuracy and

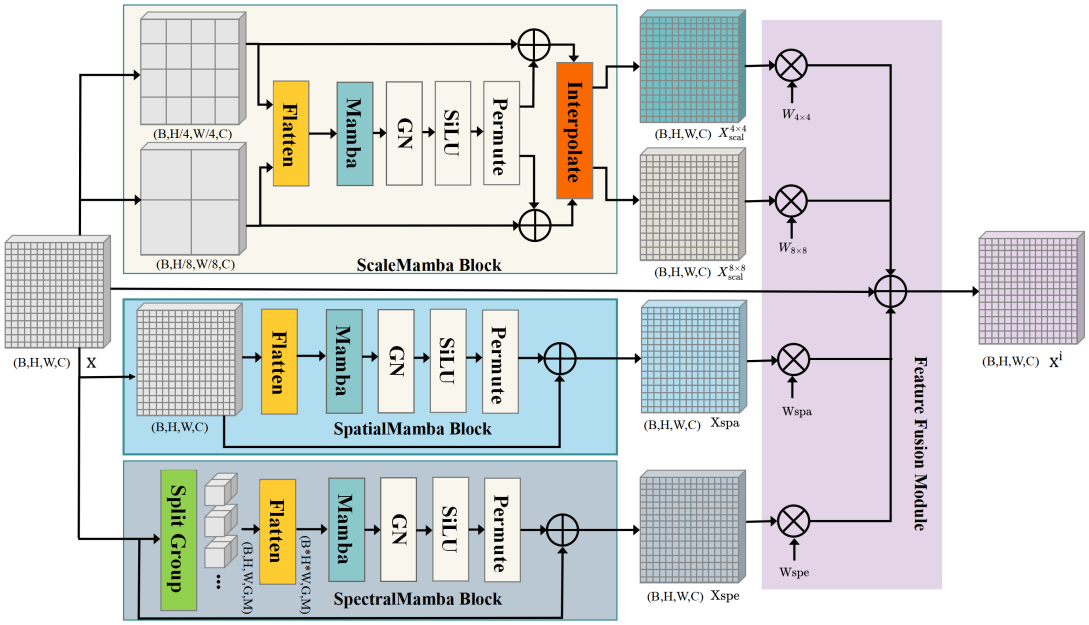

Figure 2. Schematic of the MultiScaleMamba model, comprising three parallel core modules—ScaleMamba Block, SpatialMamba Block, and SpectralMamba Block—followed by a Feature Fusion Module.

robustness, highlighting its strong potential for practical applications.

II. Methodology

A. Framework Overview

MambaMSF (Figure. 1) is an end-to-end deep learning model that directly predicts from raw input data, eliminating preprocessing and manual feature engineering. It embeds inputs into a patch format and processes them through alternating MultiScaleMamba Blocks and Pooling layers, where the Mamba module captures multi-scale feature dependencies. The extracted features are then fed into the classification head for final predictions. This design simplifies the workflow while preserving the model's ability to handle complex classification tasks.

B. MultiScaleMamba Model

The MultiScaleMamba model proposed in this study consists of four core components: SpatialMamba Block, SpectralMamba Block, ScaleMamba Block and Feature Fusion Module, as illustrated in Figure 2.

1) SpatialMamba Block: The SpatialMamba Block processes spatial features by first flattening the spatial dimensions of the input into a sequence. It then applies the Mamba model to capture long-range dependencies along the single-pixel dimension within the input space.

$$x_p = \text{Mamba}(\text{Flatten}(x)), \tag{1}$$

$$x_{spa} = \text{Reshape}(\text{SiLU}(\text{GN}(x_p))) + x, \tag{2}$$

where GN is the group normalization layer. Finally, a residual connection is employed to preserve the input information. This improves gradient propagation and reduces the degradation issue in deep networks.

2) SpectralMamba Block: The SpectralMamba module is developed to capture cross-channel spectral dependencies. Specifically, spectral features are divided into G groups, and the relationships among these groups are modeled to refine the spectral representations. The updated spectral features, guided by the extracted relationships, enhance feature learning. Finally, the processed output is combined with the input through a residual connection, preserving original information and improving the network's expressive power. The spectral feature extraction process in the SpectralMamba module is expressed as:

$$x_e = \text{Mamba}(\text{Flatten}(\text{SplitGroup}(x))), \tag{3}$$

$$x_{spe} = \text{Reshape}(\text{SiLU}(\text{GN}(x_e))) + x. \tag{4}$$

3) ScaleMamba Block: To maximize the use of spatial information in hyperspectral images, we designed a multi-scale pixel block feature module (Figure 2). Specifically, the input data of hyperspectral image undergoes 4×4 and 8×8 pixel downscaling operations to reduce the spatial resolution. The ScaleMambaBlock then processes the reduced feature maps to extract spatial dimension features. For example, consider the 4×4 scale:

$$x_4 = \text{Mamaba}(\text{Flatten}(P(x))), \tag{5}$$

$$x_{scal}^{4\times4} = \text{Interpolate}(\text{Reshape}(\text{SiLU}(\text{GN}(x_4))) + P(x)), \tag{6}$$

where P is the patch embedding module.

By extracting features at multiple spatial resolutions, this method captures multi-level spatial information from the input data. It generates feature maps at various scales and enhances

TABLE 1
PAVIAU DATASET, CATEGORY AND SAMPLE SETTING.

ID	Category	Train	Validation	Test
1	Asphalt	30	10	6591
2	Meadows	30	10	18609
3	Gravel	30	10	2059
4	Trees	30	10	3024
5	Metal sheets	30	10	1305
6	Bare soil	30	10	4989
7	Bitumen	30	10	1290
8	Bricks	30	10	3642
9	Shadows	30	10	907
Total		270	90	42416

classification performance. Finally, bilinear interpolation is used to upsample the output to the original resolution. This ensures consistency with feature maps from other dimensions.

4) Feature Fusion Module: Both spatial and spectral properties are crucial for hyperspectral image classification. Integrating these properties enhances classification performance. To address this, we designed a Feature Fusion Module, as shown in Figure 2. The module adaptively evaluates the importance of spatial and spectral features to guide the fusion process. The fusion process is expressed as follows:

$$x = x + x_{spe} \cdot w_{spe} + x_{spa} \cdot w_{spa} + x_{scal}^{4\times4} \cdot w_{4\times4} + x_{scal}^{8\times8} \cdot w_{8\times8} \quad (7)$$

Where w_{spa}, $w_{4\times4}$, and $w_{8\times8}$ represent the spatial fusion weights across different scales, and w_{spe} denotes the spectral fusion weights. These weights are randomly initialized and then updated through backpropagation to determine the final fusion results.

III. EXPERIMENTS

A. Datasets and Settings

To thoroughly evaluate the effectiveness of the proposed model, three widely recognized hyperspectral datasets are employed: the Pavia University Dataset (PaviaU), HanChuan [10], and HongHu [10].

1) PaviaU: Captured by the ROSIS sensor over the University of Pavia, this image has 103 spectral bands and a 610×340 resolution. The detailed information of PaviaU dataset is shown in Table. 1.

2) HongHu: Images were captured on November 20, 2017, in Honghu, Hubei, China, using a 17 mm Headwall Hyperspec sensor on a DJI Matrice 600 Pro UAV. The scene is a complex agricultural area with diverse crops, as shown in Table 2. The image size is 940×475 pixels with 270 bands (400–1000 nm).

3) HanChuan: The WHU-Hi-HanChuan dataset was collected on June 17, 2016, in Hanchuan, Hubei, using an Aibot X6 UAV with a 17 mm Headwall Nano-Hyperspec sensor. It covers urban and rural areas with buildings, water, and farmland, featuring seven crops like strawberrie, watermelons, and Cowpea, as shown in Table 3. The image size is 1217×303 pixels with 274 bands (400–1000 nm).

To reduce bias from random training sample selection, the experiment is repeated ten times, and results are averaged. In each trial, 30 training samples and 10 validation samples are randomly chosen from each class, with the remaining samples used for testing. The batch size is set to 1, and the Adam optimizer is used with a learning rate of 0.0003. The number of groups (G) and hidden dimensions (D) are set to 4 and 128, respectively, and the timescale and projection parameters in the Mamba module follow the settings in MambaHSI [8]. All experiments are run on a system with NVIDIA GeForce RTX 3090 GPUs, Intel Xeon Platinum 8352V CPUs, and 128 GB of RAM, using Python 3.9 and PyTorch 1.13.1 on Ubuntu.

TABLE 2
HONGHU DATASET,CATEGORY AND SAMPLE SETTING.

ID	Category	Train	Validation	Test
1	Red roof	30	10	14001
2	Road	30	10	3472
3	Bare soil	30	10	21781
4	Cotton	30	10	163245
5	Cotton firewood	30	10	6178
6	Rape	30	10	44517
7	Chinese cabbage	30	10	24063
8	Pakchoi	30	10	4014
9	Cabbage	30	10	10779
10	Tuber mustard	30	10	12354
11	Brassica parachinensis	30	10	10975
12	Brassica chinensis	30	10	8914
13	Small BC	30	10	22467
14	Lactuca sativa	30	10	7316
15	Celtuce	30	10	962
16	Film covered lettuce	30	10	7222
17	Romaine lettuce	30	10	2970
18	Carrot	30	10	3177
19	White radish	30	10	8672
20	Garlic sprout	30	10	3446
21	Broad bean	30	10	1288
22	Tree	30	10	4000
Total		660	220	385813

TABLE 3
HANCHUAN DATASET,CATEGORY AND SAMPLE SETTING.

ID	Category	Train	Validation	Test
1	Strawberry	30	10	44695
2	Cowpea	30	10	22713
3	Soybean	30	10	10247
4	Sorghum	30	10	5313
5	Waters pinach	30	10	1160
6	Watermelon	30	10	4493
7	Greens	30	10	5863
8	Trees	30	10	17938
9	Grass	30	10	9429
10	Red roof	30	10	10476
11	Gray roof	30	10	16871
12	Plastic	30	10	3639
13	Bare soil	30	10	9076
14	Road	30	10	18520
15	Bright object	30	10	1096
16	Water	30	10	75361
Total		480	160	256890

B. Results and Analysis

To assess the effectiveness of the proposed MambaMSF framework, its performance is compared with several benchmark methods, including SVM [11], FullyContNet [12], GSC-ViT [13], and MambaHSI [8]. The evaluation relies on quantitative metrics such as overall accuracy (OA), average accuracy

TABLE 4
QUANTITATIVE RESULTS FOR THE UNIVERSITY OF PAVIA DATASET, HONGHU DATASET AND HANCHUAN DATASET. BOLD IS THE BEST RESULT.

Methods	Pavia University			HongHu			HanChuan		
	OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)	OA(%)	AA(%)	Kappa(%)
SVM [11]	79.25±2.39	85.18±1.15	73.48±2.88	67.63±1.65	65.15±0.56	61.52±1.68	66.51±1.48	63.89±0.78	61.93±1.45
FullyContNet [12]	90.23±2.38	92.04±3.00	87.24±3.08	89.85±1.25	90.25±0.87	87.19±1.26	88.58±1.69	87.92±0.65	86.72±1.91
GSC-ViT [13]	91.11±1.57	92.11±0.61	88.27±1.98	91.24±0.14	91.99±0.35	89.89±0.21	83.90±0.57	76.14±0.42	81.18±0.65
MambaHSI [8]	95.36±0.83	95.67±0.79	95.25±2.55	94.24±1.47	94.56±0.66	94.24±1.64	89.08±1.74	87.79±1.13	86.67±4.07
MambaMSF (Ours)	**96.74±1.14**	**97.11±0.63**	**96.63±2.50**	**95.15±0.47**	**95.03±0.30**	**95.05±1.26**	**91.52±1.84**	**90.22±1.14**	**87.47±3.11**

(AA), and the kappa coefficient. These metrics collectively measure the classification's accuracy and consistency. Additionally, an ablation study is performed to validate the effectiveness of the multi-scale module.

1) Results of PaviaU dataset: MambaMSF achieves state-of-the-art performance on the PaviaU dataset (Table 4), outperforming all baseline methods, including MambaHSI, the runner-up, highlighting the efficacy of multi-scale fusion. Traditional methods like SVM show limited performance, while FullyContNet and GSC-ViT fall short of Mamba-based models, emphasizing the superiority of the proposed approach.

2) Results of HongHu dataset: On the HongHu dataset (Table 4), MambaMSF outperforms MambaHSI with a slight improvement, while GSC-ViT and FullyContNet remain competitive but inferior. SVM underperforms, further confirming MambaMSF's robustness in spectral-spatial feature extraction.

3) Results of HanChua dataset: MambaMSF demonstrates superior performance on the HanChuan dataset (Table 4), achieving the highest scores across all metrics and surpassing MambaHSI, particularly in complex scenarios, which underscores its adaptability. GSC-ViT and FullyContNet exhibit moderate performance but fail to match Mamba-based methods, revealing their constraints in modeling intricate spectral-spatial relationships. SVM yields significantly lower results compared to deep learning approaches, highlighting the inadequacy of traditional non-deep learning methods for complex hyperspectral classification tasks.

4) Visualisation of classification results: As shown in Table 5, the MambaMSF model achieves superior classification performance on three hyperspectral datasets: PaviaU, HongHu, and HanChuan. Its classification map closely aligns with the Ground Truth, exhibiting clear boundaries and reduced noise. Compared to SVM, FullyConvNet, GSC-ViT, and MambaHSI, MambaMSF demonstrates greater stability in complex scenes, highlighting its robustness in hyperspectral data processing.

TABLE 5
COMPARISON OF CLASSIFICATION RESULTS ON VARIOUS DATASETS.

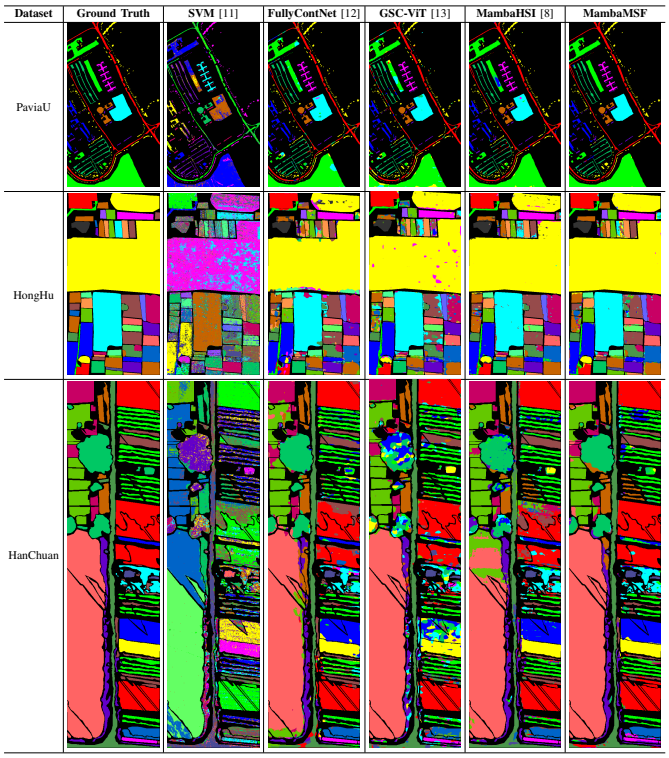

TABLE 6
ABLATION STUDIES OF KEY COMPONENTS OF OUR METHODOLOGY.

SpaMB	SpeMB	4×4 SpaMB	8×8 SpaMB	PaviaU (OA%)	HongHu (OA%)	HanChuan (OA%)
✓				95.21	89.87	86.35
	✓			88.46	75.77	79.30
		✓		94.01	93.42	89.26
			✓	92.31	92.21	89.31
✓	✓			95.36	94.24	89.08
✓	✓	✓		96.65	94.98	89.44
✓	✓	✓	✓	**96.74**	**95.15**	**91.52**

C. Ablation Studies

Ablation studies (Table 6) demonstrate the contributions of MambaMSF's components across three datasets. The Spatial-Mamba branch alone achieves strong results, highlighting its effectiveness in spatial feature extraction. The SpectralMamba branch performs less competitively, suggesting that spectral processing depends on spatial context. The addition of 4×4 or 8×8 SpatialMamba branches individually improves performance, supporting the benefit of multi-scale spatial analysis. Combining both branches enhances classification accuracy, with the 4×4 branch providing further improvement. The full configuration, incorporating all components, achieves the best performance, particularly on the HanChuan dataset. The 8×8 branch significantly boosts results, emphasizing its importance for complex spatial structures. These findings confirm the critical role of multi-scale fusion in MambaMSF's success.

IV. Conclusions

To fully leverage multi-scale information and capitalize on the advantages of the Mamba model, this paper proposes MambaMSF, a novel framework for hyperspectral image classification. The experimental results demonstrate that MambaMSF is an effective solution, outperforming both traditional and deep learning-based methods. By integrating multi-scale spectral-spatial information and leveraging the efficiency of the Mamba architecture, MambaMSF achieves superior performance. Ablation studies confirm the individual contributions of each component, with their combined integration yielding optimal results, particularly on datasets requiring robust spatial modeling. Future research could focus on dynamic scale adaptation and spectral refinement to further enhance the framework's capabilities.

References

[1] J. Li, X. Huang, and L. Tu, "Whu-ohs: A benchmark dataset for large-scale hersepctral image classification," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, p. 103022, Sep. 2022.

[2] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 270–294, Jun. 2022.

[3] V. Kumar, R. S. Singh, M. Rambabu, and Y. Dua, "Deep learning for hyperspectral image classification: A survey," *Computer Science Review*, vol. 53, p. 100658, Aug. 2024.

[4] B. Deng, S. Jia, and D. Shi, "Deep metric learning-based feature embedding for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 2, pp. 1422–1435, Feb. 2020.

[5] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[6] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[7] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First Conference on Language Modeling*, Aug. 2024.

[8] Y. Li, Y. Luo, L. Zhang, Z. Wang, and B. Du, "Mambahsi: Spatial–spectral mamba for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[9] H. Zhang, Y. Zhu, D. Wang, L. Zhang, T. Chen, Z. Wang, and Z. Ye, "A survey on visual mamba," *Applied Sciences*, vol. 14, no. 13, p. 5683, Jan. 2024.

[10] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "Whu-hi: Uav-borne hyperspectral with high spatial resolution (h2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with crf," *Remote Sensing of Environment*, vol. 250, p. 112012, Dec. 2020.

[11] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[12] D. Wang, B. Du, and L. Zhang, "Fully contextual network for hyperspectral scene parsing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[13] Z. Zhao, X. Xu, S. Li, and A. Plaza, "Hyperspectral image classification using groupwise separable convolutional vision transformer network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.