remote sensing

MDPI

Article

AFA–Mamba: Adaptive Feature Alignment with Global–Local Mamba for Hyperspectral and LiDAR Data Classification

Sai Li [1,2] and Shuo Huang [3,4,*]

[1] College of Mechanical and Electrical Engineering, Zaozhuang University, Zaozhuang 277160, China; lisai@uzz.edu.cn
[2] Zaozhuang Robot Autonomous Positioning and Navigation Technology Innovation Center, Zaozhuang 277160, China
[3] Key Laboratory of Infrared System Detection and Imaging Technology, Chinese Academy of Sciences, Shanghai 200083, China
[4] Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China
[*] Correspondence: huangshuo@mail.sitp.ac.cn

Abstract: The joint classification of hyperspectral imagery (HSI) and LiDAR data is an important task in the field of remote sensing image interpretation. Traditional classification methods, such as support vector machine (SVM) and random forest (RF), have difficulty capturing the complex spectral–spatial–elevation correlation information. Recently, important progress has been made in HSI-LiDAR classification using Convolutional Neural Networks (CNNs) and Transformers. However, due to the large spatial extent of remote sensing images, the vanilla Transformer and CNNs struggle to effectively capture global context. Moreover, the weak misalignment between multi-source data poses challenges for their effective fusion. In this paper, we introduce AFA–Mamba, an Adaptive Feature Alignment Network with a Global–Local Mamba design that achieves accurate land cover classification. It contains two main core designs: (1) We first propose a Global–Local Mamba encoder, which effectively models context through a 2D selective scanning mechanism while introducing local bias to enhance the spatial features of local objects. (2) We also propose an SSE Adaptive Alignment and Fusion (A^2F) module to adaptively adjust the relative positions between multi-source features. This module establishes a guided subspace to accurately estimate feature-level offsets, enabling optimal fusion. As a result, our AFA–Mamba consistently outperforms state-of-the-art multi-source fusion classification approaches across multiple datasets.

Keywords: hyperspectral image (HSI); light detection and ranging (LiDAR) data; joint classification; data fusion

check for updates

Citation: Li, S.; Huang, S. AFA–Mamba: Adaptive Feature Alignment with Global–Local Mamba for Hyperspectral and LiDAR Data Classification. *Remote Sens.* **2024**, *16*, 4050. https://doi.org/10.3390/rs16214050

Academic Editor: Salah Bourennane

Received: 1 September 2024
Revised: 18 October 2024
Accepted: 28 October 2024
Published: 30 October 2024

1. Introduction

Remote sensing and imaging technologies have developed rapidly in recent years, as more types of data sources have been applied to the extraction and analysis of geomorphological information [1–6]. The joint classification of hyperspectral imagery (HSI) and LiDAR data has attracted more and more interest from researchers [7]. Hyperspectral remote sensing is a form of technology that is capable of acquiring fine spectral information, which contains data in tens or even hundreds of spectral bands for achieving accurate identification of the ground [8]. LiDAR technology determines the distance and position of a target object by transmitting laser pulses and measuring their return time, thus obtaining elevation information from the ground [9–11]. With the increasing improvements of both forms of imaging technology, HSI-LiDAR joint classification plays an important role in precision agriculture, mineral development [12], environmental monitoring [13], urban planning [14], and other fields. By improving the accuracy of land cover and object classification, this joint method can enhance decision-making processes in resource management, disaster prevention, and land-use planning.

Early remote sensing technologies [15,16] were relatively limited and the type of data acquired was homogenous. As an important field of remote sensing applications, landscape classification and recognition are largely restricted in accuracy and efficiency by their data sources [16]. For example, HSI captures the radiation variations of a target object within a continuous spectral range and combines them with spatial information to achieve an accurate response to the physical and chemical properties of the object [17]. However, in some complex scenarios, different objects in the image may have similar spectral curves, and the lower spatial resolution of HSI makes it challenging to distinguish structures [18]. In such cases, combining elevation information from LiDAR data can enable the accurate identification of different land cover categories [19]. Therefore, studying how to effectively leverage the strengths of both data types for the joint classification of HSI and LiDAR data is significant for raising the accuracy and efficiency of land cover classification.

In past decades, many researchers have focused on transferring modeling methods commonly used in machine learning and pattern recognition to remote sensing image classification tasks [20,21], such as support vector machine (SVM) [22], random forest (RF) [23], Extreme Learning Machine (ELM) [24], etc. In addition, to better utilize spectral–spatial information, researchers have designed a series of ingenious classification models. Hang et al. [25] pioneered a novel approach termed matrix-based spatial spectral feature representation. Wang et al. [26] introduced an advanced approach known as discriminative multi-kernel learning, which learns to determine the optimal combination of kernels from a set of predefined basic kernels by maximizing the divisibility of the reproduced kernels in Hilbert space. It can greatly improve HSI classification performance without strictly limiting the selection of basic kernels. However, these traditional methods often require a certain amount of prior knowledge and complex parameter tuning, which can lead to suboptimal performance when dealing with complicated samples and strongly non-linear data [18].

Recently, Deep Neural Networks (DNNs) have emerged as a dominant force, demonstrating remarkable efficacy in the realm of remote sensing image analysis and processing [8,27–30]. In particular, Convolutional Neural Networks (CNNs), which are able to accurately distinguish between different feature classes by automatically learning and extracting the deep features of images, greatly improve the discriminability of remote sensing target features. Compared with traditional classification methods based on manual feature extraction, DNNs can handle more complex data as well as adapt to different shooting angles and scale variations [31]. For example, Paoletti et al. [30] proposed a novel convolutional residual pyramid network to achieve fast convergence and high accuracy classification models under complex HSI remote sensing data. Furthermore, Roy et al. [29] introduced a novel superimposed hybrid architecture comprising both 3D-CNN and 2D-CNN, which significantly enhanced the fusion of spectral and spatial information. However, the local receptive field of CNN limits its ability to capture the global information of the spectrum, resulting in suboptimal classification accuracy. To further enhance the model's global feature extraction capability, researchers [32–35] introduced the Vision Transformer (ViT) for HSI classification, capturing global spectral information and better representing the relationship between spectral and spatial features. For instance, Mei et al. [35] presented the Group-Aware Hierarchical Transformer. This novel model underscores both local and global interactions of spectral–spatial information by introducing a novel grouped pixel embedding. Although ViT can represent the global dependence of spectra well, its computational quadratic complexity makes it difficult to use for large-scale datasets. Recently, a new method known as Mamba [36] has been introduced, employing state-space models (SSMs) to efficiently capture global semantic information with minimal computational overhead, thus achieving linear complexity operations. Mamba-based methods [37–39] have shown great potential in remote sensing classification tasks. For example, RS–Mamba [39] introduced an innovative omnidirectional selective scanning module, which selectively scans remote sensing images in multiple directions. This enables the extraction of large-scale spatial features from various orientations. Zhou's team proposes a novel centralized

Mamba–Cross-Scan (MCS) mechanism for converting HSI images into sequence data to enhance feature generation and focusing for fine-grained recognition of feature categories.

Nevertheless, in recent years, single-source data classification has become increasingly unable to meet the identification needs of complex landform scenarios in terms of data comprehensiveness, credibility, and prediction accuracy. Benefiting from the powerful capabilities of deep learning models in feature extraction and fusion, CNN-based and ViT-based methods show great advantages in the joint classification of HSI and LiDAR data. Hang's team [40] introduced a coupled CNN model capable of learning distributed spectral–spatial features from HSI while simultaneously capturing elevation information from LiDAR data. This innovative model integrates heterogeneous features through a parameter sharing strategy, marking a substantial departure from conventional approaches. Zhang et al. [41] designed an Interleaved Perceptual Convolutional Neural Network (IP-CNN), which can introduce HSI perception constraints and LiDAR perception constraints into the integration of multi-source structural information, and achieved satisfactory results in small sample training. Lu et al. [42] unveiled a revolutionary method called Coupled Adversarial Learning-based Classification (CALC). Their approach pioneers an adversarial setup between a dual generator and a discriminator. This mechanism adeptly extracts similar higher-order semantic information and modality-specific complementary details, introducing a paradigm shift in classification methodologies. Zhao et al. [43] proposed a new two-branch approach that combines a CNN and Transformer encoder in a hierarchical form to achieve effective joint classification of heterogeneous information from multiple sources. Additionally, Sun and colleagues [44] developed a groundbreaking multi-scale lightweight fusion network based on this, distinctively free from attention mechanisms. This novel architecture not only drastically reduces training parameters but also effectively captures multi-scale depth and high-order features.

Although DNN-based methods have made significant progress in the joint classification of HSI and LiDAR data, the existing methods still face a series of challenges when dealing with complex feature environments [45–47]. Initially, given the varied imaging mechanisms of distinct sensors, the model necessitates the separate processing of diverse source data to ensure the effective representation of specific information. Second, the application scenarios of HSI and LiDAR sensors are distinct [48], resulting in different performance focuses for different data types. This requires the training of classification models to fully consider differences and flexibly capture fine-grained local details and global information. Finally, for formal variations in different data features, the classification model also requires a strong adaptive complementary ability to realize the effective fusion and complementarity of spectral–spatial–elevation features, so as to achieve accurate scene classification.

To tackle the aforementioned challenges, we introduce a comprehensive joint classification method for HSI and LiDAR data, termed Adaptive Feature Alignment Network with a Global–Local Mamba. It contains two branches: an HSI branch and a LiDAR feature processing branch. Specifically, we propose an SSE feature extraction module to process HSI and LiDAR data separately to mine more spectral–spatial and elevation features. It can fully extract adaptive feature information while considering data discretization. In addition, to capture semantic information at multiple scales, we also propose Global–Local Mamba modules for the two types of data sources to achieve dynamic awareness of spectral–space–elevation information. Finally, to eliminate the inherent distinction between the expressions of the two types of data in the upper and lower branches, we design an SSE Adaptive Alignment and Fusion (A^{2}F) module that alleviates the feature differences and spatial mismatches between the data sources by learning the discriminative features of the both types of data to adapt to the calibration differences.

To summarize, the main contributions are as follows:

(1) We propose a joint HSI-LiDAR classification method called Adaptive Feature Alignment Network with a Global–Local Mamba (AFA-Mamba), which uses a hybrid two-

branch CNN architecture to accurately extract 3D spectral–spatial information from HSI data while simultaneously capturing 2D elevation information from LiDAR data.

(2) The proposed Global–Local Mamba module is designed to be both efficient and effective in processing data. It operates by dynamically capturing and analyzing spectral, spatial, and elevation information, allowing the model to adaptively focus on different types of information depending on the context.

(3) We design a novel spectral–spatial–elevation Adaptive Alignment and Fusion module to adaptively recalibrate the differences by learning the discriminative features of the two types of data, thereby effectively mitigating the problems due to feature differences and spatial mismatches between the data sources. It ensures the accuracy and consistency of HSI and LiDAR information in the fusion process.

(4) AFA–Mamba demonstrates superior classification performance compared to several existing SOTA methods. The experimental results across all three datasets consistently validate the outstanding performance of our approach.

This paper's subsequent sections are structured as follows: Section 2 offers a comprehensive overview of our AFA–Mamba, elucidating its core components and operational principles. Moving forward, Section 3 meticulously outlines the experimental datasets, delineates the experimental setup, and conducts a comprehensive analysis of the classification outcomes. Lastly, Section 4 succinctly summarizes the study's conclusions and paves the way for future research directions.

2. Methodology

Figure 1 illustrates the comprehensive workflow of our proposed AFA–Mamba. It includes HSI and LiDAR Data Preprocessing, a spectral–spatial–elevation (SSE) Feature Extraction Module, a Global–Local Mamba, and an SSE Adaptive Alignment and Fusion (A^2F) Module. We introduce the details of each module in the following sections.

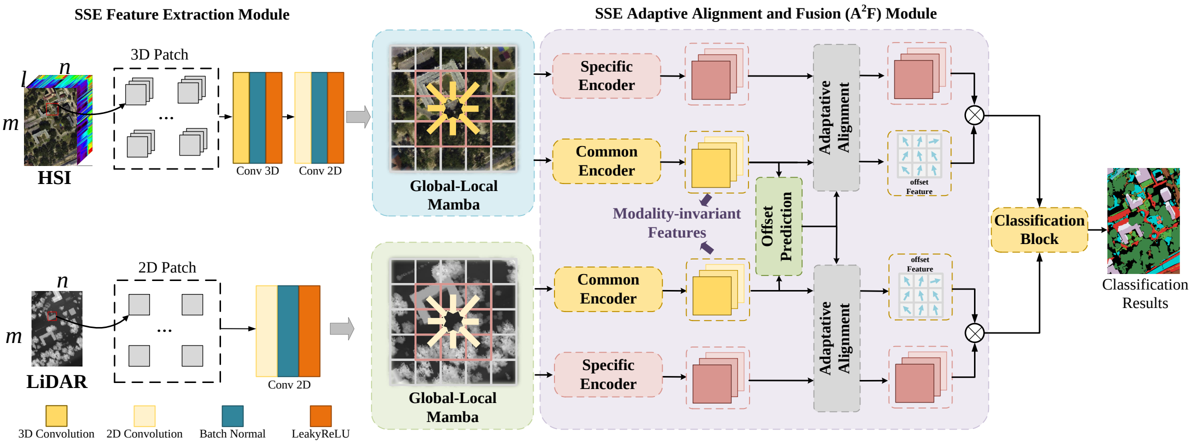


Figure 1. The architecture of our proposed AFA–Mamba. The proposed method consists of two branches: the upper branch performs fine feature extraction for the spectral–spatial information from HSI, and the lower branch performs correlation feature extraction for the elevation information from LiDAR. Finally, the feature level of the multi-source information is calibrated by the A^2F module to realize the high-precision classification of the landform categories.

2.1. HSI and LiDAR Data Preprocessing

This module is provided with HSI data, denoted as $X_H \in \mathbb{R}^{m \times n \times l}$, and LiDAR data, denoted as $X_L \in \mathbb{R}^{m \times n}$, covering the same area of the Earth's surface, where m and n represent the spatial dimensions, and l represents the number of spectral bands in the HSI data. The HSI data typically contain numerous spectral bands capable of conveying more valuable information, with each pixel being representable by a one-hot category vector. However, the large size of spectral data leads to expensive computational costs. Therefore, we extracted the first k principal components from HSI through PCA to reduce the number of the spectral band from l to k while keeping its spatial dimensions unchanged, which is defined as $X_H^{pca} \in \mathbb{R}^{m \times n \times k}$.

Next, for $X_H^{pca} \in \mathbb{R}^{m \times n \times k}$ and $X_L \in \mathbb{R}^{m \times n}$, we employed a window with a patch size of $s \times s$, performed 3D and 2D patch extraction, and obtained the 3D small patches $X_H^P \in \mathbb{R}^{s \times s \times k}$ and 2D small patches $X_L^P \in \mathbb{R}^{s \times s}$, respectively. We determined the identity of each patch by its central pixel. For edge pixels that may not meet the window size, we performed a padding operation on these pixels and defined the width as $(s-1)/2$. Lastly, we filtered out the pixel blocks with label 0, and then partitioned the remaining samples into training and test sets.

2.2. Spectral–Spatial–Elevation Feature Extraction Module

By exploiting the distinct advantage of CNNs in context modeling and feature extraction, they excel in managing vast remote sensing datasets. CNNs efficiently capture spectral–spatial correlations in HSI data and comprehensively extract elevation details from LiDAR data. Hence, our approach employs a 3D-CNN to extract intricate spectral–spatial features from high-dimensional 3D patches, enabling precise local modeling. Concurrently, we utilize a separate 2D-CNN to focus specifically on extracting elevation features from LiDAR data.

As illustrated in Figure 1, for the HSI data $X_H^{pca} \in \mathbb{R}^{m \times n \times k}$, we first used Conv3-D to extract discriminative spectral–spatial features from the HSI data, and convert the spatial dimensions of the generated feature cube into a two-dimensional vector. Subsequently, we utilized Conv2-D to mitigate the redundancy inherent in both spectral and spatial information. Unlike HSI data processing, we employed two Conv2-D convolutions to extract the surface elevation information for the LiDAR data $X_L \in \mathbb{R}^{m \times n}$, whose dimensions and convolution kernel size are 16@3 × 3 and 64@3 × 3, respectively. To speed up the training process and increase non-linear learning capabilities, we added layer normalization and ReLU activation functions after each convolution layer.

2.3. Global–Local Mamba Encoder

In hyperspectral classification tasks, remote sensing images are typically captured by satellites from a top-down perspective, resulting in large spatial features with arbitrary orientations. Therefore, the global context modeling of remote sensing images is crucial for classification tasks. However, relying solely on global features is often insufficient for identifying ground vegetation, as it usually requires local representation enhancement to handle small regions with severe boundary adhesion. Although state-space models (SSMs) have shown promise in long-sequence modeling, they face challenges in effectively combining local invariants and global context in visual data. To address this, we propose Global–Local Mamba, which introduces local bias to enhance the spatial features of local objects.

State-space models (SSMs) are a type of continuous system that map one one-dimensional function or sequence to another through specific implicit latent states. In other words, it is a model that uses latent states to transform and model sequential data. Given a one-dimensional sequence $x(t) \in \mathbb{R}$, which is projected through a hidden state $h(t) \in \mathbb{R}^m$ to form a new one-dimensional sequence $y(t) \in \mathbb{R}$, the entire system can be represented as follows:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t) + \mathbf{D}x(t). \end{aligned} \tag{1}$$

where **A** is the state matrix, which defines how the current hidden state affects its own rate of change. **B** and **C** are the input matrix and output matrix, respectively. **D** is the feed-through (or direct transmission) matrix.

In practice, the continuous system described above should be discretized under the zero-order hold assumption, converting the matrices **A** and **B** into their discrete forms for a time scale $\Delta \in \mathbb{R}^+$:

$$\begin{aligned}\overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}),\\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I})\cdot\Delta\mathbf{B}.\end{aligned} \tag{2}$$

where Δ is the step size. Therefore, the discretized version of SSMs can be expressed as:

$$h_t = \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \quad y_t = \mathbf{C}h_t + \mathbf{D}x_t. \tag{3}$$

However, the current system remains static when handling different inputs. To overcome this limitation, Mamba introduces a selective state-space model that allows parameters to adapt based on the input, enhancing selective information processing across sequences. This parameter selection mechanism can be expressed as:

$$\overline{\mathbf{B}} = f_{\mathbf{B}}(x_t), \quad \overline{\mathbf{C}} = f_{\mathbf{C}}(x_t), \quad \Delta = \theta_{\mathbf{A}}(\mathbf{P} + f_{\mathbf{A}}(x_t)). \tag{4}$$

Here, $f_{\mathbf{B}}(x_t)$, $f_{\mathbf{C}}(x_t)$, and $f_{\mathbf{A}}(x_t)$ are linear functions that expand features to the dimension of the hidden state. Since SSMs are tailored for long sequences, they have limitations in capturing detailed local information. In addition, VMamba [25] and Vim, propose specific position-aware scanning strategies to preserve the structure of 2D images. However, these directed sequences overlook the visual information within the pixel neighborhood. We explore a Global–Local Mamba Encoder, where global perception is received before focusing on details, thereby compensating for the lack of local information.

As shown in Figure 2, our Global–Local Mamba Encoder adopts the linear and state-space model flow, inspired by the usage of similar structures in Transformers and Mamba. Furthermore, we introduce a component, $\mathrm{eg}(\mathbf{L}_p)$, to enhance local bias and correct the causal relationships between neighborhood data, thereby improving the original SSM by maintaining local 2D dependencies:

$$\begin{aligned}h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t,\\ y_t &= \mathbf{C}h_t + \mathbf{D}x_t + \mathrm{eg}(\mathbf{L}_p).\end{aligned} \tag{5}$$

where $\mathrm{eg}(\mathbf{L}_p)$ operates independently of the hidden state space.

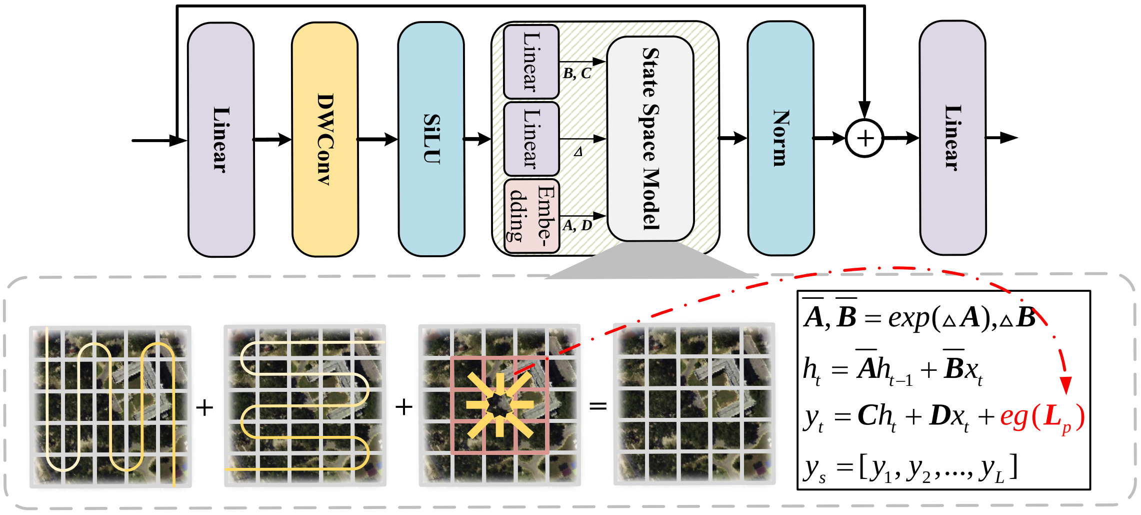

Linear
DWConv
SiLU
Linear
Linear
Embe-dding
B, C
Δ
A, D
State Space Model
Norm
+
Linear
+
+
=
$\overline{A}, \overline{B} = exp(\Delta A), \Delta B$
$h_t = \overline{A}h_{t-1} + \overline{B}x_t$
$y_t = Ch_t + Dx_t + eg(L_p)$
$y_s = [y_1, y_2, ..., y_L]$

Figure 2. Illustration of Global–Local Mamba Encoder. DWConv denotes Depth-Wise Convolution and SiLU denotes the Sigmoid Linear Unit function. Also, **A** is the state matrix, **B** is the input matrix, **C** is the output matrix, and **D** is the feed-through matrix. We introduce a selective state-space model with the premise of allowing the spatial features of localized objects to be enhanced by adjusting parameters according to the inputs. This enhances selective information processing across sequences.

2.4. SSE Adaptive Alignment and Fusion Module

Efficiently fusing multi-source remote sensing data stands as a pivotal factor in significantly enhancing remote sensing classification performance. However, the fusion of discriminative features from different sources remains a major challenge due to the inherent feature variation and spatial misalignment. To this end, we propose a spectral–spatial–elevation (SSE) Adaptive Alignment and Fusion (A^{2}F) module to adaptively adjust the relative positions between multi-source features. It achieves optimal fusion by building a guided subspace to accurately estimate the feature-level offset.

Figure 3 illustrates the detailed workflow of the SSE A^{2}F module. Specifically, we extract modality-invariant features and modality-specific features from both the HSI and elevation feature. To mitigate misalignment in the multi-source feature space, we use the modality-invariant features to predict spatial offsets in the multi-source data. These predicted offsets then adaptively adjust the spatial positions of the modality-specific features to achieve optimal feature alignment.

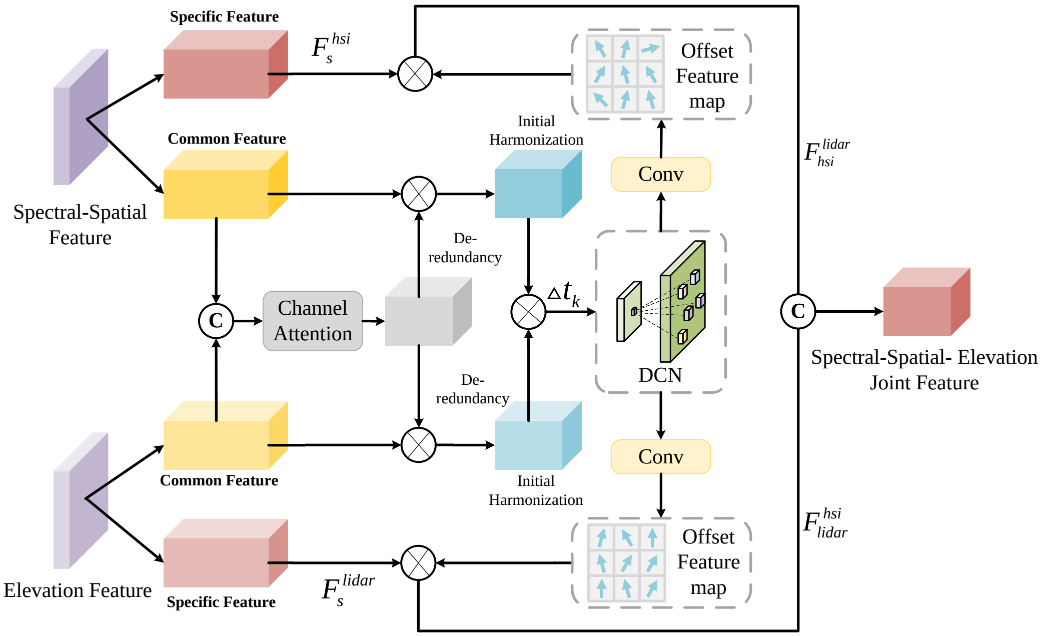


Figure 3. Illustration of SSE A^{2}F Module. F in the figure denotes the individual feature representations of the intermediate processes of the network. The module consists of two sets of upper and lower feature inputs: spectral–spatial features and elevation features. The module can adaptively adjust the relative positions between multi-source features, and achieve optimal fusion by constructing a bootstrap subspace to accurately estimate feature-level offsets.

For each input of the spectral–spatial–elevation features $\mathcal{F}_{hsi} \in \mathbb{R}^{H\times W\times C}$ and $\mathcal{F}_{lidar} \in \mathbb{R}^{H\times W\times C}$, we fuse their modality-invariant features to form a guided subspace $\mathbb{C} = [\mathbb{C}_{hsi}, \mathbb{C}_{lidar}]$. Then, through channel attention projection to this guided subspace, we amplify the information-rich bands while reducing the influence of irrelevant bands.

$$\mathbb{C} = Concat(\mathcal{F}_{hsi}, \mathcal{F}_{lidar}), \quad \hat{\mathbb{C}} = \mathcal{C}_m([\mathbb{C}_{hsi}, \mathbb{C}_{lidar}]; \theta^c). \tag{6}$$

where $\mathcal{C}_m$ is the channel attention operation. θ^c denotes the assigning of a separate parameter to each feature.

To learn the spatial shift between modalities from the guided subspace as a strong prior for subsequent alignment and fusion, we use deformable convolutions to achieve implicit offset compensation. Taking into account the varying contributions of each spectral and sub-band region to classification, we first introduce a modulation scalar Δt_k learned from $\hat{\mathbb{C}}$, which dynamically aggregates information from the surrounding area of the corresponding position p. As depicted in Figure 3, given the center-sampled value $x(p)$ in the modality-specific feature $\mathcal{F}_s^{\text{hsi}}$ or $\mathcal{F}_s^{\text{lidar}}$ and the modulation scalar Δt_k, the corresponding value $y(p)$ in the aligned feature F_{hsi}^{lidar} can be derived as follows:

$$y(p) = \sum_{k=1}^{K} \mathrm{w}_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta t_k \tag{7}$$

where K, p_k, and w_k represent the number of kernel weights, and the fixed offset of the k-th position, respectively. Finally, the two aligned features F_{hsi}^{lidar} and F_{lidar}^{hsi} are fused and sent to the classification head.

2.5. Classification Block

Following the SSE A^{2}F module processing, the output is passed through a multi-layer perceptron (MLP) layer for the final classification. This MLP comprises two linear layers featuring Gaussian Error Linear Units (GELUs). Notably, the last linear layer integrates a softmax function to derive conclusive labels for classification.

3. Experiment and Analysis

In this section, we elaborate on the various configurations and results of our experiments. First, we give the specific dataset, experimental configurations, and evaluation indicators for fair comparison. Then, we conduct quantitative experiments on three representative multi-modal datasets to demonstrate the advanced performance of our AFA–Mamba. In addition, comprehensive ablation studies are performed to explain the role of each component in the proposed AFA–Mamba. Finally, the quantitative and visual results demonstrate the superior performance of our AFA–Mamba over existing SOTA methods in remote sensing classification tasks.

3.1. Dataset Description

Following the previous methods [28,42], we select three datasets that are widely utilized for remote sensing classification tasks: MUUFL Gulfport, Trento, and the Augsburg datasets, which contain include two types of data (HSI and LiDAR). Table 1 reports the details of all datasets used, including the number of samples and the land cover ground truth categories.

(1) MUUFL Gulfport Dataset: Gathered at the University of Southern Mississippi Gulf Park campus, this dataset utilizes a reflective optical system spectrometer sensor for imaging. The dataset integrates both hyperspectral imaging (HSI) and LiDAR data. The HSI data feature 72 spectral bands spanning from 0.38 to 1.05 μm. Additionally, the LiDAR data are composed of two rasters that function at a wavelength of 1.06 μm. The MUUFL Gulfport dataset has a pixel size of 325 × 220 and covers 11 distinct land cover categories. Due to the significant noise present in the first and last 8 spectral bands, we deleted these bands during training to improve data quality. Figure 4 presents a visual representation of the MUFFL dataset, including pseudo-color composite images (HSI data), grayscale images (LiDAR data), and depictions of the various land cover categories within the dataset.

(2) Trento Dataset: This dataset originates from the rural area surrounding Trento, a city in southern Italy, and includes six distinct scenes. The hyperspectral imaging data and LiDAR data were acquired using the AISA Eagle system's hyperspectral imaging Eagle sensor and the Optech ALTM 3100EA sensor, respectively. Both sensors provide a spatial resolution of 1 m, with a pixel size of 600 × 166. The HSI data includes 63 spectral channels spanning from 402.89 nm to 989.09 nm, offering a spectral resolution between 0.42 μm and 0.99 μm. The LiDAR data are composed of a single raster. Figure 5 displays the image types for the HSI and LiDAR data and the ground truth land cover categories.

(3) Augsburg Dataset: The Augsburg dataset captures land cover in the German city of Augsburg by integrating data from hyperspectral imaging and LiDAR sources. Both data sources use 30 m spatial resolution. The HSI data contain 180 spectral bands with wavelengths ranging from 0.4 μm (UV and visible) to 2.5 μm (NIR). The LiDAR data consist of a single raster layer, both captured at a wavelength of 1.06 μm. As shown in Figure 6, it has a pixel size of 332 × 485 and consists of seven different categories.

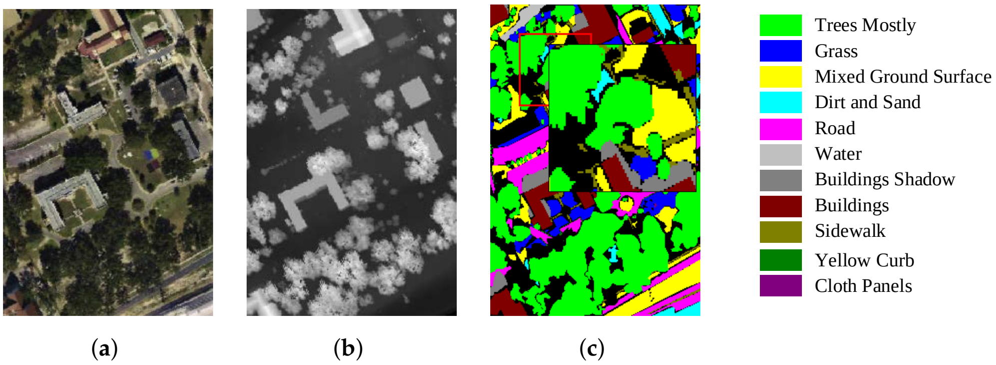


Figure 4. Visual representation of the MUUFL Gulfport dataset. (**a**) False-color map. (**b**) Grayscale image for the LiDAR. (**c**) Ground truth map of various land cover categories.

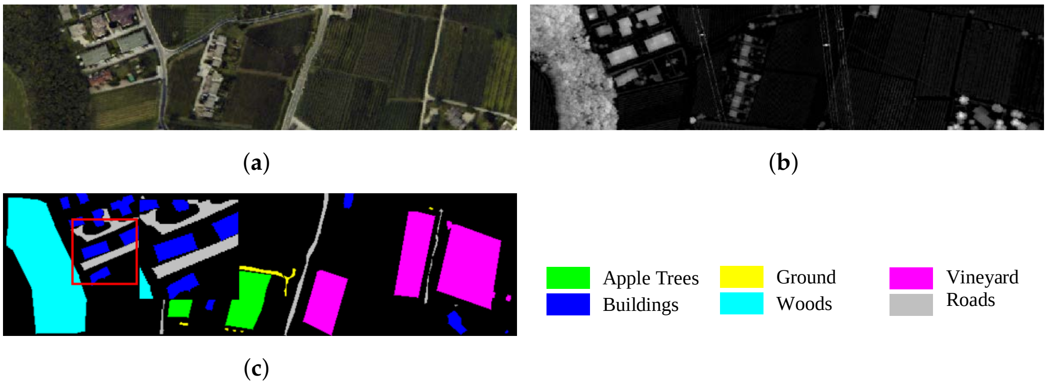


Figure 5. Visual representation of the Trento dataset. (**a**) False-color map. (**b**) Grayscale image for the LiDAR. (**c**) Ground truth map of various land cover categories.

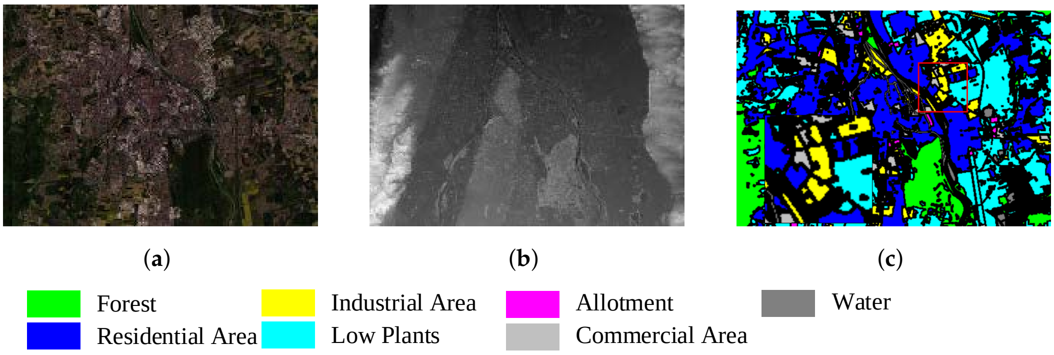


Figure 6. Visual representation of the Augsburg dataset. (**a**) False-color map. (**b**) Grayscale image for the LiDAR. (**c**) Ground truth map of various land cover categories.

Table 1. Training and test samples in MUUFL Gulfport, Trento, and Augsburg.

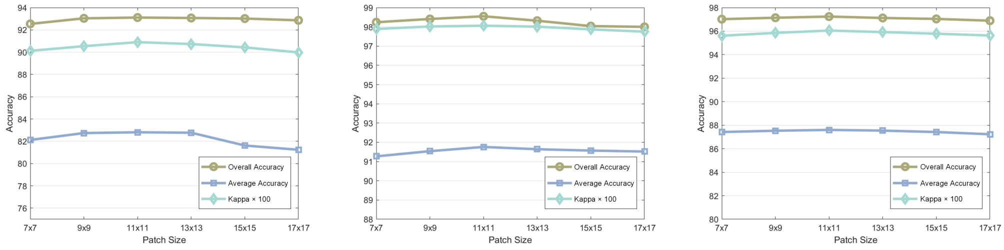
ID	MUUFL Gulfport			Trento			Augsburg		
	Land Cover Class	**Training**	**Test**	**Land Cover Class**	**Training**	**Test**	**Land Cover Class**	**Training**	**Test**
C01	Trees Mostly	1163	22,083	Apple Trees	21	4013	Forest	676	12,831
C02	Grass	214	4056	Buildings	15	2888	Residential Area	1517	28,812
C03	Mixed Ground Surface	345	6537	Ground	3	476	Industrial Area	193	3658
C04	Dirt and Sand	92	1734	Woods	46	9077	Low Plants	1343	25,514
C05	Road	335	6352	Vineyard	53	10,448	Allotment	29	546
C06	Water	24	442	Roads	16	3158	Commercial Area	83	1562
C07	Buildings Shadow	112	2121				Water	77	1453
C08	Buildings	312	5928						
C09	Sidewalk	70	1315						
C10	Yellow Curb	10	173						
C11	Cloth Panels	14	255						
	Total	2691	50,990	Total	154	30,060	Total	3918	74,376

3.2. Experimental Setting

(1) Evaluation Metrics: To assess the performance of our proposed AFA–Mamba, we utilized four commonly used evaluation metrics for a fair comparison: average accuracy (AA), overall accuracy (OA), Kappa coefficient (k), and per-class accuracy. We want the score of each metric to be as high as possible, which means the classification accuracy of our model is more accurate. To ensure fair comparisons, all experiments were conducted using separate training and test sets.

(2) Environment Configuration: We implemented our AFA–Mamba by PyTorch 2.2.0. Training was conducted using the Adam optimizer with the parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The model trained for 100 epochs, with a batch size of 64 and an initial learning rate of 1×10^{-3}. The training process utilized an NVIDIA Geforce RTX 4070ti 16 GB GPU. Regarding the traditional method among the comparison methods, we completed the experiment on the MATLAB platform.

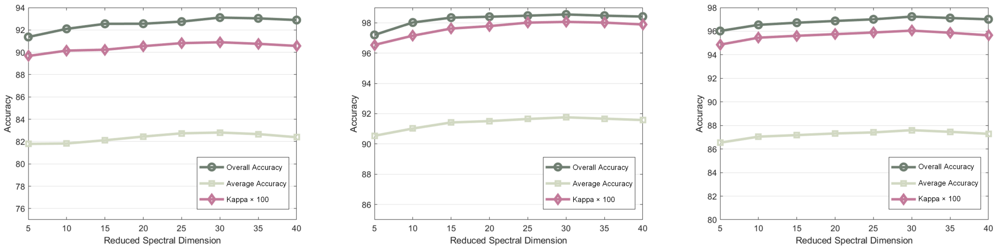
Parameter Setting Adjustment: (1) Patch Size: For joint HSI and LiDAR data classification methods, choosing an appropriate patch size means that the model can consider more neighborhood pixels with limited computational cost, which may better capture the spatial distribution, band changes, and elevation information characteristics of ground objects. In the dataset presented above, we fixed all parameter settings except patch size. In addition, in order to select the appropriate patch size, we evaluated the accuracy effect of the patch sizes in the set {7, 9, 11, 13, 15, 17} in sequence. As shown in Figure 7, an excessively large patch size will cause the method, especially the scalar fusion module, to have more complex inputs, resulting in a reduced network fitting effect and thus reduced accuracy. A patch size that is too small will result in insufficient data context information, thereby affecting global information retention and resulting in insufficient accuracy. It can be seen from this experiment that a patch size of 11 is the most appropriate.



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

Figure 7. Effect of patch size on the OA, AA, and Kappa coefficient. (**a**) MUUFL Gulfport. (**b**) Trento. (**c**) Augsburg. It can be seen that our method shows different sensitivities to the patch size on different datasets, thus better tuning the hyperparameters of the model to the optimum.

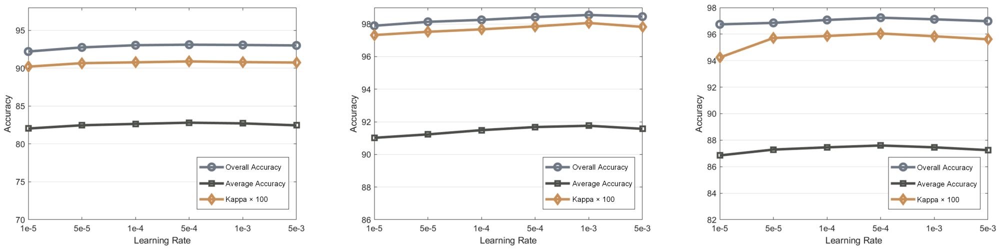
Although HSI data contain hundreds of continuous spectral bands and provide rich ground object information, due to their wide range of spectral imaging and high response characteristics, many redundant bands are often retained in the image, which can easily cause dimensional disaster. Therefore, we utilized PCA to retain the most important spectral data to achieve a balance between spectral expression ability and computational efficiency. As shown in Figure 8, too small a number of retained spectra will result in a large loss of spectral information, and low-resolution spatial data and elevation information cannot support sufficient classification accuracy. However, an excessively large number of retained spectra will introduce a certain amount of redundant noise, accompanied by high computational costs, resulting in a decrease in classification accuracy. We selected the number of retained spectral bands from the set {5, 10, 15, 20, 25, 30, 35, 40} for accuracy effect evaluation. When we retained the main 30 bands of HSI data, they could be effectively combined with LiDAR data to improve classification accuracy.



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

Figure 8. Effect of reducing spectral dimensionality on the OA, AA, and Kappa coefficient. (**a**) MUUFL Gulfport. (**b**) Trento. (**c**) Augsburg. It can be seen that our method shows different sensitivities to reductions on different datasets, thus better tuning the hyperparameters of the model to the optimum.

Learning rate: The learning rate controls the magnitude of the weight updates during model training, influencing how quickly or slowly the model learns. A higher learning rate can lead to oscillations during training, potentially hindering the model's ability to converge consistently. On the contrary, a lower learning rate will slow down the training process, leading to increased training time and higher computational cost. Therefore, in the hyperspectral image classification task, an appropriate learning rate helps to stabilize the fitting effect of the training process, thus affecting the final classification performance. We selected the learning rate from the set {1 $\times 10^{-5}$, 5 $\times$ 10^{-5}, 1 $\times$ 10^{-4}, 5 $\times$ 10^{-4}, 1 $\times$ 10^{-3}, 5 $\times$ 10^{-3}} for accuracy effect evaluation. As shown in Figure 9, it can be seen from this experiment that for the MUUFL Gulfport and Augsburg datasets, setting the learning rate to 5 $\times$ 10^{-4} and the Trento dataset to 1 $\times$ 10^{-3} can achieve better classification results.



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

Figure 9. Effect of learning rate on the OA, AA, and Kappa coefficient. (**a**) MUUFL Gulfport. (**b**) Trento. (**c**) Augsburg. It can be seen that our method shows different sensitivities to the learning rate on different datasets, thus better tuning the hyperparameters of the model to the optimum.

3.3. Ablation Study

To assess the impact of each component within AFA–Mamba, a series of experiments was conducted using the MUFFL dataset. The sample setup is shown in Table 1. The experiments included PCA, the spectral–spatial–elevation feature extraction module (SSE-FE), the HSI-Global–Local Mamba module (HSI-GL–Mamba), the LiDAR–Global–Local Mamba module (LiDAR-GL–Mamba), and the Adaptive Alignment and Fusion Module (A^2F) analysis. Table 2 reports the classification performance resulting from combining different modules.

No PCA was used in Case 1, and the raw spectra of the data were directly fed into the model. Due to the massive amount of invalid band information brought in, the accuracy of the model degraded severely. The OA, AA, and K accuracies all degraded by more than 3%. This illustrates the importance of filtering bands for de-redundancy in hyperspectral tasks.

Case 2 replaced the SSE-FE module with a regular 2D-CNN, which coarsely extracted spectral–spatial–elevation information. The OA, AA, and K accuracies were somewhat affected, decreasing by 1.86%, 1.57%, and 1.55. This illustrates the improvement of our designed feature enhancement extraction module with respect to a single CNN.

Table 2. Ablation experimental results. The results show that each module of our method works and is consistent with the motivation for the experiment.

Cases	**Component**					**Indicators**		
	PCA	**SSE-FE**	**HSI-GL-Mamba**	**LiDAR-GL-Mamba**	**A^{2}F**	**OA (%)**	**AA (%)**	**$k * 100$**
1	×	✓	✓	✓	✓	89.87	79.41	87.62
2	✓	CNN	✓	✓	✓	91.25	81.24	89.35
3	✓	✓	×	✓	✓	90.49	80.12	88.14
4	✓	✓	×	✓	✓	90.48	80.54	87.02
5	✓	✓	✓	×	✓	90.05	81.28	88.53
6	✓	✓	✓	✓	×	91.23	80.36	88.15
7	✓	✓	Original Mamba	✓	✓	91.52	81.62	89.32
8	✓	✓	✓	Original Mamba	✓	90.73	80.54	88.95
9	✓	✓	✓	✓	✓	**93.11**	**82.81**	**90.90**

Case 3 removed the SSE-FE module and directly used the most basic patch for the features of two different types of data. We can observe that the model extraction ability dropped significantly, which is reflected in three important indicators, among which the OA, AA, and K accuracy all dropped by more than 2%. This illustrates the importance of the SSE-FE module in improving classification accuracy.

Case 4 and Case 5 eliminated the HSI-GL–Mamba and LiDAR-GL–Mamba modules, respectively, which are of great significance in our feature retention. Both modules can differentiate between global and local information based on the feature map. Here, there is only one serial convolutional branch for replacement, which does not allow for the efficient extraction of more comprehensive features from HSI and LiDAR data. Therefore, the accuracy of the three classifications dropped significantly. This demonstrates that the HSI-GL–Mamba and LiDAR-GL–Mamba modules explore deep spectral–spatial–elevation semantic information, facilitating the effective integration and interaction of features across multiple branches.

Case 6 removed the A^{2}F module and adds the spectral–spatial and elevation features directly into the classifier. It is obvious that although the Mamba structure is able to obtain sufficiently rich features, due to the inherent differences and spatial offsets in the two features, direct addition cannot effectively fuse the discriminative features of these features, thus leading to a decrease in classification accuracy. In Case 7, the A^{2}F module we proposed was able to eliminate this challenge, thus achieving the optimal classification accuracy and greatly improving the three classification indicators.

Cases 7 and 8 replace the GL–Mamba of the upper-branch HSI and the GL–Mamba of the lower-branch LiDAR with the original Mamba. It is evident that while the original Mamba structure can capture global contextual information, it lacks the local bias necessary to refine spatial details. As a result, the classification accuracy suffers, as the model is unable to fully exploit the fine-grained spatial and spectral variations present in both data sources.

3.4. Classification Result and Analysis

To demonstrate the superiority of our proposed AFA–Mamba compared to other advanced methods, we carefully selected some representative and excellent classification methods and divided them into two categories: the HSI-based classification methods, including RF [20], SVM [21], 2D-CNN [49], HybridSN [29], GAHT [35], and MiM [50], and the joint HSI and LiDAR data fusion classification methods, including CoupledCNN [40], CALC [42], HCTnet [43], M2FNet [44], and HLMamba [51]. For a fair comparison, the default parameters of all methods followed the corresponding references, and the training set partitioning and other parameters were consistent with the configuration described above. We repeated the experiment ten times to obtain its mean and variance.

(1) Quantitative Results and Analysis:

Tables 3–5 report the quantitative comparison results of our AFA–Mamba against other advanced methods across the three datasets. We conducted each experiment 10 times and calculated the average and standard deviation of the results to ensure the reliability and fairness of the comparison. In the table, the best results are highlighted in bold red. Our method consistently achieved the highest classification performance across all datasets for the OA, AA, and Kappa metrics.

For example, Table 3 provides the detailed results of each comparison method alongside our proposed AFA–Mamba using the MUUFL Gulfport dataset. Our method achieves the highest performance in all aspects. In contrast to traditional approaches like RF and SVM, deep learning methods can learn a wider range of features to improve classification accuracy. The 2D-CNN, HyBridSN, GAHT, and MiM methods based on HSI data classification extract effective features in HSI data, but their OA results are 1.32%, 2.84%, 4.00%, and 1.48% lower than our AFA–Mamba, respectively. The main reason for this is that their limited receptive fields ignore global features. Moreover, our method outperforms joint HSI and LiDAR data classification approaches in key metrics such as OA, AA, and k. It also demonstrates competitive performance in the average accuracy of each class. Specifically, our AFA–Mamba surpasses CoupledCNN, CALC, HCTnet, M2FNet, and HLMamba in AA by 2.53%, 4.79%, 3.58%, 2.57%, and 1.1%, respectively. This is attributed to the multiple receptive fields of the local–global branch of our method and the efficient spectral–elevation feature correction fusion strategy. Tables 4 and 5 report the results of all compared methods on the Trento and Augsburg datasets. We can draw similar conclusions as above from the two tables. Obviously, our proposed method still maintains an advanced level in all aspects.

The Trento dataset has a small number of categories and an uneven distribution of sample numbers. Therefore, the OA and Kappa of each method achieved high accuracy, while AA had greater room for improvement. It can be observed from Table 4 that the joint classification methods of HSI and LiDAR data achieved good results, especially HCTnet and M2FNet. The dual Transformer encoder branch used by HCTnet better combines the characteristics of the two types of remote sensing data. However, due to its more complex fusion mode, it is prone to under-fitting in the Trento dataset with a small number of samples. As for our proposed AFA–Mamba, the MFR module can better retain the detailed information of the data even in scenarios where the number of samples is insufficient. Therefore, our OA and Kappa are 0.56% and 0.75% higher than HCTnet, respectively. In addition, M2FNet benefits from its multi-scale feature extraction and retains more spectral–spatial–elevation information, and the Global–Local Mamba modules we proposed fully enhance the awareness of global–local information, so it is ultimately better than M2FNet and achieved OA, AA, and Kappa values that are 0.77%, 0.19%, and 1.03% higher, respectively.

Table 3. Classification accuracy of comparison experiment in MUUFL Gulfport using various methods. Nos. 1 to 11 are the classification accuracies for each landform category.

No.	Only HSI Input						HSI and LiDAR Input					
	RF [20]	**SVM [21]**	**2D-CNN [49]**	**HybridSN [29]**	**GAHT [35]**	**MiM [50]**	**CoupledCNN [40]**	**CALC [42]**	**HCTnet [43]**	**M2FNet [44]**	**HLMamba [51]**	**Ours**
1	96.54 ± 1.13	**97.64 ± 1.07**	97.15 ± 0.32	97.07 ± 0.3	96.21 ± 0.74	95.75 ± 1.22	97.23 ± 0.52	97.46 ± 1.03	97.13 ± 0.67	96.95 ± 0.41	97.12 ± 0.55	96.94 ± 0.31
2	79.11 ± 4.30	87.39 ± 3.04	84.61 ± 1.28	81.91 ± 7.99	84.15 ± 5.42	85.96 ± 1.32	82.21 ± 4.00	**88.41 ± 3.19**	86.31 ± 1.60	86.12 ± 2.54	86.40 ± 3.12	87.43 ± 4.3
3	83.76 ± 3.14	86.14 ± 4.83	86.63 ± 3.36	83.61 ± 4.03	85.32 ± 3.87	88.23 ± 1.2	87.64 ± 2.73	84.82 ± 5.44	89.40 ± 1.51	88.86 ± 1.81	88.39 ± 1.72	**91.68 ± 1.16**
4	89.36 ± 2.72	93.64 ± 2.01	**94.92 ± 0.87**	92.39 ± 6.93	89.52 ± 4.55	91.65 ± 3.09	92.90 ± 2.68	89.39 ± 3.48	90.40 ± 1.52	89.85 ± 1.64	91.59 ± 1.62	91.42 ± 2.76
5	85.99 ± 4.11	90.39 ± 2.39	90.78 ± 0.97	90.74 ± 1.03	86.95 ± 3.36	91.16 ± 0.67	93.93 ± 1.45	93.03 ± 1.98	93.62 ± 1.03	94.01 ± 0.70	92.02 ± 0.67	**94.45 ± 1.06**
6	89.37 ± 4.77	95.42 ± 2.86	95.44 ± 3.61	84.47 ± 7.03	94.40 ± 3.03	96.44 ± 0.31	94.57 ± 2.87	90.16 ± 3.86	83.62 ± 5.34	93.28 ± 1.75	**96.24 ± 1.03**	91.9 ± 3.24
7	80.31 ± 3.69	83.18 ± 2.30	81.88 ± 2.17	78.41 ± 4.29	75.55 ± 7.75	81.09 ± 2.15	**88.77 ± 1.06**	83.61 ± 4.23	86.04 ± 1.99	84.89 ± 2.76	85.82 ± 0.93	86.36 ± 1.99
8	95.70 ± 2.09	96.86 ± 1.18	96.39 ± 0.92	95.82 ± 1.45	93.81 ± 2.86	95.8 ± 0.51	96.21 ± 1.43	96.60 ± 0.95	96.68 ± 0.47	**97.14 ± 0.53**	95.72 ± 0.37	96.58 ± 0.79
9	28.39 ± 4.11	40.75 ± 5.58	54.18 ± 5.44	48.30 ± 7.19	36.06 ± 7.04	48.19 ± 5.47	45.31 ± 6.11	50.78 ± 9.34	52.11 ± 4.16	**57.64 ± 3.08**	48.55 ± 2.82	53.44 ± 4.87
10	5.53 ± 2.84	13.43 ± 7.18	23.60 ± 5.94	26.63 ± 5.65	11.22 ± 5.41	20.56 ± 1.91	16.18 ± 3.78	11.56 ± 8.38	20.00 ± 2.02	24.57 ± 4.70	26.25 ± 4.58	**27.11 ± 5.54**
11	70.00 ± 8.30	80.47 ± 7.53	86.71 ± 4.74	92.59 ± 4.18	61.84 ± 7.68	64.62 ± 6.62	88.08 ± 6.92	72.39 ± 11.3	76.20 ± 2.84	69.37 ± 5.69	61.31 ± 2.3	**93.57 ± 1.77**
OA(%)	88.92 ± 0.79	91.76 ± 0.51	91.79 ± 0.45	90.27 ± 1.08	89.11 ± 0.88	91.63 ± 0.22	92.09 ± 0.61	91.94 ± 0.65	92.45 ± 0.28	92.53 ± 0.19	92.01 ± 0.27	**93.11 ± 0.39**
AA(%)	73.09 ± 1.02	78.66 ± 1.33	81.12 ± 0.83	79.19 ± 1.96	74.05 ± 1.29	78.25 ± 0.89	80.28 ± 1.91	78.02 ± 1.87	79.23 ± 0.85	80.24 ± 0.80	79.04 ± 0.66	**82.81 ± 1.31**
$k \times 100$	85.27 ± 1.04	89.05 ± 0.70	89.10 ± 0.61	87.13 ± 1.43	85.59 ± 1.11	88.91 ± 0.29	89.52 ± 0.81	89.30 ± 0.86	90.00 ± 0.36	90.12 ± 0.25	89.42 ± 0.35	**90.90 ± 0.52**

Table 4. Classification accuracy of comparison experiment in Trento using various methods. Nos. 1 to 6 are the classification accuracies for each landform category.

No.	Only HSI Input						HSI and LiDAR Input					
	RF [20]	**SVM [21]**	**2D-CNN [49]**	**HybridSN [29]**	**GAHT [35]**	**MiM [50]**	**CoupledCNN [40]**	**CALC [42]**	**HCTnet [43]**	**M2FNet [44]**	**HLMamba [51]**	**Ours**
1	94.62 ± 4.91	96.28 ± 4.61	97.41 ± 1.87	96.33 ± 5.09	98.69 ± 0.79	90.96 ± 4.84	97.91 ± 3.45	98.56 ± 1.30	99.30 ± 0.81	99.08 ± 0.82	98.95 ± 0.78	**99.43 ± 0.40**
2	82.93 ± 12.81	92.68 ± 3.33	85.49 ± 3.74	76.90 ± 7.43	79.42 ± 6.69	48.14 ± 11.14	94.17 ± 1.46	94.49 ± 2.80	**98.10 ± 2.18**	90.54 ± 5.53	84.29 ± 4.99	94.29 ± 2.20
3	6.02 ± 10.70	9.08 ± 5.07	39.56 ± 13.64	45.35 ± 16.13	31.97 ± 7.16	54.38 ± 9.24	18.30 ± 8.51	51.41 ± 23.48	63.57 ± 18.91	**65.8 ± 12.73**	34.3 ± 10.73	58.42 ± 15.5
4	99.90 ± 0.10	99.88 ± 0.18	99.85 ± 0.20	99.86 ± 0.26	99.93 ± 0.05	99.82 ± 0.19	98.77 ± 0.80	99.99 ± 0.02	**100.00 ± 0.00**	99.97 ± 0.04	99.96 ± 0.05	99.99 ± 0.02
5	99.69 ± 0.63	**100.00 ± 0.00**	99.81 ± 0.19	94.96 ± 3.27	99.99 ± 0.02	95.35 ± 2.5	99.91 ± 0.10	99.96 ± 0.08	99.76 ± 0.32	99.96 ± 0.07	99.57 ± 0.22	99.99 ± 0.01
6	78.31 ± 5.50	81.83 ± 4.12	83.76 ± 2.30	71.88 ± 4.81	92.73 ± 3.19	93.28 ± 3.45	98.20 ± 0.52	88.22 ± 3.64	89.75 ± 2.98	94.07 ± 2.25	85.66 ± 4.98	**98.43 ± 0.44**
OA(%)	93.74 ± 1.03	95.41 ± 0.75	95.48 ± 0.65	91.67 ± 1.71	95.98 ± 0.50	90.71 ± 1.25	97.27 ± 0.53	97.25 ± 0.56	97.99 ± 0.27	97.78 ± 0.61	95.64 ± 0.58	**98.55 ± 0.35**
AA(%)	76.91 ± 2.78	79.96 ± 1.11	84.31 ± 2.55	80.88 ± 3.25	83.79 ± 1.64	80.32 ± 2.70	84.54 ± 1.21	88.77 ± 3.76	91.75 ± 3.07	91.57 ± 2.29	83.79 ± 2.23	**91.76 ± 2.68**
$k \times 100$	91.57 ± 1.41	93.83 ± 1.03	93.94 ± 0.87	88.86 ± 2.31	94.61 ± 0.68	87.61 ± 1.66	96.35 ± 0.72	96.32 ± 0.76	97.31 ± 0.36	97.03 ± 0.82	94.16 ± 0.78	**98.06 ± 0.47**

Table 5. Classification accuracy of comparison experiment in Augsburg using various methods. Nos. 1 to 7 are the classification accuracies for each landform category.

No.	Only HSI Input						HSI and LiDAR Input					
	RF [20]	**SVM** [21]	**2D-CNN** [49]	**HybridSN** [29]	**GAHT** [35]	**MiM** [50]	**CoupledCNN** [40]	**CALC** [42]	**HCTnet** [43]	**M2FNet** [44]	**HLMamba** [51]	**Ours**
1	98.58 ± 0.41	98.89 ± 0.32	99.15 ± 0.18	98.88 ± 0.67	97.76 ± 0.52	99.24 ± 0.2	**99.40 ± 0.21**	98.76 ± 0.30	99.09 ± 0.17	98.23 ± 0.72	99.1 ± 0.22	99.16 ± 0.21
2	96.52 ± 3.12	97.86 ± 1.51	97.75 ± 0.36	97.29 ± 0.76	98.13 ± 0.62	98.18 ± 0.25	99.17 ± 0.20	**99.34 ± 0.12**	98.80 ± 0.26	98.37 ± 0.32	98.22 ± 0.52	98.83 ± 0.36
3	77.82 ± 5.20	82.22 ± 3.44	80.18 ± 2.86	76.87 ± 4.22	85.70 ± 3.21	80.8 ± 1.42	82.25 ± 4.99	91.69 ± 3.16	91.02 ± 2.63	**91.93 ± 2.04**	83.15 ± 4.0	89.38 ± 2.6
4	98.77 ± 0.21	99.02 ± 0.35	**99.16 ± 0.10**	98.94 ± 0.19	97.57 ± 0.48	99.05 ± 0.07	98.98 ± 0.21	99.08 ± 0.17	98.60 ± 0.26	98.43 ± 0.56	98.8 ± 0.88	98.92 ± 0.26
5	41.93 ± 14.77	61.47 ± 9.84	69.41 ± 5.14	67.53 ± 4.44	65.46 ± 10.06	69.87 ± 5.52	78.99 ± 7.96	**85.35 ± 3.15**	81.92 ± 5.59	80.95 ± 5.01	78.17 ± 3.83	83.86 ± 2.96
6	22.25 ± 5.65	32.39 ± 4.92	48.68 ± 5.70	40.33 ± 4.76	67.74 ± 5.92	51.02 ± 5.47	55.43 ± 6.88	58.67 ± 9.48	68.07 ± 4.11	71.54 ± 3.23	59.42 ± 4.46	**76.2 ± 4.67**
7	57.71 ± 3.36	60.30 ± 2.27	65.87 ± 1.24	64.21 ± 2.38	60.10 ± 4.20	67.12 ± 3.30	64.84 ± 2.56	63.78 ± 1.69	68.31 ± 2.62	68.29 ± 3.22	**68.62 ± 1.51**	66.81 ± 3.53
OA(%)	94.01 ± 1.07	95.29 ± 0.63	95.75 ± 0.11	95.07 ± 0.45	95.64 ± 0.30	96.00 ± 0.13	96.57 ± 0.28	97.12 ± 0.15	97.03 ± 0.20	96.77 ± 0.31	96.29 ± 0.58	**97.24 ± 0.14**
AA(%)	70.51 ± 2.25	76.02 ± 1.83	80.03 ± 1.18	77.72 ± 0.95	81.78 ± 2.02	80.76 ± 0.61	82.72 ± 0.85	85.24 ± 1.29	86.54 ± 1.26	86.82 ± 1.07	83.64 ± 0.64	**87.60 ± 0.93**
$k \times 100$	91.40 ± 1.46	93.24 ± 0.87	93.91 ± 0.16	92.93 ± 0.63	93.75 ± 0.43	94.26 ± 0.18	95.08 ± 0.40	95.87 ± 0.23	95.75 ± 0.29	95.37 ± 0.45	94.69 ± 0.81	**96.05 ± 0.20**

The Augsburg dataset has a high spatial resolution and complex ground object information, which places higher requirements on the model's local–global information balance capability. In addition, due to the addition of LiDAR data elevation information, the joint classification method has additional advantages over the single-source data classification methods. Specifically, as shown in Table 5, CALC uses dual adversarial networks to conduct adversarial training on ground object space and achieves high OA accuracy. This is due to the fact that the adversarial strategy is conducive to the effective combination of space and elevation information. However, our proposed SFCF module achieves effective correction of the fused data by introducing additional offsets, which enhances the coupling degree of spatial–elevation information. Therefore, it is still 0.12%, 1.36%, and 0.18% higher than the CALC method in terms of OA, AA, and Kappa, respectively. Similarly, compared with other methods, the efficient and stable fusion caused by the offset correction of our method can also be proven to be more robust. In summary, by comparing with current classic and advanced methods in three different types of datasets, it is proved that our AFA–Mamba has efficient joint classification performance.

(2) Visual Evaluation and Analysis: Figures 10–12 show that the visualization results of various methods can be used for qualitative comparison. From the results, we can significantly observe the classification differences between different methods. It is worth mentioning that our AFA–Mamba can generate more accurate noise-free feature classification maps.

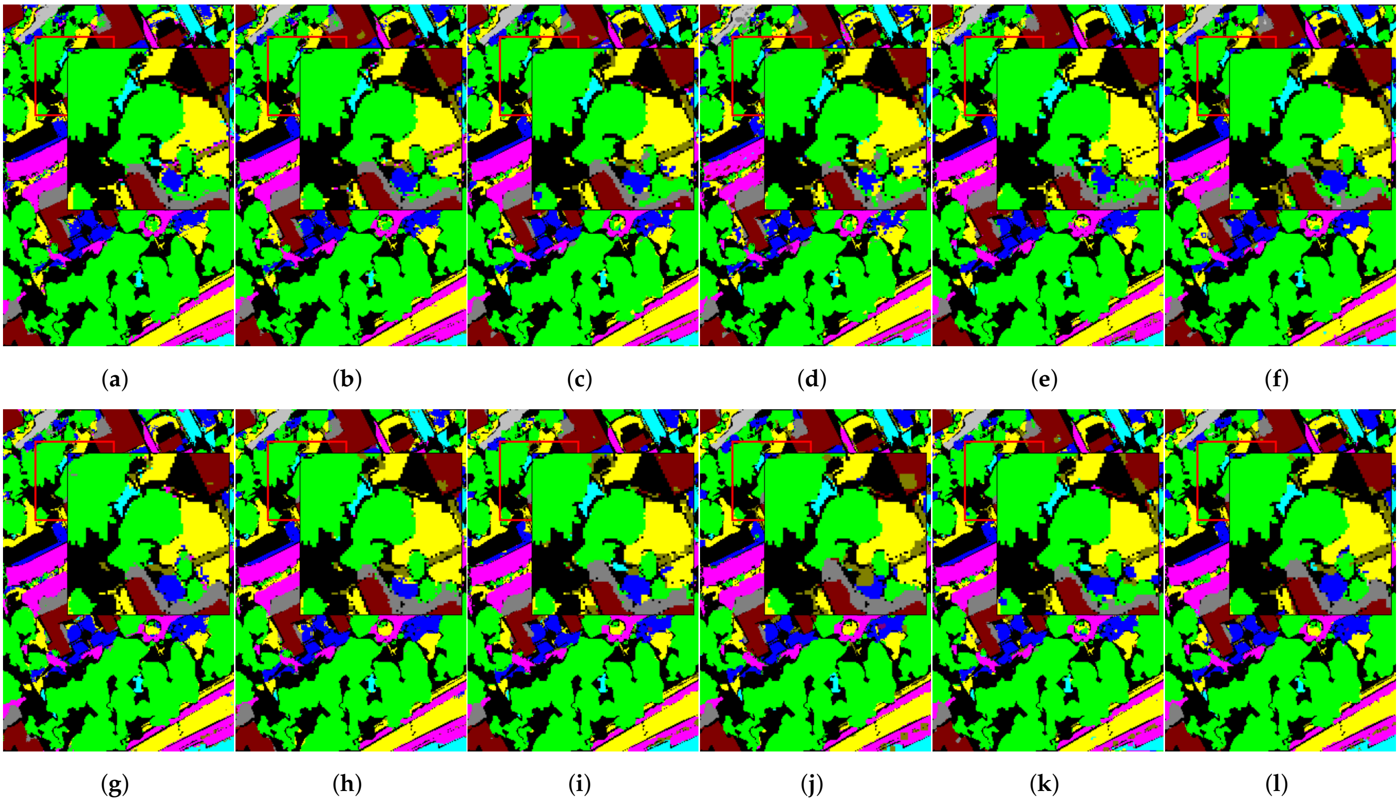
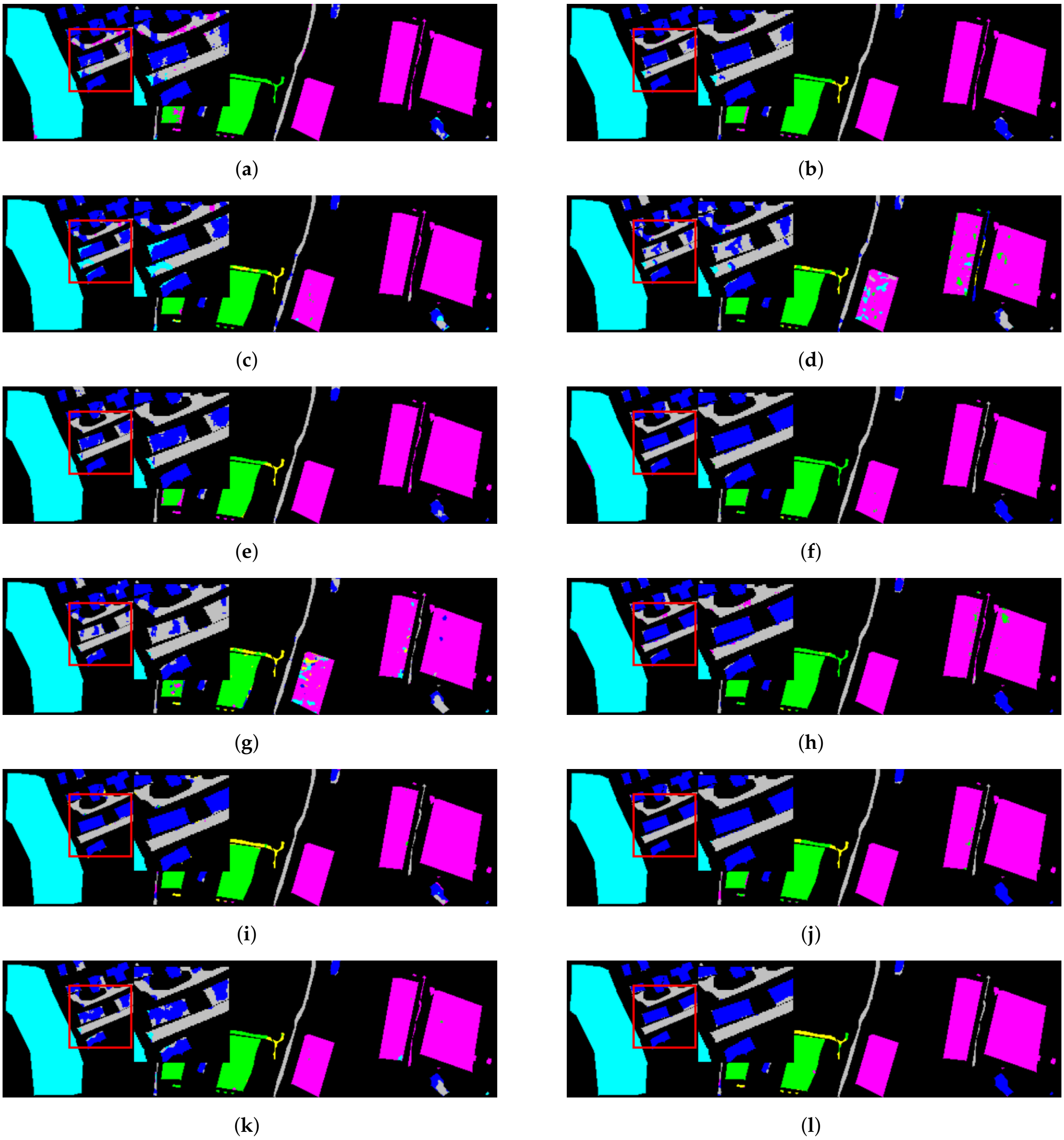


Figure 10. Maps depicting the classification of MUUFL Gulfport using various methods. (**a**) RF. (**b**) SVM. (**c**) 2D-CNN. (**d**) HybridSN. (**e**) GAHT. (**f**) MiM. (**g**) CoupledCNN. (**h**) CALC. (**i**) HCTnet. (**j**) M2FNet. (**k**) HLMamba. (**l**) AFA–Mamba.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**) (**h**) (**i**) (**j**) (**k**) (**l**)

Figure 11. Maps depicting the classification of Trento using various methods. (**a**) RF. (**b**) SVM. (**c**) 2D-CNN. (**d**) HybridSN. (**e**) GAHT. (**f**) MiM. (**g**) CoupledCNN. (**h**) CALC. (**i**) HCTnet. (**j**) M2FNet. (**k**) HLMamba. (**l**) AFA–Mamba.

Specifically, the classification method that only uses HSI data has unclear boundaries and introduces more noise. The multi-source data joint classification method significantly reduces this phenomenon, but performs poorly in some ground areas. On the other hand, the classification results obtained by our method have clear boundaries and high classification accuracy. For example, the visualization results of the MUUFL Gulfport dataset are presented in Figure 10, and it can be easily seen that most methods have blur and a lot of noise when distinguishing small and dense areas of multiple categories, while the classification map of our AFA–Mamba is closer to the real ground truth.

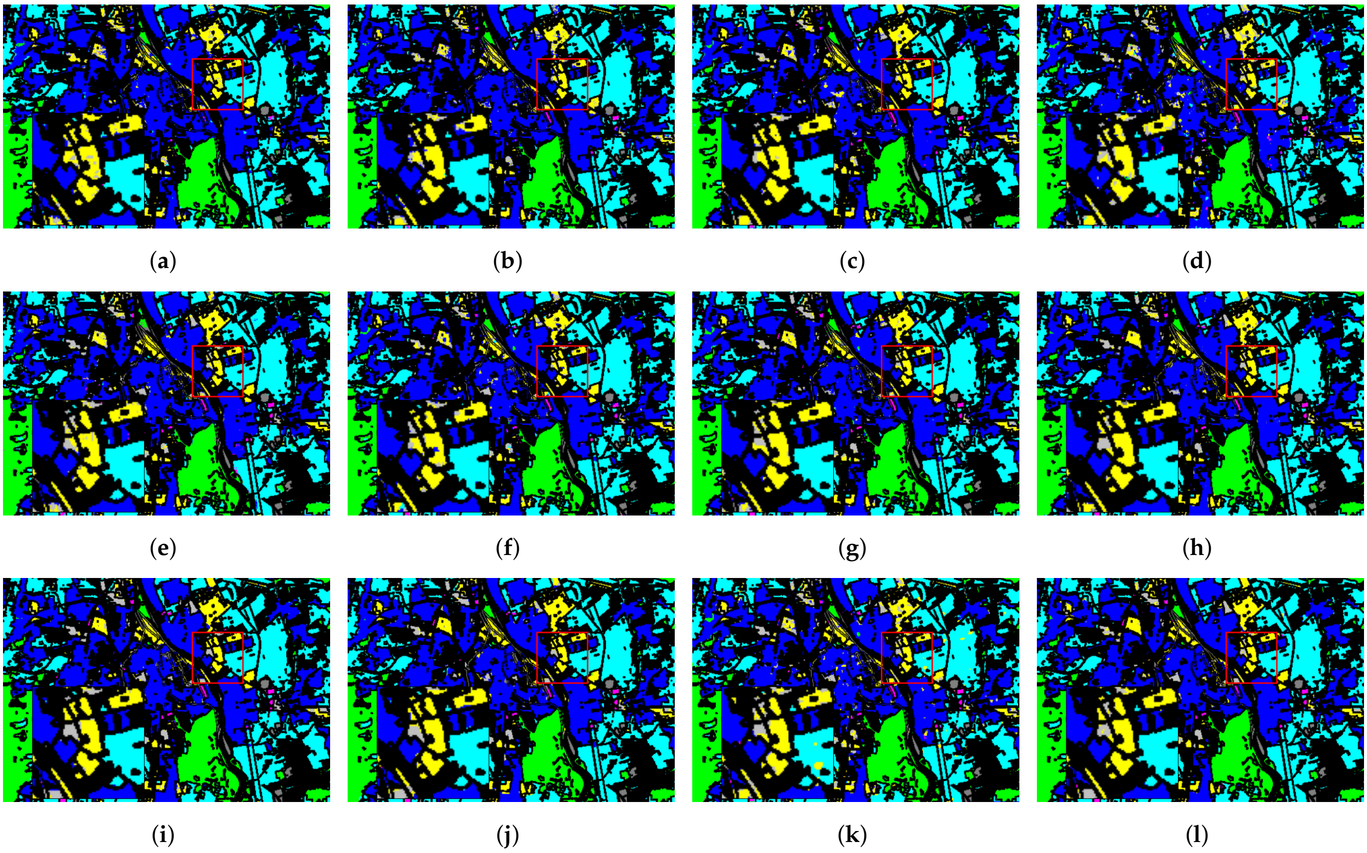


Figure 12. Maps depicting the classification of Augsburg using various methods. (**a**) RF. (**b**) SVM. (**c**) 2D-CNN. (**d**) HybridSN. (**e**) GAHT. (**f**) MiM. (**g**) CoupledCNN. (**h**) CALC. (**i**) HCTnet. (**j**) M2FNet. (**k**) HLMamba. (**l**) AFA–Mamba.

Figure 11 illustrates the visualization results of the comparative method in the Terno dataset. This dataset has larger ground objects and fewer categories than the MUUFL Gulfport dataset, so it is easier to classify. We can clearly see that most comparison methods still have the problem of introducing a large amount of noise, while our method still preserves high-precision classification effects. In addition, the classification results of the Augsburg dataset with higher resolution are depicted in Figure 12; our AFA–Mamba still maintains the optimal classification performance in denser scenes with more categories.

4. Conclusions

In this paper, we propose a novel and efficient Adaptive Feature Alignment Network with a Global–Local Mamba (AFA–Mamba) to achieve the efficient fusion of spectral–spatial and elevation features, greatly improving the accuracy of remote sensing classification tasks. Specifically, we propose a novel SSE feature extraction module to explore deep spectral–spatial–elevation semantic information through multi-branch feature extraction. In addition, Global–Local Mamba modules are proposed to enhance ground objects that are sensitive to spectral and elevation information. In order to eliminate the representation differences and spatial misalignment of multi-source features, we propose the SSE Adaptive Alignment and Fusion (A^2F) module to effectively learn the discriminative features of heterogeneous data and achieve the adaptive calibration of spatial differences. Extensive experiments demonstrate the advancement and robustness of our method.

Author Contributions: Conceptualization, S.L. and S.H.; methodology, S.L. and S.H.; software, S.L.; formal analysis, S.L.; investigation, S.H.; data curation, S.H.; writing—original draft preparation, S.L.; writing—review and editing, S.H.; visualization, S.L. and S.H.; supervision, S.L. and S.H.; project administration, S.L.; funding acquisition, S.L. and S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Shandong Province (Grant number: ZR202103050458); and the start-up foundation for doctoral research of Zaozhuang University (Grant number: 1020714).

Data Availability Statement: All relevant data are within the paper.

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

References

1. Kahraman, S.; Bacher, R. A comprehensive review of hyperspectral data fusion with lidar and sar data. *Annu. Rev. Control* **2021**, *51*, 236–253. [CrossRef]
2. Zhou, Z.; Zheng, C.; Liu, X.; Tian, Y.; Chen, X.; Chen, X.; Dong, Z. A dynamic effective class balanced approach for remote sensing imagery semantic segmentation of imbalanced data. *Remote Sens.* **2023**, *15*, 1768. [CrossRef]
3. Wang, J.; Hu, J.; Liu, Y.; Hua, Z.; Hao, S.; Yao, Y. El-nas: Efficient lightweight attention cross-domain architecture search for hyperspectral image classification. *Remote Sens.* **2023**, *15*, 4688. [CrossRef]
4. Su, Z.; Wan, G.; Zhang, W.; Guo, N.; Wu, Y.; Liu, J.; Cong, D.; Jia, Y.; Wei, Z. An Integrated Detection and Multi-Object Tracking Pipeline for Satellite Video Analysis of Maritime and Aerial Objects. *Remote Sens.* **2024**, *16*, 724. [CrossRef]
5. Zhang, G.; Fang, W.; Zheng, Y.; Wang, R. SDBAD-Net: A spatial dual-branch attention dehazing network based on meta-former paradigm. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 60–70. [CrossRef]
6. Fang, W.; Zhang, G.; Zheng, Y.; Chen, Y. Multi-Task Learning for UAV Aerial Object Detection in Foggy Weather Condition. *Remote Sens.* **2023**, *15*, 4617. [CrossRef]
7. Kuras, A.; Brell, M.; Rizzi, J.; Burud, I. Hyperspectral and lidar data applied to the urban land cover machine learning and neural-network-based classification: A review. *Remote Sens.* **2021**, *13*, 3393. [CrossRef]
8. Audebert, N.; Le Saux, B.; Lefèvre, S. Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 159–173. [CrossRef]
9. Li, N.; Ho, C.P.; Wang, I.T.; Pitchappa, P.; Fu, Y.H.; Zhu, Y.; Lee, L.Y.T. Spectral imaging and spectral LIDAR systems: Moving toward compact nanophotonics-based sensing. *Nanophotonics* **2021**, *10*, 1437–1467. [CrossRef]
10. Khodadadzadeh, M.; Li, J.; Prasad, S.; Plaza, A. Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 2971–2983. [CrossRef]
11. Ghamisi, P.; Benediktsson, J.A.; Phinn, S. Land-cover classification using both hyperspectral and LiDAR data. *Int. J. Image Data Fusion* **2015**, *6*, 189–215. [CrossRef]
12. Murphy, R.J.; Taylor, Z.; Schneider, S.; Nieto, J. Mapping clay minerals in an open-pit mine using hyperspectral and LiDAR data. *Eur. J. Remote Sens.* **2015**, *48*, 511–526. [CrossRef]
13. Voss, M.; Sugumaran, R. Seasonal effect on tree species classification in an urban environment using hyperspectral data, LiDAR, and an object-oriented approach. *Sensors* **2008**, *8*, 3020–3036. [CrossRef]
14. Liu, L.; Coops, N.C.; Aven, N.W.; Pang, Y. Mapping urban tree species using integrated airborne hyperspectral and LiDAR remote sensing data. *Remote Sens. Environ.* **2017**, *200*, 170–182. [CrossRef]
15. Maxwell, A.E.; Warner, T.A.; Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* **2018**, *39*, 2784–2817. [CrossRef]
16. Gómez-Chova, L.; Tuia, D.; Moser, G.; Camps-Valls, G. Multimodal classification of remote sensing images: A review and future directions. *Proc. IEEE* **2015**, *103*, 1560–1584. [CrossRef]
17. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep learning for hyperspectral image classification: An overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [CrossRef]
18. Kumar, B.; Dikshit, O.; Gupta, A.; Singh, M.K. Feature extraction for hyperspectral image classification: A review. *Int. J. Remote Sens.* **2020**, *41*, 6248–6287. [CrossRef]
19. Ghamisi, P.; Höfle, B.; Zhu, X.X. Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *10*, 3011–3024. [CrossRef]
20. 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]
21. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
22. Pal, M.; Mather, P.M. Support vector machines for classification in remote sensing. *Int. J. Remote Sens.* **2005**, *26*, 1007–1011. [CrossRef]
23. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
24. Huang, F.; Lu, J.; Tao, J.; Li, L.; Tan, X.; Liu, P. Research on optimization methods of ELM classification algorithm for hyperspectral remote sensing images. *IEEE Access* **2019**, *7*, 108070–108089. [CrossRef]
25. Hang, R.; Liu, Q.; Song, H.; Sun, Y. Matrix-based discriminant subspace ensemble for hyperspectral image spatial–spectral feature fusion. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 783–794. [CrossRef]

26. Wang, Q.; Gu, Y.; Tuia, D. Discriminative multiple kernel learning for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3912–3927. [CrossRef]
27. Tejasree, G.; Agilandeeswari, L. An extensive review of hyperspectral image classification and prediction: Techniques and challenges. *Multimed. Tools Appl.* **2024**, *83*, 80941–81038. [CrossRef]
28. Ahmad, M.; Shabbir, S.; Roy, S.K.; Hong, D.; Wu, X.; Yao, J.; Khan, A.M.; Mazzara, M.; Distefano, S.; Chanussot, J. Hyperspectral image classification—Traditional to deep models: A survey for future prospects. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *15*, 968–999. [CrossRef]
29. 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]
30. Paoletti, M.E.; Haut, J.M.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.J.; Pla, F. Deep pyramidal residual networks for spectral–spatial hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 740–754. [CrossRef]
31. Paoletti, M.E.; Haut, J.M.; Plaza, J.; Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 279–317. [CrossRef]
32. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]
33. Zou, J.; He, W.; Zhang, H. Lessformer: Local-enhanced spectral-spatial transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [CrossRef]
34. Yuan, D.; Yu, D.; Qian, Y.; Xu, Y.; Liu, Y. S2Former: Parallel Spectral–Spatial Transformer for Hyperspectral Image Classification. *Electronics* **2023**, *12*, 3937. [CrossRef]
35. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
36. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
37. 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*, 1–16. [CrossRef]
38. Ma, X.; Zhang, X.; Pun, M.O. RS^{3}Mamba: Visual State Space Model for Remote Sensing Image Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*. [CrossRef]
39. Zhao, S.; Chen, H.; Zhang, X.; Xiao, P.; Bai, L.; Ouyang, W. Rs-mamba for large remote sensing image dense prediction. *arXiv* **2024**, arXiv:2404.02668. [CrossRef]
40. Hang, R.; Li, Z.; Ghamisi, P.; Hong, D.; Xia, G.; Liu, Q. Classification of hyperspectral and LiDAR data using coupled CNNs. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4939–4950. [CrossRef]
41. Zhang, M.; Li, W.; Tao, R.; Li, H.; Du, Q. Information fusion for classification of hyperspectral and LiDAR data using IP-CNN. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]
42. Lu, T.; Ding, K.; Fu, W.; Li, S.; Guo, A. Coupled adversarial learning for fusion classification of hyperspectral and LiDAR data. *Inf. Fusion* **2023**, *93*, 118–131. [CrossRef]
43. Zhao, G.; Ye, Q.; Sun, L.; Wu, Z.; Pan, C.; Jeon, B. Joint classification of hyperspectral and LiDAR data using a hierarchical CNN and transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 1–16. [CrossRef]
44. Sun, L.; Wang, X.; Zheng, Y.; Wu, Z.; Fu, L. Multiscale 3-D–2-D Mixed CNN and Lightweight Attention-Free Transformer for Hyperspectral and LiDAR Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef]
45. Mäyrä, J.; Keski-Saari, S.; Kivinen, S.; Tanhuanpää, T.; Hurskainen, P.; Kullberg, P.; Poikolainen, L.; Viinikka, A.; Tuominen, S.; Kumpula, T.; et al. Tree species classification from airborne hyperspectral and LiDAR data using 3D convolutional neural networks. *Remote Sens. Environ.* **2021**, *256*, 112322. [CrossRef]
46. Mohla, S.; Pande, S.; Banerjee, B.; Chaudhuri, S. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 92–93.
47. Wang, X.; Feng, Y.; Song, R.; Mu, Z.; Song, C. Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data. *Inf. Fusion* **2022**, *82*, 1–18. [CrossRef]
48. Dong, W.; Zhang, T.; Qu, J.; Xiao, S.; Zhang, T.; Li, Y. Multibranch feature fusion network with self-and cross-guided attention for hyperspectral and LiDAR classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
49. Zhao, W.; Du, S. Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4544–4554. [CrossRef]
50. Zhou, W.; Kamata, S.I.; Wang, H.; Wong, M.S.; Hou, H.C. Mamba-in-Mamba: Centralized Mamba-Cross-Scan in Tokenized Mamba Model for Hyperspectral Image Classification. *Neurocomputing* **2024**, *613*, 128751. [CrossRef]
51. Liao, D.; Wang, Q.; Lai, T.; Huang, H. Joint Classification of Hyperspectral and LiDAR Data Based on Mamba. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [CrossRef]

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