Joint Classification of Hyperspectral and LiDAR Data Based on Mamba

Diling Liao, *Student Member, IEEE*, Qingsong Wang, Tao Lai, and Haifeng Huang, *Member, IEEE*

***Abstract*— With the increasing number of remote sensing (RS) data sources, the joint utilization of multimodal data in Earth observation tasks has become a crucial research topic. As a typical representative of RS data, hyperspectral images (HSIs) provide accurate spectral information, while rich elevation information can be obtained from light detection and ranging (LiDAR) data. However, due to the significant differences in multimodal heterogeneous features, how to efficiently fuse HSI and LiDAR data remains one of the challenges faced by existing research. In addition, the edge contour information of images is not fully considered by existing methods, which can easily lead to performance bottlenecks. Thus, a joint classification network of HSI and LiDAR data based on Mamba (HLMamba) is proposed. Specifically, a gradient joint algorithm (GJA) is first performed on LiDAR data to obtain the edge contour data of the land distribution. Subsequently, a multimodal feature extraction module (MFEM) was proposed to capture the semantic features of HSI, LiDAR, and edge contour data. Then, to efficiently fuse multimodal features, a novel deep learning (DL) framework called Mamba, was introduced, and a multimodal Mamba fusion module (MMFM) was constructed. By efficiently modeling the long-distance dependencies of multimodal sequences, the MMFM can better explore the internal features of multimodal data and the interrelationships between modalities, thereby enhancing fusion performance. Finally, to validate the effectiveness of HLMamba, a series of experiments were conducted on three common HSI and LiDAR datasets. The results indicate that HLMamba has superior classification performance compared to other state-of-the-art DL methods. The source code of the proposed method will be available publicly at https://github.com/Dilingliao/HLMamba.**

***Index Terms*— Hyperspectral images (HSIs), joint classification, light detection and ranging (LiDAR), Mamba, multimodal.**

I. Introduction

REMOTE sensing (RS) technology, with its capability to provide real-time, high spatiotemporal resolution earth surface data, is gradually increasing in importance and influence. As a typical representative of RS data, hyperspectral images (HSIs) have been proven to have enormous value in multiple application fields, including surface change detection [1], vegetation monitoring [2], ecological environment monitoring [3], resource management [4], and geology [5].

Received 13 June 2024; revised 17 July 2024 and 17 August 2024; accepted 8 September 2024. Date of publication 12 September 2024; date of current version 26 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62273365 and in part by the Xiaomi Young Talents Program and the Project under Grant 2019ZT08X751. *(Corresponding author: Qingsong Wang.)*

The authors are with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 518000, China (e-mail: liaodling@mail2.sysu.edu.cn; wangqs5@mail.sysu.edu.cn; lait3@mail.sysu.edu.cn; huanghaifeng@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3459709

How to use HSI to achieve precise classification of land features, which has always been one of the key tasks in RS image processing. In HSI classification tasks, early work mainly relied on a single data source. For example, support vector machines (SVMs) [6], random forest (RF) [7], and extreme learning machines (ELMs) [8] are classic methods used for RS land cover classification. In the preprocessing stage of data processing, principal component analysis (PCA) is a new dimensionality reduction strategy [9]. Although the computational complexity has been significantly reduced, the classification performance is still poor. In recent years, deep learning (DL) has been widely applied in various visual tasks due to its inherent ability in automated feature engineering [10], [11], [12]. In HSI classification tasks, popular DL frameworks include autoencoders (AEs) [13], convolutional neural networks (CNNs) [14], [15], [16], [17], [18], [19], generative adversarial networks (GANs) [20], [21], [22], recurrent neural networks (RNNs) [23], capsule networks (CapsNet) [24], and graph convolutional networks (GCNs) [25], [26], [27]. Although existing DL methods have made breakthrough progress, their performance in complex land classification is still not satisfactory. Especially in urban areas, different land cover may use the same material composition, resulting in significant similarity in the spectra emitted by different land cover (such as urban roads and buildings).

Thanks to the improvement of RS imaging technology, the availability of RS images has been improved. Usually, different RS sensors can be used to capture the different characteristics of coverage materials in the same geographical coverage area, generating RS images with unique attributes. For example, detailed spectral information of land cover can be obtained using HSI sensors [28], accurate height information of the earth's surface and land cover can be measured with light detection and ranging (LiDAR) [29], and synthetic aperture radar (SAR) sensors can capture geometric information of the amplitude and phase of ground objects [30]. Research has shown that the joint use of RS images obtained from multiple sensors is beneficial for analyzing land cover information from multiple perspectives [31]. LiDAR data contain accurate elevation information, which helps HSI construct a complete feature representation and effectively improves the discrimination ability of land cover categories with similar spectral responses at different altitudes. However, feature extraction and fusion of HSI and LiDAR data remains a challenging research topic.

Generally, DL-based joint classification networks consist of three parts: multimodal feature extraction, feature fusion, and classification. In the joint classification of HSI and LiDAR,

many excellent DL frameworks have emerged. Notably, CNNs are among the most popular frameworks.

In the aspect of multimodal feature extraction, considering the heterogeneity of HSI and LiDAR data features is crucial for effective information fusion. This is due to the inconsistent amount of image data obtained by different sensors, the different data structures, and the irrelevant physical attributes. Xu et al. [32] proposed a dual-branch CNN, consisting of a dual-channel CNN module and a cascaded CNN module. The former extracts spatial and spectral information from HSI, while the latter extracts elevation information from LiDAR data. Experimental results have shown that using dual-branch CNN can effectively improve model performance. Similarly, Hang et al. [16] proposed a coupled CNN framework for HSI and LiDAR data fusion, utilizing weight-sharing CNN modules to separately extract features from HSI and LiDAR. In addition, CoupledCNN employs both feature-level and decision-level fusion methods in the fusion stage. Although this method has achieved good classification performance, the classification map still suffers from excessive smoothness due to insufficient feature richness and contextual information. Therefore, a new multiscale DL network with self-calibrated convolution (MSNetSC) has been proposed [33]. On one hand, using multiscale self-calibration CNN modules to obtain features from different receptive fields enhances the representation ability of multimodal features. On the other hand, attention fusion modules are employed to fuse various contextual features. Additionally, to address the issues of insufficient extraction of multiscale spatial-spectral information and limitations in feature fusion using existing methods, Gao et al. [34] proposed an adaptive multiscale spatial-spectral enhancement network (AMSSE-Net).

In the aspect of feature fusion, although the above methods fully extract features from multimodal data, they only fuse derived features through feature stacking. This approach ignores the correlation between heterogeneous features, leading to poor decision-making ability of the model. Therefore, an unsupervised HSI and LiDAR feature extraction fusion network [35] has been proposed to explore the correlation of multisource data. In the same way, to enhance cross-modal information exchange, Fang et al. [36] proposed a spatial-spectral enhancement module (S2EM). In addition, to explore the complementarity of multisource information, a new dual-channel spatial, spectral, and multiscale attention convolutional long short-term memory neural network has been proposed [34], called A3CLNN. Zhang et al. [37] proposed an interleaving perception CNN (IP-CNN) for fusing information from multiple sources of data. Cai et al. [38] proposed a graph attention-based multimodal fusion (GAMF) network to explore the correlation and complementarity between different data sources. To learn features from different modalities and exchange information in a more compact way, Wu et al. [39] proposed a cross-channel reconstruction module (CCR-Net), which provides a new perspective for handling complex real-world scenarios and diverse channel data.

In the aspect of classifier design, Ghamisi et al. [40] combined logistic regression (LR) and CNN to design a DL-based classifier.

Although CNN methods excel at representing spatial features, they struggle with the large amount of spectral sequence information contained in HSI. In recent years, Transformers have garnered attention from many researchers due to their powerful ability to model long sequences. In computer vision tasks, the vision Transformer (ViT) [41] was the first successful application of the Transformer framework. Subsequently, there has been a significant amount of outstanding work in the HSI classification task. For example, Hong et al. [42] proposed a spectral Transformer (SF) model for capturing spectral long-range dependencies. Roy et al. [43] attempted to apply ViT to HSI classification tasks and proposed a multimodal fusion Transformer (MFT). To enhance spatial representation ability, Sun et al. [44] proposed the spectral-spatial feature tokenization Transformer (SSFTT). Especially, Transformer was also introduced into the joint classification task of HSI and LiDAR, enhancing the representation ability of long sequence features. Zhao et al. [45] designed a dual-branch network consisting of hierarchical CNN and Transformer (HCT), fully demonstrating the Transformer's powerful remote dependency modeling ability. Additionally, in [46], a new global-local Transformer network (GLTNet) is proposed to capture complex local and global spectral-spatial relationships. To boost the interaction between information, Roy et al. [47] designed a morphological Transformer (MorFormer). Additionally, they also introduced a multimodal DL framework, called Cross-HL [48]. Feng et al. [49] proposed a dynamic scale hierarchical fusion network (DSHFNet) based on multiattention, which addresses the limitations of current mainstream methods in feature extraction and fusion.

Undoubtedly, HSI and LiDAR joint classification methods based on CNN or Transformer have demonstrated superior classification performance. Additionally, a new DL framework, Mamba, has recently been recognized as an effective way to balance global receptive fields and computational efficiency [50], [51], [52]. However, they still have some limitations. The general summary is as follows:

1) Many methods only employ simple CNN structures to extract general information from HSI and LiDAR data, overlooking the diversity of features that could be leveraged.
2) While the Transformer exhibits exceptional long-distance dependency modeling ability and enhances multimodal data representation during fusion, its secondary complexity for images introduces substantial computational overhead.
3) Existing Mamba work focuses solely on single data source processing in HSI classification tasks, resulting in poor classification and generalization performance.

To overcome the limitations of CNN, Transformer, and Mamba, this article proposes a joint classification network of HSI and LiDAR based on Mamba (HLMamba). Specifically, we first employ the gradient joint algorithm (GJA) to obtain edge contour data from LiDAR, which serves as one of the inputs for the multimodal feature extractor. Then, we introduce the Mamba framework for multimodal data fusion. Finally, we design a CNN-based classifier. It is worth noting that CNNs have a strong ability to extract fine features, but it is difficult

to obtain long-distance spatial information. On the contrary, similar to a Transformer, a Mamba can easily obtain long-distance dependencies. Thus, a CNN-based feature extractor is adopted by HLMamba to extract fine features, and a Mamba fusion module is designed to obtain long-distance dependencies within and between modalities. This modeling strategy can be conducive to combining the advantages of the two frameworks.

The main contributions can be summarized as follows:

1) A GJA was proposed to derive edge contour data from LiDAR data, serving as one of the inputs to the multimodal feature extraction module (MFEM). MFEM comprehensively considers the edge contour information of land cover distribution, enhancing the diversity of multimodal features.
2) A low parameter and time-complexity multimodal Mamba fusion module (MMFM) has been introduced to better mine intra and intermodal features. Similar to Transformer, Mamba also possesses powerful long-distance modeling capabilities.
3) To validate the superiority of the HLMamba method, extensive experiments were conducted on three common datasets: Houston, MUUFL, and Trento. The results demonstrate that HLMamba outperforms other state-of-the-art methods. Ablation experiments further confirm the effectiveness of key modules of HLMamba. For instance, compared to Transformer, Mamba requires approximately 30% of the training parameters and 10% less runtime. As we know, this is the first application of the Mamba framework in the joint classification task of HSI and LiDAR.

The remainder of this article is structured as follows. In Section II, a detailed introduction is provided to the overall structure of HLMamba and the principles of each module. Section III presents extensive experimental results and discussions, including ablation experiments, quantitative comparisons, and visual comparisons. Section IV summarizes the work and highlights future research directions.

II. Methodology

In this section, a detailed introduction will be provided to the overall framework of HLMamba, as shown in Fig. 1.

From the figure, it can be observed that the entire process is divided into four stages, namely the stage of input, the stage of multimodal feature extraction, the stage of multimodal feature fusion, and the stage of classification. Specifically, in the stage of the input, HSI, LiDAR, and edge contour data based on GJA are included. In the second stage, spectral-spatial, elevation, and edge contour features are extracted from three types of data using MFEM. In the third stage, intra and intermodal relationships are captured using MMFM. The final stage involves classification, and a classifier is designed using an fully connected (FC) layer in this article.

A. Stage of Input

Assuming the HSI image $X_1 \in \mathbb{R}^{H\times W\times L}$ and LiDAR data $X_2 \in \mathbb{R}^{H\times W\times 1}$ cover the same land area. Among them, H and W represent the height and width of the image, respectively, and L represents the spectral dimension of the HSI image. It is worth noting that an algorithm GJA is proposed for LiDAR data, resulting in the extraction of edge contour data $X_3 \in \mathbb{R}^{H\times W\times 1}$.

Next, we will further introduce GJA. As is well known, HSI contains rich spectral-spatial information, while LiDAR data contains elevation information of land cover. In the joint classification task of HSI and LiDAR, the complementary features of the two modal data contribute to the improvement of classification performance, especially for land cover categories with similar spectral responses but different altitudes. However, although complementary information and enhanced features are beneficial for improving classification accuracy, most existing methods lack feature diversity and are prone to performance bottlenecks. Therefore, in order to improve the diversity of input features, a GJA is proposed for obtaining edge contour data.

For images, gradients can represent the degree of variation of the image at each point. By combining the elevation information of LiDAR data and using a first-order gradient change to obtain the edge and contour information of land cover distribution, it is beneficial to further improve the ability to distinguish land cover, especially for samples in shaded and nonshaded areas of buildings.

In scientific computing and data analysis, gradients represent the rate of change of a function in various directions. A gradient is a vector whose components are the partial derivatives of a function in various directions. For a scalar function f, its gradient can be expressed as

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_N}\right). \tag{1}$$

Among them, $(\partial f/\partial x_k)$ represents the partial derivative of the function f in the kth direction.

In the gradient calculation of 2-D images, both internal and external points are considered. This article employs the central difference method to calculate the gradient of internal points, which is expressed as

$$\left(\frac{\partial f}{\partial x}\right)_k = \frac{f(x_{k+1}) - f(x_{k-1})}{2h}. \tag{2}$$

In the above equation, h is the spacing between sample points. Due to uniform spacing, (2) can be simplified as

$$\left(\frac{\partial f}{\partial x}\right)_k = \frac{f_{k+1} - f_{k-1}}{2h}. \tag{3}$$

Assuming the row and column directions are represented by n and m, respectively, the gradient is calculated using a two-point difference along n and m. This calculation process can be expressed as

$$\begin{cases} \left(\dfrac{\partial f}{\partial x}\right)_n = \dfrac{f_{n+1} - f_{n-1}}{2h} \\ \left(\dfrac{\partial f}{\partial x}\right)_m = \dfrac{f_{m+1} - f_{m-1}}{2h}. \end{cases} \tag{4}$$

For boundary points, forward or backward differences are utilized for gradient calculation, as there may be insufficient

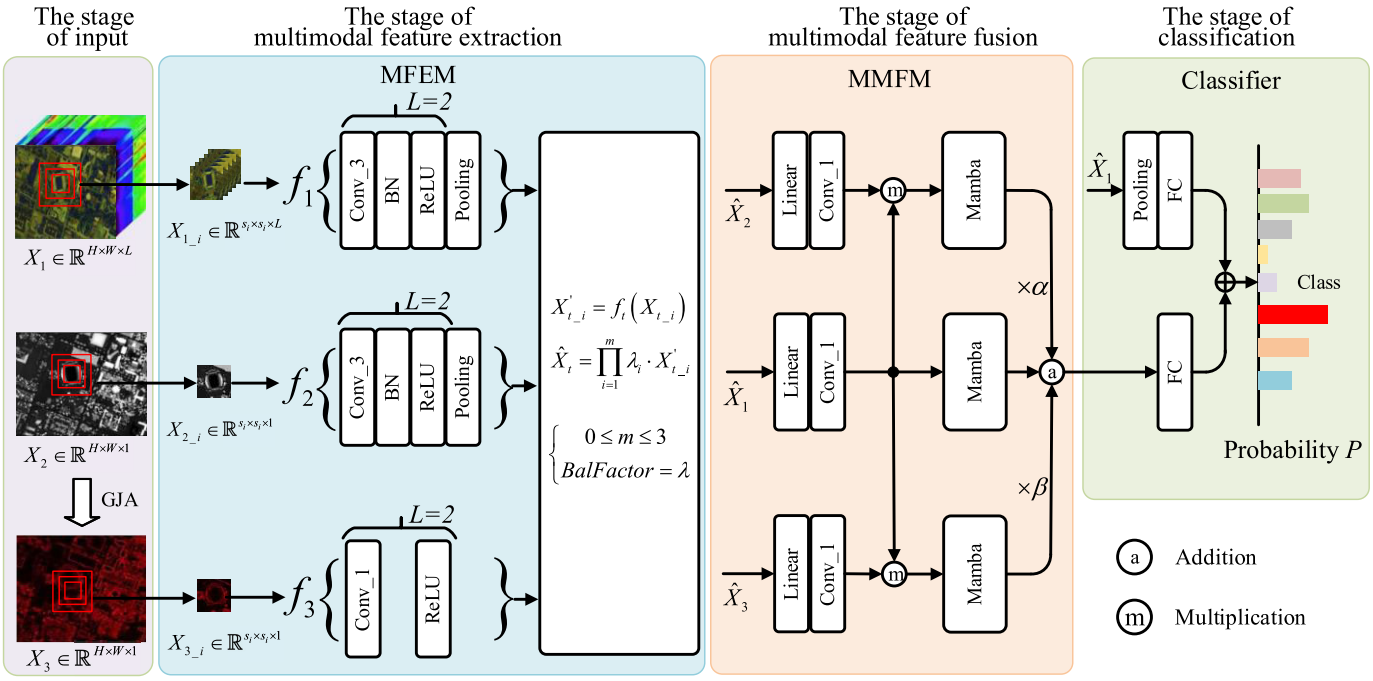

Fig. 1. Overall structure of HLMamba. From the figure, it can be observed that the entire process is divided into four stages, namely the stage of input, the stage of multimodal feature extraction, the stage of multimodal feature fusion, and the stage of classification. HSI, LiDAR, and edge contour data are represented by X_1, X_2, and X_3, respectively. The output of each category is probability P.

neighboring points. The calculation formula is expressed as

$$\begin{cases} \left(\dfrac{\partial f}{\partial x}\right)_0 = \dfrac{f_1 - f_0}{h} \\ \left(\dfrac{\partial f}{\partial x}\right)_{(n-1)\text{or}(m-1)} = \dfrac{f_{(n-1)\text{or}(m-1)} - f_{(n-2)\text{or}(m-2)}}{h} \end{cases} \tag{5}$$

where $(\partial f/\partial x)_0$ is the result of forward difference and $(\partial f/\partial x)_{(n-1)\text{or}(m-1)}$ is the result of backward difference.

Considering the correlation between terrain changes and the surrounding terrain, combining the gradients of directional rows and columns can more comprehensively reflect the change information of terrain data. Fortunately, the modulus of gradients can represent the comprehensive situation of gradient changes in directional n and m. Therefore, the specific formula of GJA can be expressed as follows:

$$X_3 = \sqrt{\underbrace{\left(\sum_{i=0}^{n-1}\left(\frac{\partial f}{\partial x}\right)_i\right)^2}_{\text{The direction of row}} + \underbrace{\left(\sum_{j=0}^{m-1}\left(\frac{\partial f}{\partial x}\right)_j\right)^2}_{\text{The direction of coloum}}}. \tag{6}$$

After the LiDAR data is calculated using (6), data X_3 containing the contour information of the ground features is obtained. Finally, HSI X_1, LiDAR data X_2, and edge contour data X_3 are sent to the next stage by the stage of input.

B. Stage of Multimodal Feature Extraction

In recent years, CNNs, as the mainstream backbone architecture, have fully demonstrated the powerful extraction ability of spatial structural information and local contextual information. In order to obtain rich spectral-spatial, elevation, and edge contour information, an MFEM is constructed using CNN in the second stage, as shown in the second part of Fig. 1. It is worth noting that to further enhance the CNN capture of detailed information, spatial multiscale patch segmented is performed on the input data X_1 before MFEM to obtain $X_{1\_i} \in \mathbb{R}^{s_i \times s_i \times L}$. Here, i represents the scale factor, and s_i represents the spatial size of the ith scale factor. Similarly, the results of X_2 and X_3 processing are denoted as $X_{2\_i} \in \mathbb{R}^{s_i \times s_i \times L}$ and $X_{3\_i} \in \mathbb{R}^{s_i \times s_i \times L}$, respectively. Next, X_{1_i}, X_{2_i}, and X_{3_i} are sent to MFEM to obtain results $\hat{X}_1$, $\hat{X}_2$, and $\hat{X}_3$, respectively. The process of MFEM can be expressed as follows:

$$\hat{X}_t = \prod_{i=1}^{3} \lambda_i \cdot f_t\left(X_{t\_i}\right) \tag{7}$$

where $0 < t \leq 3$, $f_t(\cdot)$ represents the convolutional function, which comprises the Conv_kernel layer, BatchNormalization (BN) layer, activation function ReLU layer, and pooling layer. Additionally, λ_i represents the trainable weighted value corresponding to the ith scale factor. In the designed MFEM, the depth of the CNN is denoted by L. Usually, a deeper model can obtain more abstract features, but it is likely to fall into the local optimal solution. Therefore, to avoid such situations, $L = 2$ is taken.

It is worth noting that $f_3(\cdot)$ only consists of the Conv_kernel layer and activation function ReLU layer, as the introduction of the BN layer would weaken the strength of the edge contour data.

C. Stage of Multimodal Feature Fusion

In the stage of HSI and LiDAR multimodal feature fusion, considering the computational efficiency of Mamba and its

ability to capture long-range dependencies of multimodal features, an MMFM is proposed in this article. It should be noted that the Mamba framework is an improved version of state space models (SSMs) [53], and the concept of SSM originates from continuous linear time-invariant systems.

Assuming the input sequence $x(t)$ is given, SSM maps it to $y(t) \in \mathbb{R}$ by hiding the state $h(t) \in \mathbb{R}^N$. This process can be represented by ordinary differential equations as follows:

$$\begin{cases} h'(t) = Ah(t) + Bx(t) \\ y(t) = Ch(t) \end{cases} \tag{8}$$

where $A \in \mathbb{R}^{N\times N}$ is the state transition matrix as well as $B \in \mathbb{R}^{N\times 1}$ and $C \in \mathbb{R}^{N\times 1}$ are the mapping matrices. However, as (9) describes a continuous system, it is not compatible with deep models of discrete sequences. Therefore, we use the zero-order hold discretization technique to transform the continuous parameters A and B into discrete matrices $\bar{A}$ and $\bar{B}$. The specific process can be expressed as follows:

$$\begin{cases} \bar{A} = \exp(\Delta A) \\ \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B \\ \quad \approx (\Delta A)^{-1}(\Delta A)(\Delta B) \\ \quad = \Delta B. \end{cases} \tag{9}$$

In the above equation, Δ is a scale parameter of time. After discretization, the SSM discrete system can be represented as follows:

$$\begin{cases} h_t = \bar{A}h_{t-1} + \bar{B}x_t \\ y_t = Ch_t. \end{cases} \tag{10}$$

Usually, to enhance computational efficiency, convolution is employed to accelerate the linear recursive process mentioned above. Finally, this process can be represented as follows:

$$y = x * \bar{K}. \tag{11}$$

Among them, $K = (C\bar{B}, C\bar{A}\bar{B}, \ldots, C\bar{A}^{L-1}\bar{B})$ is the structured convolution kernel, and L is the length of the sequence.

Although SSM has linear time complexity, capturing long-range dependency relationships is difficult with it. Therefore, a selective SSM (S6) was designed by Mamba to overcome this limitation of SSM. The Mamba structure is shown in Fig. 2. It is worth noting that the mapping matrix in S6 depends on the input sequence, allowing selective attention to each input unit. Specifically, $B \in \mathbb{R}^{B\times L\times N}$, $C \in \mathbb{R}^{B\times L\times N}$, and $\Delta \in \mathbb{R}^{B\times L\times D}$ can be dynamically calculated through the input sequence $x \in \mathbb{R}^{B\times L\times D}$. The relationship between input x and output y can be expressed as follows:

$$y = \text{Mamba}(x) \tag{12}$$

where Mamba$(\cdot)$ is the Mamba function. To better understand the Mamba process, it can be specifically expressed as

$$\begin{cases} \text{resz} = \sigma(f_{\text{linear}}(f_{\text{norm}}(x))) \\ z = \sigma(f_{\text{3DConv}}(f_{\text{linear}}(f_{\text{norm}}(x)))) \\ y = f_{\text{linear}}(f_{\text{norm}}(\text{SSM}(z)) \cdot \text{resz}) + x. \end{cases} \tag{13}$$

In the above equation, $\sigma(\cdot)$ is the SiLU activation function. $f_{\text{linear}}(\cdot)$, $f_{\text{norm}}(\cdot)$, and $f_{\text{3DConv}}(\cdot)$ represent linear, normalized, and convolutional functions, respectively, with a kernel size of $1 \times 1 \times 1$. x is the input sequence and y is the output sequence of Mamba.

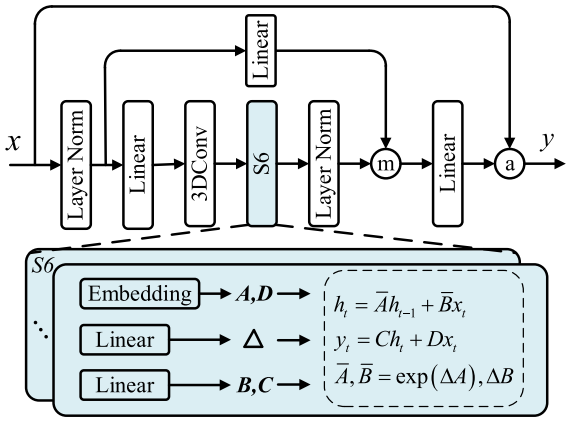

Fig. 2. Structure of Mamba.

The structure of MMFM is shown in Fig. 1. First, the three input data $\hat{X}_1$, $\hat{X}_2$, and $\hat{X}_3$ are convolved to obtain advanced features of spectral-spatial $\hat{X}_{1\_\text{Conv}}$, elevation $\hat{X}_{2\_\text{Conv}}$, and edge contour $\hat{X}_{3\_\text{Conv}}$. Next, $\hat{X}_{1\_\text{Conv}}$ is fused into the features $\hat{X}_{2\_\text{Conv}}$ and $\hat{X}_{3\_\text{Conv}}$ to combine features from different modalities. Then, the three results are sent to Mamba to fully explore the feature representation capabilities intra and intermodalities. Additionally, two learnable parameters are used to weight the elevation and edge contour branch data separately. Finally, the three high-level semantic features are further fused. The process of MMFM can be represented as follows:

$$\begin{cases} \hat{X}_{1\_M} = \underbrace{\text{Mamba}\left(f_{\text{Conv}}\left(f_{\text{linear}}\left(\hat{X}_1\right)\right)\right)}_{\text{HSI spectral-spatial features}} \\ \hat{X}_{2\_M} = \underbrace{\text{Mamba}\left(f_{\text{Conv}}\left(f_{\text{linear}}\left(\hat{X}_2\right)\right) \cdot f_{\text{Conv}}\left(f_{\text{linear}}\left(\hat{X}_1\right)\right)\right)}_{\text{LiDAR elevation features}} \\ \hat{X}_{3\_M} = \underbrace{\text{Mamba}\left(f_{\text{Conv}}\left(f_{\text{linear}}\left(\hat{X}_3\right)\right) \cdot f_{\text{Conv}}\left(f_{\text{linear}}\left(\hat{X}_1\right)\right)\right)}_{\text{LiDAR based edge contour features}} \end{cases} \tag{14}$$

$$X = \hat{X}_{1\_M} + \alpha \cdot \hat{X}_{2\_M} + \beta \cdot \hat{X}_{3\_M}. \tag{15}$$

In the above equation, X represents the output of the MMFM model after fusion. α and β are learnable parameters.

The steps of the MMFM are shown in Algorithm 1 in detail.

D. Stage of the Classification

In order to improve computational efficiency and avoid overfitting, a classifier is designed using a simple linear structure during the classification stage in this article. Furthermore, considering that spectral-spatial features can enhance the discriminative ability of the classifier, the classifier also incorporates the spectral-spatial feature $\hat{X}_1$ of HSI. The process of the stage of the classification can be expressed as follows:

$$X_{\text{class}} = f_{\text{linear}}(X) + f_{\text{linear}}\left(f_{\text{pooling}}\left(\bar{X}_1\right)\right). \tag{16}$$

X_{class} represents the output features of the classifier, and the feature length is the number of categories.

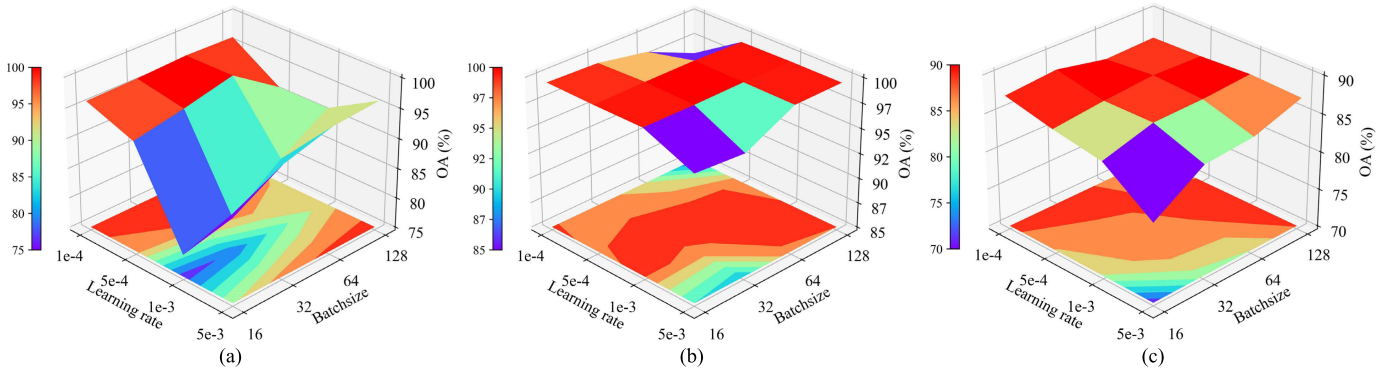

Fig. 3. Impact of learning rate and batch size on performance. (a) Houston dataset, (b) MUUFL dataset, and (c) Trento dataset.

Algorithm 1 MMFM

Input: $\hat{X}_1, \hat{X}_2, \hat{X}_3$:$(B, L, D)$, $T = 500$, $L = 2$

1: **Initialization:**

$A : (D, N) \leftarrow parameter$

$B : (B, L, N) \leftarrow linear_B(\hat{X}_i)$

$C : (B, L, N) \leftarrow linear_C(\hat{X}_i)$

$\Delta : (B, L, N) \leftarrow \tau_\Delta(parameter + linear_\Delta(\hat{X}_i))$

$\bar{A}, \bar{B} : (B, L, D, N) \leftarrow discretize(\Delta, A, B)$

2: **for** $t = 1$ **to** T **do**

3: Calculate the convolutional output of $\hat{X}_i (1 \leq i \leq 3)$.

4: Calculate the mid-level fusion of HSI, LiDAR, and Edge contour.

5: Update SSM module

6: Calculate the Mamba output.

7: Calculate the weighted of learnable parameter α and

8: Calculate deep-fusion of Mamba results.

9: **end for**

Output:

Feature X after multimodal fusion.

III. Experimental Results and Analysis

To validate the effectiveness of HLMamba, extensive experiments were conducted on three HSI-LiDAR datasets. First, detailed information about the datasets used for the experiments was provided. Next, the experimental setup and evaluation metrics were outlined. Following this, the ablation experiments for the proposed method were presented, and the Mamba fusion module was compared and analyzed against the Transformer fusion module. Finally, a series of experimental comparisons between HLMamba and the state-of-the-art methods were conducted, including quantitative comparisons, visual comparisons, and accuracy comparisons with different training sample sizes.

A. Dataset Description

The experiment used three common datasets: Houston, MUUFL, and Trento datasets. Next, a detailed introduction to these three datasets will be provided.

1) Houston: [1]

This dataset consists of scenes collected by the Compact Airborne Spectral Imager (CASI) on the University of Houston campus and adjacent urban areas. It includes HSI and LiDAR-based Digital Surface Model (DSM) data, with a spatial size of 345 × 1905 pixels and a spatial resolution of 2.5 m. The HSI data contains 144 spectral bands, covering a wavelength range of 0.38–1.35 μm. In addition, the Houston dataset comprises a total of 15 029 labeled samples, representing 15 different types of land cover categories.

2) MUUFL: [2]

This dataset was collected from the Gulfport campus of the University of South Mississippi, Long Beach, MS, USA, and includes both HSI and LiDAR datasets. The spatial size of the HSI data is 325 × 220 pixels, with 64 spectral channels. It is worth noting that the LiDAR data was captured by an ALTM sensor using a 1064 nm wavelength laser. This dataset contains 53 687 labeled samples across 11 types of land cover categories.

3) Trento: [3]

This dataset was collected in a rural area of Trento in southern Italy and includes both HSI and LiDAR datasets. The spatial size is 166 × 600 pixels, with a spatial resolution of 1 m. The HSI data spans a wavelength range of 0.42–0.99 μm, containing 63 spectral bands. Additionally, the LiDAR data were obtained using a 3100EA sensor. This dataset contains 30 214 labeled samples across six types of land cover categories. The detailed category information of the three datasets is shown in Table I.

B. Experimental Setup

All experiments in this study were implemented using the Pytorch framework and were conducted on a server with a 2.1 GHz Intel[4] Xeon[4] Silver 8352V CPU and a 24 GB Nvidia GeForce RTX 4090 GPU. The HLMamba method employs an Adam optimizer with the training rounds set to 500. The spatial sizes for multiscale input are set to 8 × 8, 16 × 16, and 24 × 24.

To objectively reflect performance, three common evaluation metrics are used: overall accuracy (OA), average accuracy

[1]https://www.grss-ieee.org/technical-committees/image-analysis-and-data-fusion/?tab=past-data-fusion-contests

[2]https://github.com/GatorSense/MUUFLGulfport/

[3]https://drive.google.com/drive/folders/1HK3eL3loI4Wd-RFr1psLLmVLTVDLctGd

[4]Registered trademark.

TABLE I
DETAILED CATEGORY INFORMATION FOR THREE DATASETS

No.	Houston			MUUFL			Trento		
	Class name	Train.	Test.	Class name	Train.	Test.	Class name	Train.	Test.
C1	Healthy grass	20	1231	Trees	20	23226	Apple Trees	20	4014
C2	Stressed grass	20	1234	Mostly Grass	20	4250	Buildings	20	2883
C3	Synthetic grass	20	677	Mixed Ground Surface	20	6862	Ground	20	459
C4	Trees	20	1224	Dirt and Sand	20	1806	Woods	20	9103
C5	Soil	20	1222	Road	20	6667	Vineyard	20	10481
C6	Water	20	305	Water	20	446	Roads	20	3154
C7	Residential	20	1248	Buildings Shadow	20	2213			
C8	Commercial	20	1224	Buildings	20	6220			
C9	Road	20	1232	Sidewalk	20	1365			
C10	Highway	20	1207	Yellow Curb	20	163			
C11	Railway	20	1215	Cloth Panels	20	249			
C12	Parking Lot 1	20	1213						
C13	Parking Lot 2	20	449						
C14	Tennis Court	20	408						
C15	Running Track	20	640						
-	Total	300	14729	Total	220	53467	Total	120	30094

TABLE II
CONTRIBUTION OF HLMAMBA COMPONENTS (%)

Datasets		Baseline1	Baseline1+Mamba	Baseline2+Mamba
Houston	OA	89.92	96.54**(+6.62)**	96.80**(+0.26)**
	AA	91.16	97.14	97.35
	K×100	88.99	96.26	96.54
MUUFL	OA	78.17	86.13**(+7.96)**	87.96**(+1.83)**
	AA	79.31	84.20	83.29
	K×100	72.33	81.98	84.23
Trento	OA	98.98	99.31**(+0.33)**	99.54**(+0.23)**
	AA	97.69	98.90	98.91
	K×100	98.63	99.35	99.38

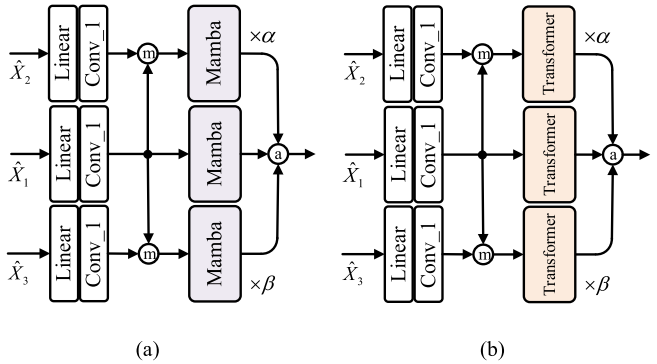

Fig. 4. Structure of multimodal fusion. (a) Mamba-based multimodal fusion module. (b) Transformer-based multimodal fusion module.

(AA), and Kappa coefficient (K). Additionally, for a fair comparison, all results are averaged over ten experiments. In the comparison method, if the original paper provides model parameters and training methods, we will take the optimal parameters from the original paper. If the original paper does not provide model parameters and training methods, we will conduct experiments to obtain the optimal parameters for this method and use the same training method as in this article.

C. Parameter Sensitivity Analysis

In the training parameters of the network, the learning rate and batch size directly influence the weight updates in the model. Usually, selecting the optimal combination of these parameters requires experimentation and optimization to achieve both fast and accurate training results. Specifically, an appropriate learning rate facilitates stable convergence of the model to global or local optimal solutions. Meanwhile, an optimal batch size enhances generalization ability and accelerates training speed. Therefore, in this section, combination experiments were conducted on three datasets to determine the optimal learning rate and batch size for the HLMamba method. The selected learning rate set is $\{1e-4, 5e-4, 1e-3, 5e-3\}$, and the selected batch size set is {16, 32, 64, 128}. The experimental results are shown in Fig. 3, where red represents the area with the highest contour value and dark blue represents the area with the lowest contour value.

From the results, it can be seen that in the Houston dataset, better results are often obtained with a smaller learning rate. In the MUUFL dataset, it is observed that batch size has little effect on performance when the learning rates are 0.001 and 0.0005. In the Trento dataset, it is evident that a smaller learning rate and a larger batch size often result in higher OA values. Therefore, a learning rate of 0.0005 and a batch size of 64 are determined to be optimal for HLMamba.

D. Ablation Experiment

1) Contribution of Each Component of HLMamba: In order to verify the effectiveness of Mamba fusion and edge contour (Edge) data in the HLMamba method, we conducted ablation experiments. Among them, Baseline1 represents the input of HSI and LiDAR data to MFEM, while Baseline2 represents the input of HSI, LiDAR, and Edge data to MFEM. The experimental results are shown in Table II. It can be observed that the

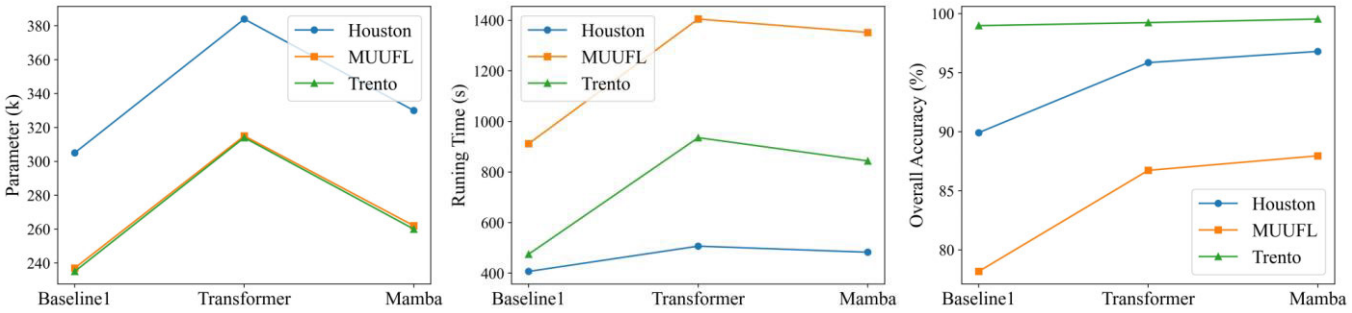

Fig. 5. Comparison results between Mamba and Transformer modules. Comparison results of (a) training parameters, (b) training time, and (c) classification accuracy OA %.

TABLE III
CLASSIFICATION ACCURACY (%) OBTAINED BY DIFFERENT METHODS ON HOUSTON DATASET (BEST RESULTS DISPLAYED IN BOLD)

No.	HSI		HSI+LiDAR							Ours
	ViT	SF	CoupledCNN	GAMF	HCT	MFT	GLTNet	Cross-HL	DSHFNet	HLMamba
C1	90.31	88.78	90.73	87.40	**98.21**	93.41	88.78	96.26	87.89	96.67
C2	96.72	98.11	94.48	98.78	82.98	94.89	98.94	**100**	98.46	**100**
C3	99.84	99.84	98.07	100	98.22	100	98.52	99.55	**100**	**100**
C4	97.02	99.37	96.89	93.79	**100**	91.91	99.50	97.30	99.59	98.36
C5	91.68	91.77	96.64	99.67	**100**	95.58	99.75	**100**	99.83	**100**
C6	93.14	99.00	88.52	**100**	88.19	**100**	99.01	99.01	98.03	98.69
C7	79.38	82.03	77.96	86.69	72.99	81.65	95.35	92.86	95.43	**97.68**
C8	68.08	70.76	90.11	76.14	83.57	90.11	83.98	75.24	**91.99**	86.19
C9	71.86	77.70	73.62	87.01	56.57	90.82	91.15	79.87	65.25	**94.64**
C10	73.45	77.55	82.35	91.79	**99.91**	92.54	92.95	94.53	89.06	96.60
C11	74.81	75.67	88.72	75.14	50.53	82.96	94.81	90.78	**95.39**	93.33
C12	67.86	70.40	77.82	91.01	56.63	92.74	97.93	95.79	90.02	**98.52**
C13	49.89	48.20	94.65	93.76	74.61	97.99	99.33	94.43	93.54	**99.77**
C14	93.48	89.93	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
C15	95.37	97.04	99.37	97.81	**100**	96.25	**100**	**100**	99.68	**100**
OA	82.32	83.90	88.61	90.36	82.38	92.00	95.17	93.35	92.50	**96.80**
AA	82.86	84.41	90.00	91.94	84.16	93.39	96.01	94.38	93.60	**97.36**
K×100	80.89	82.60	87.69	89.58	80.94	91.36	94.77	92.81	91.89	**96.54**

Mamba fusion of Baseline1 (HSI + LiDAR) can significantly improve classification accuracy, especially for Houston and MUUFL datasets. On this basis, the introduction of Edge data effectively improves classification performance. The results of the ablation experiment fully demonstrate the contribution of LiDAR-based edge data to classification performance and the effectiveness of the Mamba-based multimodal fusion method.

2) Comparison Between Mamba and Transformer Modules: In HSI classification tasks, modeling based on the Mamba framework offers the advantage of small parameters and rapid acquisition of long-distance dependency relationships. To verify this advantage, we replaced the Mamba module in the multimodal fusion stage with the traditional Transformer module and conducted experimental comparisons of parameters, runtime, and classification accuracy OA. The structure is shown in Fig. 4, and the experimental results are presented in Fig. 5.

"Baseline1" represents the MFEM without any fusion module structure. "Transformer" and "Mamba" denote the fusion modules based on Transformer and Mamba, respectively. Each line color corresponds to a different dataset. Through comparison, the following conclusions can be drawn: 1) compared with Transformer, the transformation selection process used by Mamba to obtain relationships reduces the required training parameters by approximately 30%. 2) The Mamba model exhibits a shorter runtime compared to the Transformer model, with a reduction of about 10%. 3) In terms of classification accuracy, Mamba outperforms Transformer.

E. Performance Comparison

To evaluate the performance of HLMamba, we compared it with several state-of-the-art classification methods. These include two HSI-based single-source classification methods, ViT [41] and SF [42], as well as seven HSI and LiDAR-based

TABLE IV
CLASSIFICATION ACCURACY (%) OBTAINED BY DIFFERENT METHODS ON THE MUUFL DATASET
(BEST RESULTS DISPLAYED IN BOLD)

No.	HSI		HSI+LiDAR							Ours
	ViT	SF	CoupledCNN	GAMF	HCT	MFT	GLTNet	Cross-HL	DSHFNet	HLMamba
C1	96.98	96.78	87.73	89.99	85.11	88.94	88.37	89.39	75.91	**93.67**
C2	56.08	56.33	64.54	75.01	72.18	75.97	85.48	70.87	**85.67**	80.82
C3	70.00	69.07	53.78	65.06	64.83	59.80	64.80	71.45	66.21	**71.55**
C4	61.31	60.26	87.04	90.75	93.68	81.00	86.43	90.69	**98.78**	79.07
C5	85.28	85.94	69.43	88.34	72.35	69.11	77.21	82.61	**89.89**	87.66
C6	40.62	79.90	**100**	99.77	99.55	99.77	**100**	**100**	94.61	**100**
C7	43.08	43.44	80.70	93.04	84.90	85.04	88.29	**95.30**	88.97	87.57
C8	84.96	84.63	91.27	94.11	93.24	85.77	94.01	93.29	86.76	**97.30**
C9	37.74	40.84	42.56	51.57	53.26	45.71	60.51	60.95	0.36	**65.27**
C10	21.75	25.16	73.00	**84.66**	74.33	71.77	76.68	69.32	0	58.90
C11	47.83	49.88	99.59	95.58	**99.59**	95.98	95.98	**99.59**	97.18	94.38
OA	74.90	75.44	78.31	85.14	80.46	79.87	83.70	84.86	77.86	**87.96**
AA	58.69	60.20	77.24	**84.36**	81.18	78.08	83.44	83.95	71.31	83.29
$K\times100$	68.44	69.08	72.05	80.79	75.06	74.08	79.05	80.39	72.22	**84.23**

TABLE V
CLASSIFICATION ACCURACY (%) OBTAINED BY DIFFERENT METHODS ON TRENTO DATASET
(BEST RESULTS DISPLAYED IN BOLD)

No.	HSI		HSI+LiDAR							Ours
	ViT	SF	CoupledCNN	GAMF	HCT	MFT	GLTNet	Cross-HL	DSHFNet	HLMamba
C1	58.83	62.67	98.77	95.26	99.15	97.48	98.95	99.32	93.77	**99.45**
C2	71.00	72.63	94.27	93.09	97.88	98.02	**99.23**	84.59	98.40	99.17
C3	63.69	67.28	99.12	**100**	97.82	95.42	97.82	**100**	97.82	97.38
C4	92.90	92.18	**100**	**100**	**100**	99.89	**100**	**100**	**100**	**100**
C5	88.14	88.22	99.34	98.35	98.73	99.97	**100**	99.89	**100**	**100**
C6	72.85	72.45	94.83	93.27	94.83	93.84	96.48	94.99	95.27	**97.43**
OA	80.15	80.95	98.50	97.43	98.67	98.72	99.39	97.87	98.49	**99.54**
AA	74.57	75.90	97.73	96.67	98.07	97.44	98.75	96.47	97.55	**98.91**
$K\times100$	73.95	94.95	98.00	96.58	98.22	98.29	99.18	97.16	97.98	**99.38**

multisource joint classification methods, namely CoupledCNN [16], GAMF [38], MFT [43], HCT [45], GLTNet [46], Cross-HL [48], and DSHFNet [49].

1) Quantitative Results and Analysis: Tables III–V present the OA, AA, K, and accuracy values for each category obtained by the proposed method and all comparison methods on three datasets, with the best results displayed in bold. The results indicate that the proposed method HLMamba achieves significantly better classification results than other methods. Overall, the joint classification method based on HSI and LiDAR data yields higher OA values than the HSI single-source classification method. This is because the multimodal joint classification method obtains more diverse features, improving the representation ability of high-level features. Specifically, on the Houston dataset, compared to ViT and SF, OA values of HLMamba are 14.48% and 12.90% higher, respectively, and all category accuracies are higher than these two methods. Compared with GLTNet, which has the best performance among the comparison methods, HLMamba exhibits higher OA, AA, and K values. It is noteworthy that our method achieved the highest-class accuracy values in nine categories, while GLTNet only excelled in two categories. This is attributed to the HLMamba method's incorporation of edge contour information, which mitigates the issue of insufficient edge information between categories. The proposed method continues to perform well on the MUUFL dataset. Notably, HLMamba achieved the best result in C8 (building), surpassing the suboptimal GAMF method by 3.19%. This underscores the utility of LiDAR and edge contour information in distinguishing between building and nonbuilding categories. Although good performance has been achieved, there are still some shortcomings. After introducing edge contour and elevation information, the category features of different altitudes are enhanced, which leads to suboptimal AA values obtained by HLMamba. Fortunately, HLMamba shows significant advantages in OA and K. In the Trento

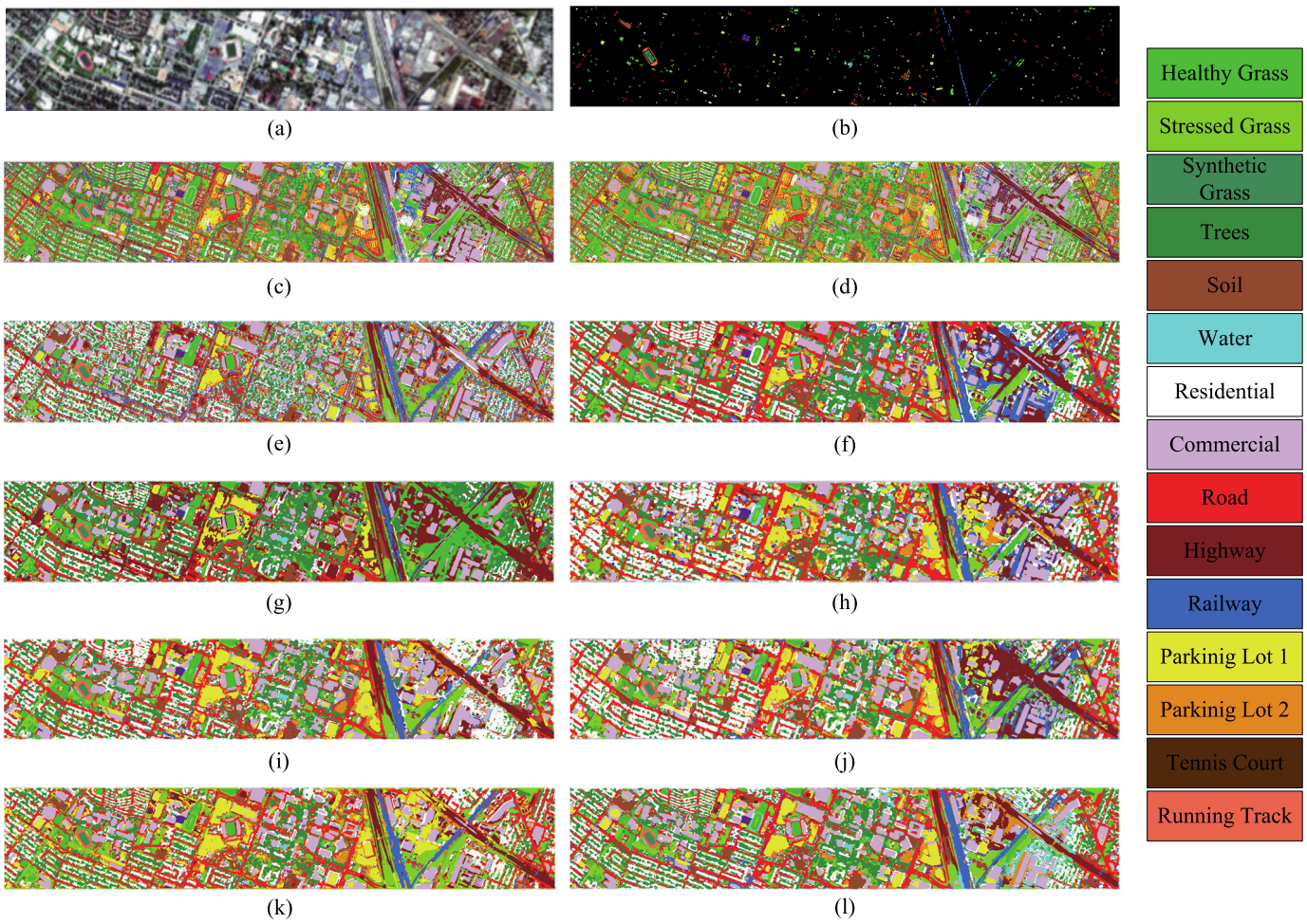

Fig. 6. Classification result graphs obtained by different methods on the Houston dataset. (a) Pseudo color map, (b) ground truth map, (c) ViT (82.32%), (d) SF (83.90%), (e) CoupledCNN (88.61%), (f) GAMF (90.36%), (g) HCT (82.38%), (h) MFT (92.00%), (i) GLTNet (95.17%), (j) Cross-HL (93.35%), (k) DSHFNet (92.50%), and (l) HLMamba (96.80%).

dataset, our method outperformed others in four categories, with the classification accuracy of the C2 (building) nearing 100%. This outcome is also attributed to HLMamba's fusion strategy involving HSI, LiDAR, and edge contour information. In addition, the OA, AA, and K values of this method achieved the best results.

2) Visualization Results and Analysis: Figs. 6–8 illustrate the classification results of all methods on the three datasets. It is evident that the HSI single-source classification method based on Transformer exhibited the poorest visualization results. Although this approach can establish spectral long-range dependencies, it struggles to capture precise spatial features. Particularly on the Houston dataset, the residential areas appear scattered and the land cover areas are small, leading to significant misclassification in the results obtained by ViT and SF. Among the HSI and LiDAR joint methods, the classification results of CoupledCNN and HCT exhibit a significant amount of salt and pepper noise, as this method does not fully consider the fusion of heterogeneous features. On the other hand, GAMF, MFT, GLTNet, Cross-HL, and DSHFNet employ feature extraction and fusion structural modeling to achieve satisfactory classification results. In contrast, our method fully considers feature diversity and fusion strategies, resulting in a clearer classification map.

In addition, to directly analyze the feature representation ability of the proposed method HLMamba, we also compared the visualization results of the representative methods GAMF based on GCN, DSHFNet based on CNN, and GLTNet based on Transformer using t-Distributed Stochastic Neighbor Embedding (t-SNE). The results are shown in Figs. 9–11. Each point corresponds to a feature, and its color represents the specified class label. It is evident that similar categories are often clustered together, minimizing intraclass variance. From the figures, it can be seen that the performance of GAMF based on GCN is relatively poor on the three datasets, which reflects its limited feature representation ability. Although DSHFNet based on CNN achieved a good interclass gap, the intraclass gap was relatively large. The GLTNet method based on Transformer achieved good visualization results, and our method HLMamba's results are more clustered within theclass. This is particularly evident for C2 (yellow) on the MUUFL dataset and C5 (purple) and C6 (brown) on the Trento dataset. The above results indicate that HLMamba can effectively utilize the advantages of multimodal data and enhance the representation ability of fused features.

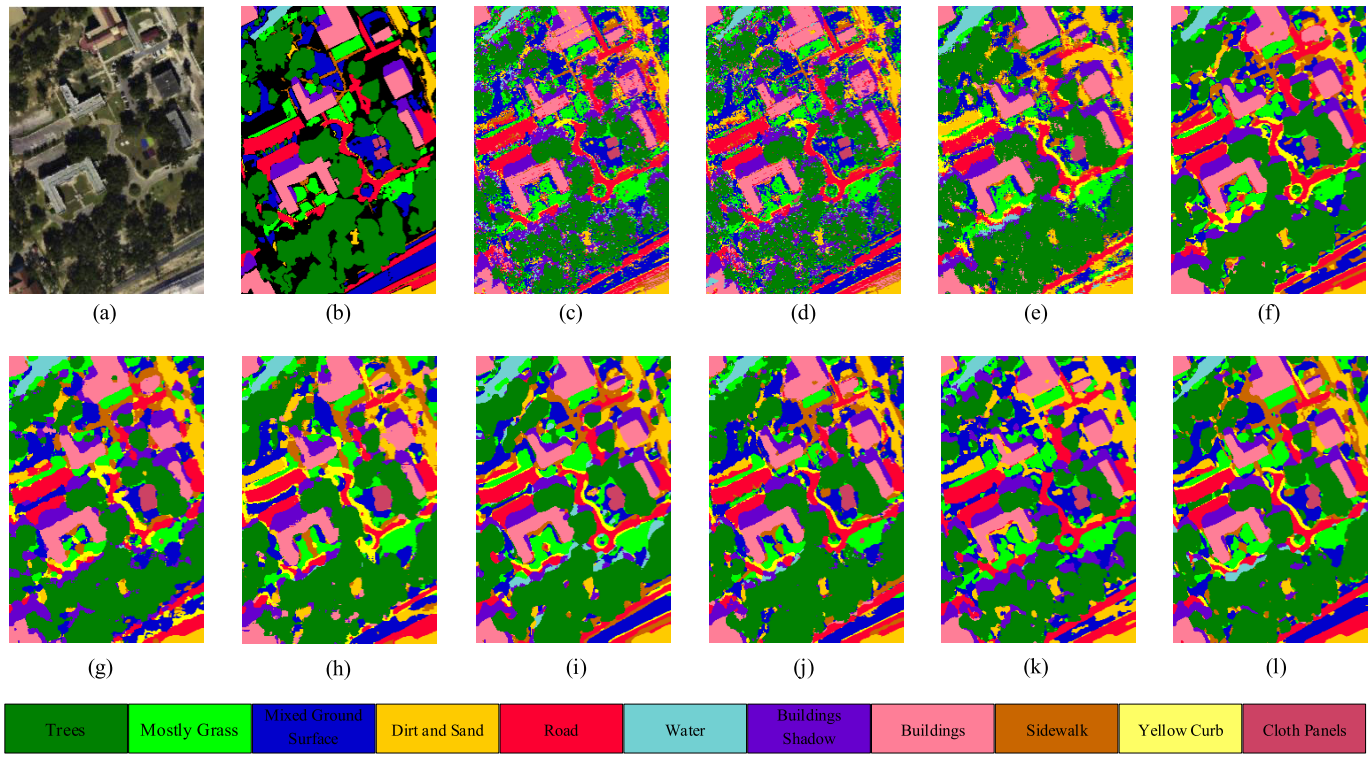

Fig. 7. Classification result graphs obtained by different methods on the MUUFL dataset. (a) Pseudo color map, (b) ground truth map, (c) ViT (74.90%), (d) SF (75.44%), (e) CoupledCNN (78.31%), (f) GAMF (85.14%), (g) HCT (80.46%), (h) MFT (79.87%), (i) GLTNet (83.70%), (j) Cross-HL (84.86%), (k) DSHFNet (77.86%), and (l) HLMamba (87.96%).

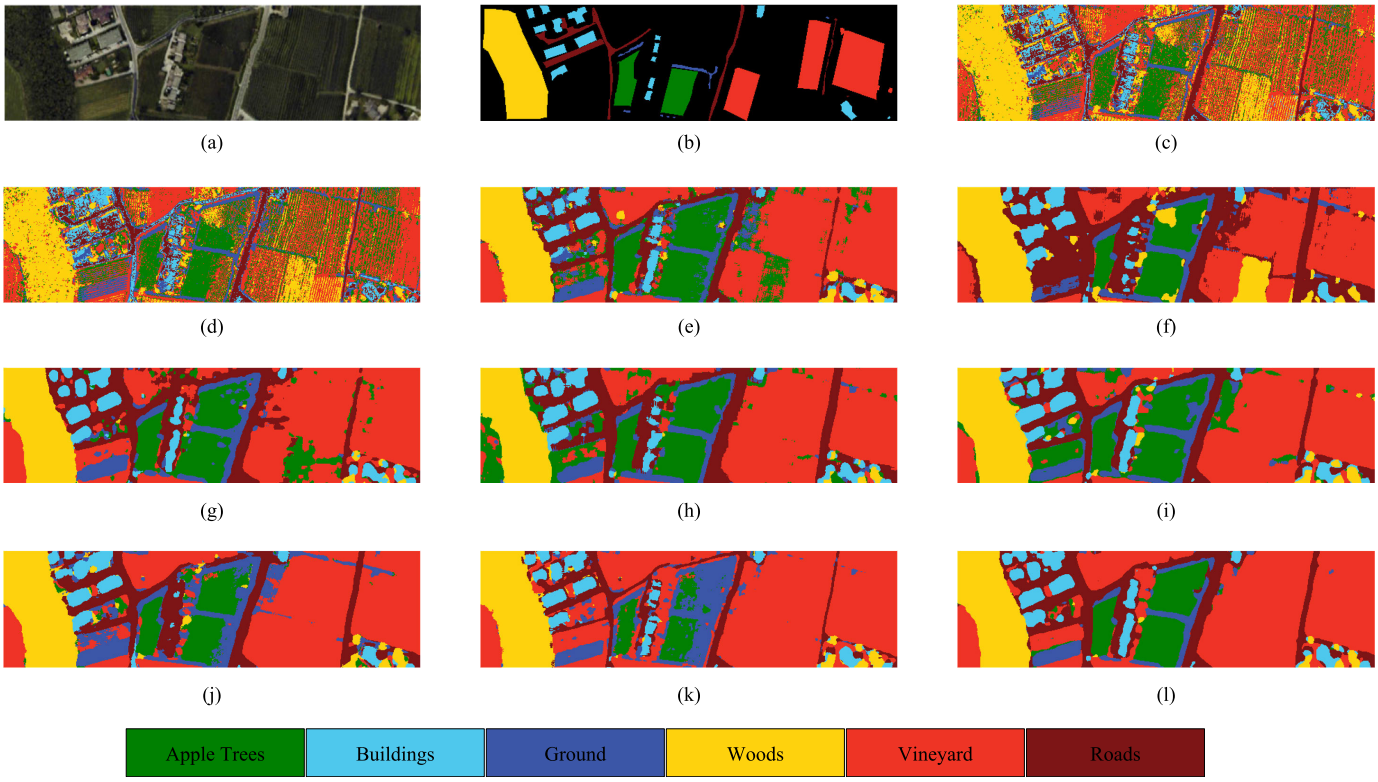

Fig. 8. Classification result graphs obtained by different methods on the Trento dataset. (a) Pseudo color map, (b) ground truth map, (c) ViT (80.15%), (d) SF (80.95%), (e) CoupledCNN (98.50%), (f) GAMF (97.43%), (g) HCT (98.67%), (h) MFT (98.72%), (i) GLTNet (99.39%), (j) Cross-HL (97.87%), (k) DSHFNet (98.49%), and (l) HLMamba (99.54%).

F. Comparison of Different Training Sample Sizes

The number of various training samples significantly influences the experimental outcomes. To analyze the classification performance of the proposed method HLMamba under different training samples, this section selected 20, 40, 60, 80, and 100 training samples to conduct experiments on all methods.

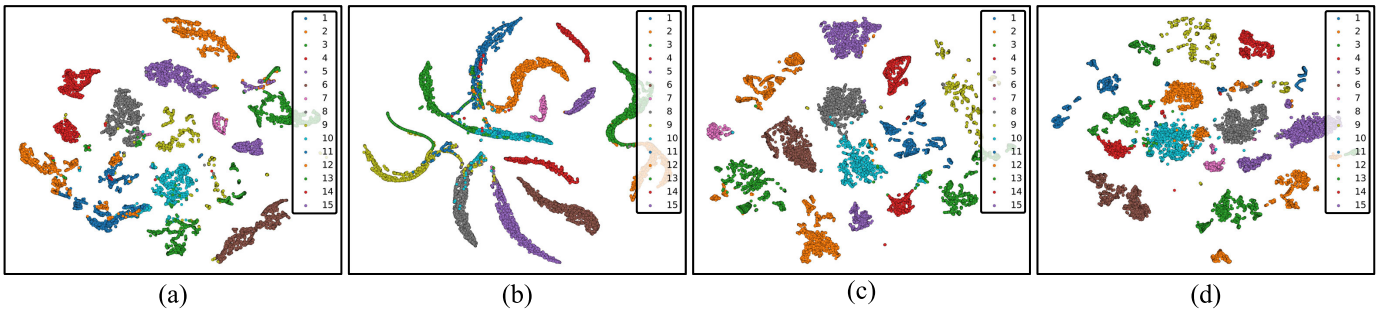

Fig. 9. t-SNE visualization results on Houston dataset. (a) GAMF, (b) DSHFNet, (c) GLTNet, and (d) HLMamba.

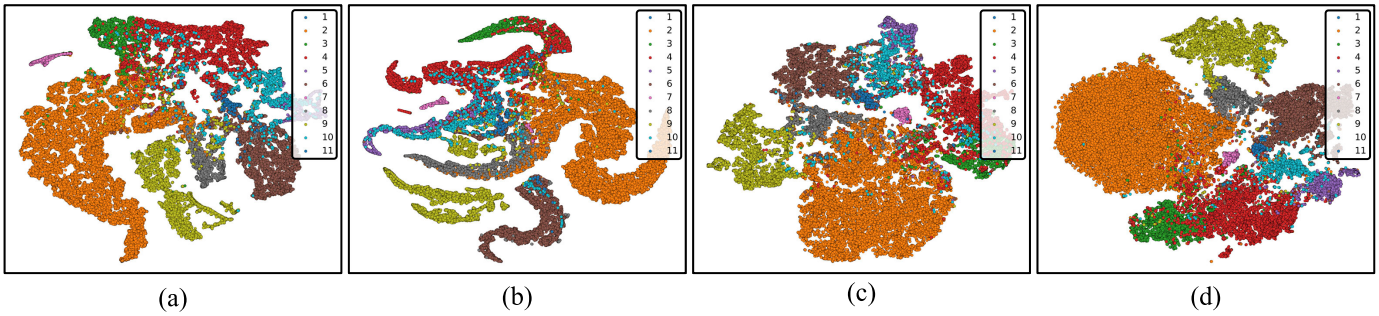

Fig. 10. t-SNE visualization results on MUUFL dataset. (a) GAMF, (b) DSHFNet, (c) GLTNet, and (d) HLMamba.

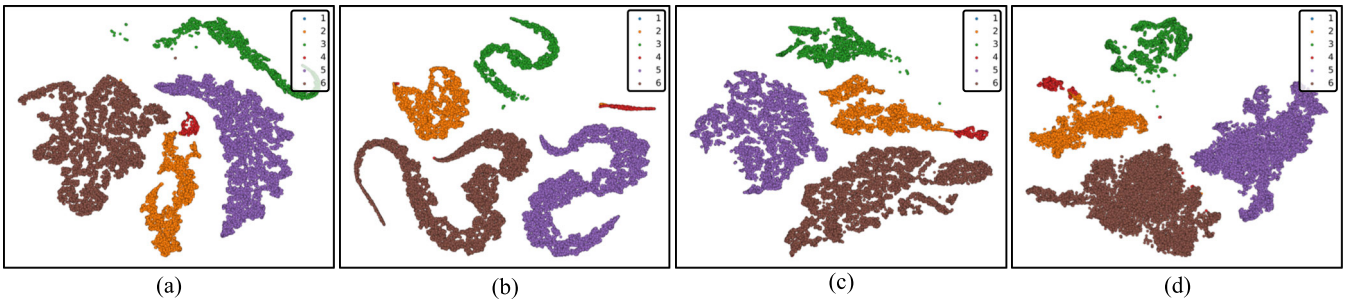

Fig. 11. t-SNE visualization results on Trento dataset. (a) GAMF, (b) DSHFNet, (c) GLTNet, and (d) HLMamba.

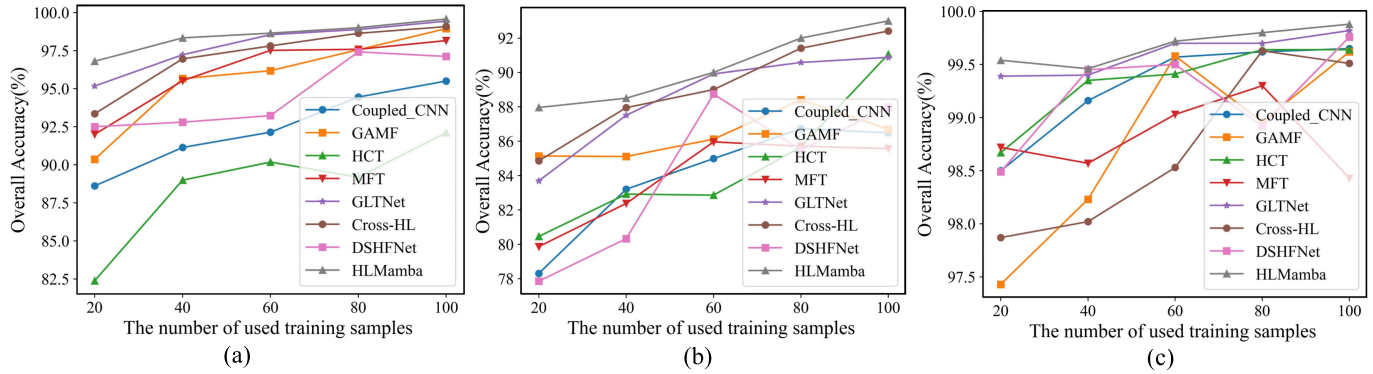

Fig. 12. Impact of different methods using different training sample sizes on the classification accuracy OA value. (a) Houston dataset, (b) MUUFL dataset, and (c) Trento dataset.

The experimental results are depicted in Fig. 12, where different colored lines represent different classification methods. From the results, it can be observed that HLMamba not only maintains good competitiveness with sufficient samples but also exhibits strong feature learning ability even with limited samples.

G. Experiment of the Disjoint Train-Test for Houston Dataset

There are usually two sampling strategies to divide the training and testing sets: random sampling, where each category is randomly divided into training and testing sets according to a certain proportion or a fixed number of samples per category, and disjoint sampling[5], which divides them into disjoint training and testing sets. In order to further observe the performance of the two sampling strategies, in addition to the random sampling results of the Houston dataset provided in Section III-E of this chapter, the experiment on disjoint sampling was also conducted on the Houston dataset. The

[5]http://dase.grss-ieee.org/

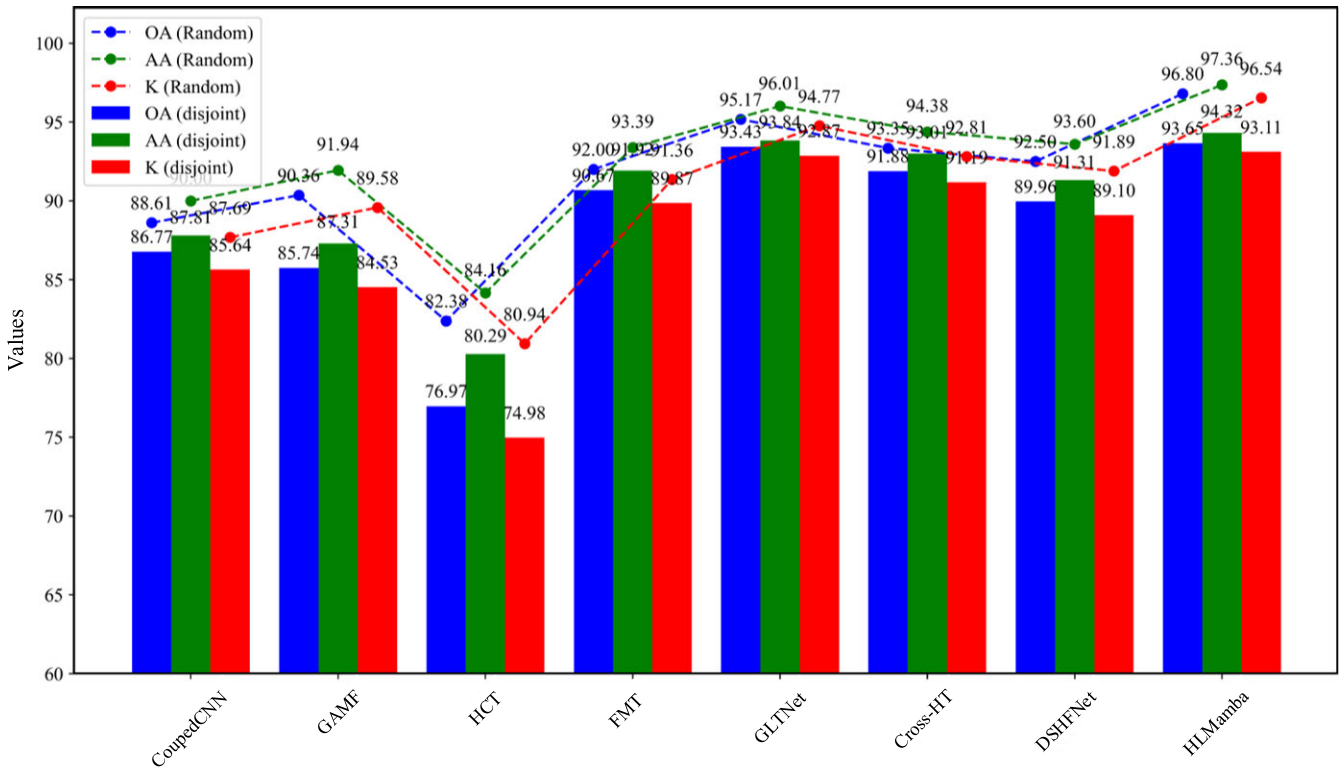

Fig. 13. Experimental results of different sampling strategies.

results are shown in Fig. 13, where the line represents the random sampling results and the bar chart represents the disjoint sampling results. In addition, blue denotes OA, green denotes AA, and red denotes K.

From Fig. 13, it is evident that the experimental results of the disjoint strategy for all methods are lower than those of random sampling. Specifically, HCT exhibited the most significant decrease, followed by GAMF, while other methods also showed a certain degree of decrease. We believe that there are two reasons that can be summarized as follows: first, HCT and GAMF exhibit strong spatial dependence, resulting in unsatisfactory results in disjoint sampling. Second, to some extent, while random sampling can enhance robustness, it may also introduce information leakage compared to disjoint sampling.

IV. Conclusion

In this article, a joint classification method based on Mamba for HSI and LiDAR data is proposed. To enrich feature diversity, edge contour data based on LiDAR is introduced, serving as input alongside HSI and LiDAR data for the feature extraction stage. Subsequently, a CNN-based MFEM is designed to extract spectral-spatial information from HSI, elevation information from LiDAR, and edge contour information. In the multimodal feature fusion stage, a low parameter and time-complexity MMFM is designed to mine intra and intermodal features. Finally, the classifier is designed with a simple linear structure. The proposed method, HLMamba, is validated on Houston, MUUFL, and Trento datasets, demonstrating its superiority and the effectiveness of edge contour information in multisource joint classification tasks. Additionally, the fusion module based on Transformer and Mamba is analyzed, revealing that: 1) Mamba requires approximately 30% of the training parameters of Transformer; 2) Mamba requires approximately 10% less runtime than Transformer; and 3) the classification accuracy obtained by the fusion module based on Mamba in the HLMamba model is better than that of Transformer. These findings suggest that in HSI-LiDAR joint classification tasks, the Mamba framework exhibits more efficient performance compared to the Transformer.

Although Transformer has been widely used in RS and achieved significant results, Mamba has shown great potential and is considered to have a promising future. Therefore, in the future, efforts will be made to fully explore the advantages of the Mamba framework and investigate further applications in multimodal RS image joint classification.

Acknowledgment

The authors would like to thank the Editor-in-Chief, the Associate Editor, and the reviewers for their insightful comments and suggestions. In addition, they would like to express their special thanks to the Organizing Committee of the 2013 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Competition for providing the hyperspectral (HS) and light detection and ranging (LiDAR) data.

References

[1] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral–spatial–temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[2] B. F. R. Davies et al., "Multi- and hyperspectral classification of soft-bottom intertidal vegetation using a spectral library for coastal biodiversity remote sensing," *Remote Sens. Environ.*, vol. 290, May 2023, Art. no. 113554.

[3] S. Kendler, I. Ron, S. Cohen, R. Raich, Z. Mano, and B. Fishbain, "Detection and identification of sub-millimeter films of organic compounds on environmental surfaces using short-wave infrared hyperspectral imaging: Algorithm development using a synthetic set of targets," *IEEE Sensors J.*, vol. 19, no. 7, pp. 2657–2664, Apr. 2019.

[4] M. Govender, K. Chetty, and H. Bulcock, "A review of hyperspectral remote sensing and its application in vegetation and water resource studies," *Water SA*, vol. 33, no. 2, pp. 145–151, Dec. 2009.

[5] P. Ghamisi et al., "The potential of machine learning for a more responsible sourcing of critical raw materials," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8971–8988, 2021.

[6] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[7] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.

[8] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.

[9] V. Laparra, J. Malo, and G. Camps-Valls, "Dimensionality reduction via regression in hyperspectral imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1026–1036, Sep. 2015.

[10] S. Zhou, M. Ke, and P. Luo, "Multi-camera transfer GAN for person re-identification," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 393–400, Feb. 2019.

[11] J. Liu, C. Gu, J. Wang, G. Youn, and J.-U. Kim, "Multi-scale multi-class conditional generative adversarial network for handwritten character generation," *J. Supercomput.*, vol. 75, no. 4, pp. 1922–1940, Apr. 2019.

[12] S. He, Z. Li, Y. Tang, Z. Liao, F. Li, and S.-J. Lim, "Parameters compressing in deep learning," *Comput., Mater. Continua*, vol. 62, no. 1, pp. 321–336, 2020.

[13] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[14] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[15] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[16] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.

[17] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[18] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.

[19] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep&dense convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 10, no. 9, p. 1454, Sep. 2018.

[20] J. Feng, H. Yu, L. Wang, X. Cao, X. Zhang, and L. Jiao, "Classification of hyperspectral images based on multiclass spatial–spectral generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5329–5343, Aug. 2019.

[21] X. Wang, K. Tan, Q. Du, Y. Chen, and P. Du, "CVA2E: A conditional variational autoencoder with an adversarial training process for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5676–5692, Aug. 2020.

[22] J. Feng et al., "Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 7, p. 1149, Apr. 2020.

[23] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, and H. Zhou, "Spatial sequential recurrent neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4141–4155, Nov. 2018.

[24] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3859–3869.

[25] F. F. Shahraki and S. Prasad, "Graph convolutional neural networks for hyperspectral data classification," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Anaheim, CA, USA, Nov. 2018, pp. 968–972.

[26] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 241–245, Feb. 2019.

[27] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[28] D. Hong et al., "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

[29] Z. Wang and M. Menenti, "Challenges and opportunities in LiDAR remote sensing," *Frontiers Remote Sens.*, vol. 2, Mar. 2021, Art. no. 641723.

[30] W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Asymmetric feature fusion network for hyperspectral and SAR image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 8057–8070, Oct. 2022.

[31] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, Sep. 2020.

[32] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018.

[33] Z. Xue, X. Yu, X. Tan, B. Liu, A. Yu, and X. Wei, "Multiscale deep learning network with self-calibrated convolution for hyperspectral and LiDAR data collaborative classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514116.

[34] H. Gao, H. Feng, Y. Zhang, S. Xu, and B. Zhang, "AMSSE-Net: Adaptive multiscale spatial–spectral enhancement network for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531317.

[35] X. Du, X. Zheng, X. Lu, and A. A. Doudkin, "Multisource remote sensing data classification with graph fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10062–10072, Dec. 2021.

[36] S. Fang, K. Li, and Z. Li, "S^2ENet: Spatial–spectral cross-modal enhancement network for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[37] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5506812.

[38] J. Cai et al., "A novel graph-attention based multimodal fusion network for joint classification of hyperspectral image and LiDAR data," *Expert Syst. Appl.*, vol. 249, Sep. 2024, Art. no. 123587.

[39] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517010.

[40] P. Ghamisi, B. Höfle, and X. X. Zhu, "Hyperspectral and LiDAR data fusion using extinction profiles and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 3011–3024, Jun. 2017.

[41] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.

[42] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[43] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515620.

[44] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.

[45] G. Zhao, Q. Ye, L. Sun, Z. Wu, C. Pan, and B. Jeon, "Joint classification of hyperspectral and LiDAR data using a hierarchical CNN and transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500716.

[46] K. Ding, T. Lu, W. Fu, S. Li, and F. Ma, "Global–local transformer network for HSI and LiDAR data joint classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5541213.

[47] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral–spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503615.

[48] S. K. Roy, A. Sukul, A. Jamali, J. M. Haut, and P. Ghamisi, "Cross hyperspectral and LiDAR attention transformer: An extended self-attention for land use and land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5512815.
[49] Y. Feng, L. Song, L. Wang, and X. Wang, "DSHFNet: Dynamic scale hierarchical fusion network based on multiattention for hyperspectral image and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5522514.
[50] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.
[51] J. X. Yang, J. Zhou, J. Wang, H. Tian, and A. W. C. Liew, "HSIMamba: Hyperpsectral imaging efficient feature learning with bidirectional state space for classification," 2024, *arXiv:2404.00272*.
[52] J. Yao, D. Hong, C. Li, and J. Chanussot, "SpectralMamba: Efficient mamba for hyperspectral image classification," 2024, *arXiv:2404.08489*.
[53] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," 2021, *arXiv:2111.00396*.

Diling Liao (Student Member, IEEE) received the B.S. degree from Zhuhai College, Jilin University, Zhuhai, China, in 2019, and the M.S. degree from Qiqihar University, Qiqihar, China, in 2023. He is currently pursuing the Ph.D. degree with Sun Yat-sen University, Shenzhen, China.

His research interests include hyperspectral image classification, deep learning, and multimodal remote sensing image processing.

Qingsong Wang received the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2011.

Since 2019, he has been working with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China, where he is mainly engaged in basic theory and engineering application research in the field of precision processing of microwave mapping. His research interests include multisystem radar positioning, image matching, and unmanned aerial vehicle (UAV) navigation.

Tao Lai received the Ph.D. degree from the National University of Defense Technology, Changsha, China, in 2010.

He is currently an Associate Professor and a Professor. His main research interests are ultrawideband synthetic aperture radar (SAR)/ISAR imaging radar system design and imaging processing, low and slow target detection phased array radar system design and signal processing, multi-input multi-output (MIMO) imaging radar signal processing, multichannel signal processing, and ground deformation monitoring SAR system design and signal processing.

Haifeng Huang (Member, IEEE) received the B.S. degree in mathematics and the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 1997 and 2005, respectively.

He is currently a Professor and the Associate Dean of the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China. His research interests include basic theory and key technology in the field of space electronics and intelligent sensing and are mainly oriented to intelligent remote sensing, surveying, oceanography, surveillance, geological hazards, and other applications.