Joint Classification of Hyperspectral and LiDAR Data Using Hierarchical Multimodal Feature Aggregation-Based Multihead Axial Attention Transformer

Fei Zhu, Cuiping Shi, *Member, IEEE*, Kaijie Shi, and Liguo Wang, *Member, IEEE*

***Abstract*— The rapid development of sensor and multimodal technology has provided more possibilities for multisource remote sensing image classification. However, some existing joint classification methods are limited to single-level feature fusion and fail to fully explore the deep correlation between cross-level features, thus limiting the effective interaction and complementarity of information between different modal data. To alleviate this issue, this article proposes a hierarchical multimodal feature aggregation-based multihead axial attention transformer (HMAT) for joint classification of hyperspectral and light detection and ranging (LiDAR) data. First, a hierarchical multimodal feature aggregation module (HMFA) is proposed to more effectively fuse spatial–spectral features of hyperspectral images (HSIs) and elevation features of LiDAR data and generate more discriminative low-dimensional feature representations. Second, a pyramid-inverted pyramid convolution module (PIP) is designed. Through the complementary feature extraction structure, PIP can more fully capture the multiscale local features in the fused feature map of hyperspectral and LiDAR data. Finally, a multihead axial attention (MHAA) component is constructed to capture information at different scales in the fused feature maps, thereby accurately modeling global dependencies. The proposed HMAT has been extensively tested on three publicly available datasets. The experimental results demonstrate that the classification performance of the proposed method outperforms that of several state-of-the-art methods.**

***Index Terms*— Axial attention, convolutional neural networks (CNNs), feature aggregation, hyperspectral, light detection and ranging (LiDAR), multimodal, transformer.**

I. Introduction

HYPERSPECTRAL images (HSIs) provide a comprehensive spectral signature for each pixel, enabling precise identification and classification of land cover types and conditions [1]. Therefore, HSIs are widely used in areas such as land monitoring [2], [3], urban planning [4], [5], medical diagnosis [6], [7], and precision agriculture [8], [9].

HSI classification (HSIC) aims to assign a specific class label to each pixel in an HSI by analyzing its spectral and spatial features [10], [11]. However, HSI data are susceptible to interference from factors such as atmospheric conditions and illumination conditions during the acquisition process, which can lead to incorrect classification. Light detection and ranging (LiDAR) data can record the elevation information of the object being measured and is less affected by atmospheric conditions; thus, it can be combined with HSI data to alleviate the limitations of using either data type alone. With the continuous breakthroughs of multimodal technology in the field of artificial intelligence, research on joint classification for HSI and LiDAR data has also become a hot topic in the field of remote sensing image classification [12], [13], [14], [15].

Early research in HSI and LiDAR data classification primarily relied on manually engineered features, which often required extensive domain knowledge and were limited in their ability to capture complex data patterns [16], [17], [18], [19]. Liao et al. [20] proposed a graph-based generalized feature fusion method. This method utilizes the original spectral information and the morphological contour features in multimodal fusion data for classification tasks. Pedergnana et al. [21] proposed a method based on extended morphological profiles, which performs the classification of HSI and LiDAR data by stacking features extracted from different modal data. Dalponte et al. [22] proposed a method based on support vector machine, aimed at classifying complex forest areas.

While these methods are computationally efficient, they exhibit a strong reliance on prior knowledge and lack the ability to adaptively capture the intrinsic characteristics of diverse data modalities.

Recent years have witnessed significant advancements in deep learning technology, particularly in the domain of multimodal artificial intelligence, including multisource remote sensing image classification [23], [24], [25], [26], [27], [28]. Convolutional neural networks (CNNs) have always been favored by researchers due to their powerful feature extraction capabilities [29], [30], [31], [32]. Hang et al. [33] proposed a method based on coupled CNNs (CoupledCNNs) by fusing HSI and LiDAR data at the feature level

Received 29 September 2024; revised 7 December 2024 and 11 January 2025; accepted 17 January 2025. Date of publication 23 January 2025; date of current version 5 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42271409 and in part by the Fundamental Research Funds in Heilongjiang Provincial Universities under Grant 145109145. *(Corresponding author: Cuiping Shi.)*

Fei Zhu and Kaijie Shi are with the Department of Communication Engineering, Qiqihar University, Qiqihar 161000, China (e-mail: 2022935750@qqhru.edu.cn; 2022910313@qqhru.edu.cn).

Cuiping Shi is with the College of Information Engineering, Huzhou University, Huzhou 313000, China (e-mail: shicuiping@zjhu.edu.cn).

Liguo Wang is with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian 116000, China (e-mail: wangliguo@hrbeu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2025.3533475

using a weight-sharing strategy and investigating multiple decision-level fusion strategies. Zhang et al. [34] proposed the interleaving perception CNN (IP-CNN), a method that combines traditional CNN with information fusion techniques to jointly classify multisource heterogeneous data. Wu et al. [35] proposed the cross-channel reconstruction network (CCR-Net), a CNN-based method that learns reconstruction strategies across different remote sensing data sources to exchange information more effectively. Fang et al. [36] proposed a method named spatial–spectral cross-modal enhancement network (S2ENet), which enhances the spatial features in HSI and spectral features in LiDAR, promoting information interaction between the two modalities. Han et al. [37] proposed a cross-modal semantic enhancement network (CMSE) to mine and fuse similar high-level semantic information and complementary discriminative information in multimodal data. Hong et al. [38] proposed an end-to-end multimodal deep learning framework called multimodal deep learning for remote sensing imagery classification (MDL-RS), which offers a foundational solution for remote sensing image classification by exploring a variety of different fusion strategies. Zhang et al. [39] proposed a three-branch CNN. This method integrates shallow and deep features using a multilevel feature fusion (MLF) module and enhances the information exchange between spatial and elevation features via a mutual guided attention (MGA) module. Wang et al. [40] proposed multiscale spatial-spectral cross-modal attention network (MS2CANet), a network that extracts features at various scales and utilizes the spatial-spectral cross-modal attention (S2CA) module to improve the interaction between different modalities.

Generative adversarial networks (GANs) significantly enhance the robustness and generalization capabilities of models through adversarial training, where a generator and a discriminator are trained in a minimax game [41], [42]. Lu et al. [43] proposed a classification method based on coupled adversarial learning (CALC), which effectively integrates high-level semantic information and complementary information from HSI and LiDAR data through coupled adversarial learning and multilevel feature processing. Yang et al. [44] proposed text-supervised multidimensional contrastive fusion network (TMCFN), a network that learns visual representations using textual information. Graph attention network (GAT) is a variant of graph neural network (GNN) that places particular emphasis on the utilization of an attention mechanism during the propagation and update of information among nodes within a graph [45], [46]. Cai et al. [47] proposed a graph attention-based multimodal fusion network (GAMF) that employs a graph attention mechanism to construct undirected graphs, thereby mitigating the long-range dependency issue in HSI and LiDAR data. Wang et al. [48] proposed a method called Markov edge decoupled fusion network (MEDFN). MEDFN optimizes graph construction through reinforcement learning, thereby improving the performance of the model.

Transformer, a deep learning model for processing sequential data, has revolutionized the field of natural language processing (NLP) [49]. Its core idea, after being refined and adapted, has also found widespread application in the field of computer vision (CV) and achieved significant success [50], [51], [52], [53]. Ding et al. [54] proposed a method called global–local transformer network (GLT), which synergizes the strengths of CNNs in capturing local spatial features and transformers in modeling long-range dependencies. Roy et al. [55] proposed a multimodal fusion transformer (MFT). MFT enhances the model's generalization capability by incorporating multimodal data as external classification tokens into the transformer encoder. Zhao et al. [56] proposed a method based on hierarchical CNN and transformer (HCT) to fuse the features of two modalities through a novel cross-token attention mechanism. Feng et al. [57] proposed a spectral–spatial–elevation fusion transformer (S2EFT) that addresses the limitations of transformers in capturing local spatial information. Roy et al. [58] proposed a cross-hyperspectral and LiDAR attention transformer (Cross-HL), which facilitates the precise exchange of information between different modalities by extending the self-attention mechanism. Song et al. [59] proposed a height information guided hierarchical fusion-and-separation network (HFSNet), which employs a dual-structure encoder to independently capture the spectral sequence information in HSI and the spatial information in LiDAR and uses height information to guide mutual learning between modalities. Zeng et al. [60] proposed a cross-modal hierarchical frequency fusion network (HFNet), which seeks to mitigate the discrepancies between different modalities through fusion in the frequency domain. Shi et al. [61] proposed a gated cross aggregation network (GCA-Net) that jointly embeds LiDAR elevation features within an HSI encoder to achieve better cross-modal alignment. Qu et al. [62] proposed a semi-supervised classification method called shared-private decoupling-based multilevel feature alignment semisupervised learning (SASS), which aims to utilize both labeled and unlabeled data to improve the classification performance. Dong et al. [63] proposed a method called contrastive constrained cross-scene model informed interpretable classification strategy (C3MI-C) for unsupervised cross-scene classification of HSI and LiDAR data. This method showed excellent performance by separating the classification task from the domain adaptation task and optimizing the alignment of features from the two domains.

Although the above methods have achieved good classification performance, their limitations still need to be further studied.

1) The existing methods mainly focus on single-level feature fusion or serial MLF, neglecting the potential benefits of deeper and cross-level feature interactions. This limitation hampers the effective integration of information from diverse data sources.
2) While the multihead self-attention mechanism (MHSA) is effective in capturing sequential dependencies, its individual attention heads lack effective interaction, limiting the model's ability to model multiscale features.

To alleviate this issue, a hierarchical multimodal feature aggregation-based multihead axial attention transformer (HMAT) is proposed for joint classification of HSI and LiDAR data. First, a hierarchical multimodal feature aggregation module (HMFA) is proposed to effectively fuse spatial–spectral features and elevation features. This module

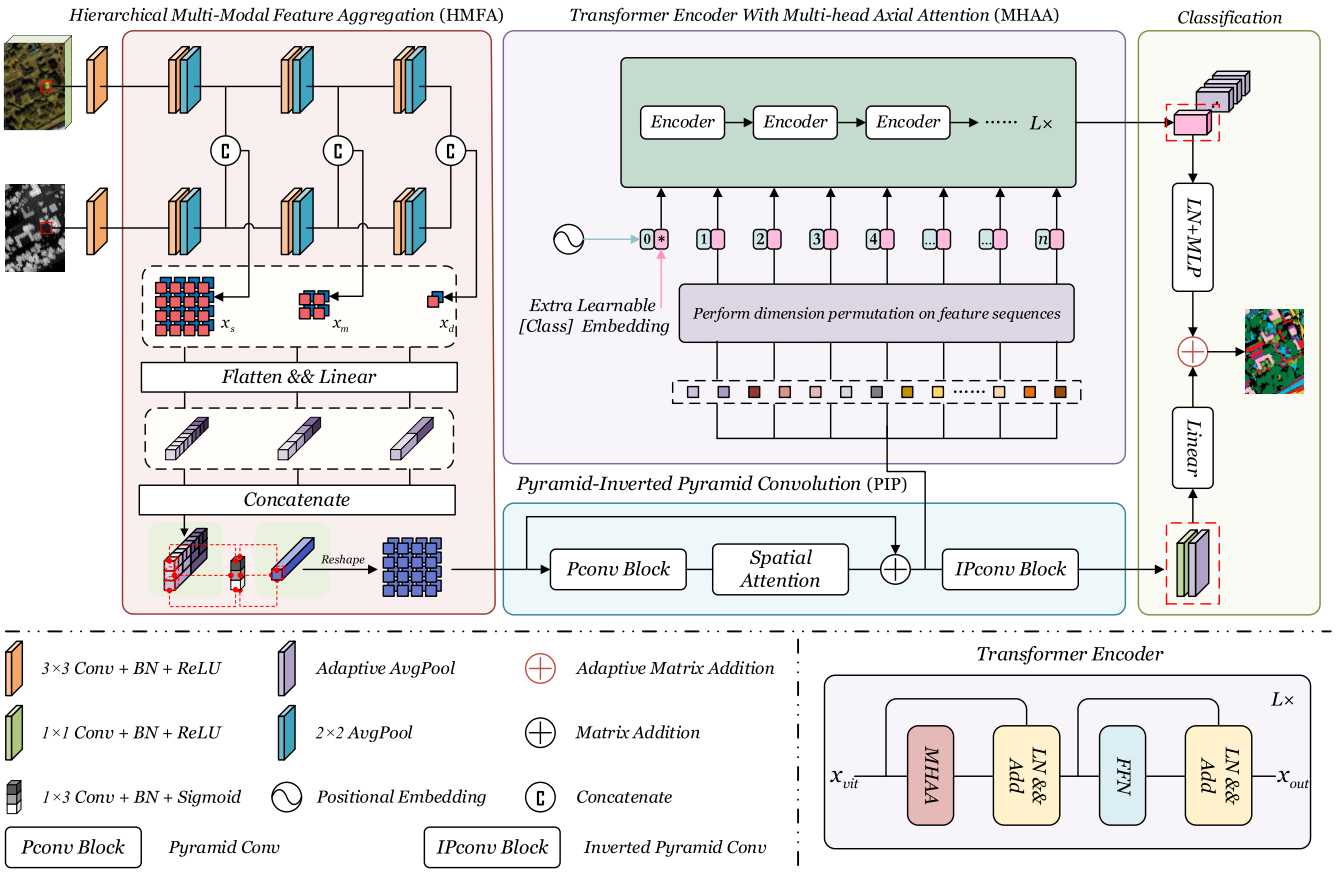

Fig. 1. Overall structure of the proposed HMAT. First, the HMFA module combines features from different levels and modalities. Next, the PIP module extracts local features and generates intermediate features for global processing. Then, the MHAA module captures global dependencies. Finally, the local and global features are fused, and the classification results are predicted.

generates more discriminative low-dimensional feature representations through cross-layer feature interaction and feature fusion from shallow to deep. Second, a pyramid-inverted pyramid convolution module (PIP) is designed. This module captures the multiscale local features in the fused feature map of hyperspectral and LiDAR data more fully through a complementary feature extraction structure, thereby further improving the performance of the model. Finally, a multi-head axial attention (MHAA) component is constructed. This module considers the multimodal information in the fused data from different scales, thereby more accurately modeling global dependencies. The proposed HMAT has been thoroughly experimentally validated on three publicly available datasets. The experimental results show that compared with some current state-of-the-art methods, the proposed method has better classification performance.

The main contributions of this article are summarized as follows.

1) An HMFA module is proposed. This module generates more discriminative low-dimensional feature representations through cross-layer feature interaction and feature fusion from shallow to deep, thereby more effectively fusing spatial–spectral features and elevation features.
2) A PIP module is designed. This module more fully captures local features in the fusion feature map of hyperspectral and LiDAR data through a complementary feature extraction structure.
3) An MHAA component is constructed. MHAA considers the global contextual information in the fused feature map of hyperspectral and LiDAR from different scales, thereby more accurately modeling global dependencies.

The remainder of this article is organized as follows. Section II presents the proposed HMAT in detail. Section III describes the experimental setup, including datasets, parameter settings, and evaluation metrics, followed by an evaluation of the proposed method. Section IV concludes this article and outlines potential future research directions.

II. Methodology

The HSI data are denoted as $\boldsymbol{X}_{\text{hsi}} \in \mathbb{R}^{h\times w\times b}$, and the LiDAR data are denoted as $\boldsymbol{X}_{\text{ela}} \in \mathbb{R}^{h\times w\times l}$, where h and w represent the height and width of the data, respectively, and b and l represent the number of bands for the HSI and LiDAR, respectively. The patches extracted from the two modal data mentioned above are denoted as $\boldsymbol{X}_1 \in \mathbb{R}^{s\times s\times b}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{s\times s\times l}$. Here, $s \times s$ represents the patch size.

A. Overall Structure

The overall structure of the proposed HMAT is shown in Fig. 1, which mainly consists of five parts: channel modulation, hierarchical multimodal feature aggregation, weighted

local feature extraction, global feature extraction, and classification stage. Specifically, since HSI data and LiDAR data have different feature representation spaces, especially with significant differences in data dimensions, directly feeding these raw data into the network would make it difficult for the network to capture effective common features, which could adversely affect subsequent processing. Therefore, in the channel modulation part, 3×3 convolution kernels are used to unify the number of channels for different modal data. In addition, batch normalization (BN) operations are applied to standardize these data, effectively eliminating scale and distribution differences caused by different data sources. The results are denoted as $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{s\times s\times c}$. This process can be represented as

$$\boldsymbol{A}, \boldsymbol{B} \leftarrow F^*_{3\times3}(\boldsymbol{X}_1), F^*_{3\times3}(\boldsymbol{X}_2) \tag{1}$$

$$F^*_{3\times3}(\boldsymbol{X}) = \delta(\mathrm{BN}(\boldsymbol{X} * \mathbf{W} + b)) \tag{2}$$

where $F^*_{3\times3}$ represents a 2-D convolutional block with a kernel size of 3, δ represents the rectified linear unit (ReLU) activation function, $*$ represents the convolution operator, $\mathbf{W}$ represents the weights of the convolution kernel, b represents the bias, and c represents the number of modulated channels.

In the hierarchical multimodal feature aggregation part, we employ the HMFA module to extract hierarchical features from multimodal data and aggregate these features into a low-dimensional feature representation that contains information from different levels.

Subsequently, we employ the PIP module to extract multiscale features from the low-dimensional features processed by HMFA and assign different weights to different feature regions through a spatial attention mechanism. The outputs of the PIP module are divided into two branches: one for the classification task and the other for extracting global features. Then, in the global feature extraction part, through the proposed MHAA component, the transformer can extract information from the input data at different scales, thereby outputting more representative global features.

In the final classification stage, we unify local and global features into the same dimension and adopt an adaptive fusion strategy to assign dynamic weights to the two types of features, thereby achieving effective fusion and improving classification accuracy. This process can be represented as

$$\boldsymbol{G} = \mathrm{LN}\big(\mathrm{MLP}\big(\boldsymbol{T}^{\mathrm{cls}}\big)\big) \tag{3}$$

$$\boldsymbol{L} = \mathrm{Linear}\Big(F'''_{\mathrm{avg}}\big(F^*_{1\times1}(\boldsymbol{U})\big)\Big) \tag{4}$$

$$\boldsymbol{P} = \eta \times \boldsymbol{G} + (1-\eta) \times \boldsymbol{L} \tag{5}$$

where $F^*_{1\times1}$ represents a 2-D convolutional block with a kernel size of 1, F'''_{avg} represents an adaptive average pooling layer, and η represents the dynamically adjusted weighting coefficient. $\boldsymbol{G}$ represents global features, $\boldsymbol{L}$ represents local features, and $\boldsymbol{P}$ represents the fusion features to be predicted. LN represents the layer normalization and MLP represents the multilayer perceptron.

B. HMFA Module

Most existing joint classification methods are limited to single-layer or serial feature fusion, which fails to fully explore the potential correlation of cross-layer features, thus hindering the efficient fusion of multimodal information. To alleviate this issue, an HMFA is designed, with its detailed structure shown in Fig. 1.

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be the input of this section. The 3×3 convolution blocks are first adopted to extract shallow features from different modal data. Subsequently, average pooling is introduced to effectively reduce the spatial dimensions of the features to $s/2 \times s/2$ (originally $s \times s$), achieving spatial compression and initial feature aggregation. Finally, the features from different modalities are stacked together using a concatenation operation. On this basis, the above process will be repeated at a deeper level. Specifically, in the middle-level feature extraction stage, the window size after the pooling operation is reduced to $s/4 \times s/4$. In the deep feature extraction stage, the window size is adjusted to $s/8 \times s/8$. This process can be represented as

$$\boldsymbol{A}_i, \boldsymbol{B}_i \leftarrow F_{\mathrm{avg}}\big(F^*_{3\times3}(\boldsymbol{A})\big), \quad F_{\mathrm{avg}}\big(F^*_{3\times3}(\boldsymbol{B})\big) \tag{6}$$

$$\boldsymbol{C}_i = \mathrm{Concat}(\boldsymbol{A}_i, \boldsymbol{B}_i) \tag{7}$$

where F_{avg} represents the average pooling layer. The stacked results are denoted as $\boldsymbol{C}_i$, where $0 \le i \le 2$. Through hierarchical feature extraction, three levels of features were captured: shallow features $\boldsymbol{C}_0 \in \mathbb{R}^{s/2\times s/2\times d}$, middle-level features $\boldsymbol{C}_1 \in \mathbb{R}^{s/4\times s/4\times d}$, and deep features $\boldsymbol{C}_2 \in \mathbb{R}^{s/8\times s/8\times d}$, $d = 2 \times c$.

To achieve deeper feature fusion, considering the inconsistency of feature dimensions at different levels, we first vectorize the features and then project them into a unified feature space through a fully connected layer. This process achieves the initial fusion of the two modal data at the feature level. The results are denoted as $\bar{\boldsymbol{C}}_i \in \mathbb{R}^{n\times d}$. To achieve deeper feature fusion, we stack the feature vectors of each level along the channel dimension to form a new 3-D tensor $\boldsymbol{D} \in \mathbb{R}^{1\times n\times d}$. Subsequently, we convolve $\boldsymbol{D}$ with a 1×3 convolution kernel to capture the local correlation between features at different levels. This process fuses multiscale features in a low-dimensional space to obtain a more expressive feature representation. Finally, we reshape the convolution output into a new tensor $\boldsymbol{E} \in \mathbb{R}^{s\times s\times d}$. This process can be represented as

$$\bar{\boldsymbol{C}}_i = \mathrm{Linear}(\mathrm{Flatten}(\boldsymbol{C}_i)) \tag{8}$$

$$\boldsymbol{D} = \mathrm{Concat}\big(\bar{\boldsymbol{C}}_0, \bar{\boldsymbol{C}}_1, \bar{\boldsymbol{C}}_2\big) \tag{9}$$

$$\boldsymbol{E} = \mathrm{Re\ shape}\big(F^*_{1\times3}(\boldsymbol{D})\big) \tag{10}$$

$$F^*_{1\times3}(\boldsymbol{D}) = \sigma(F_{\mathrm{BN}}(\boldsymbol{D} * \mathbf{W} + b)) \tag{11}$$

where $F^*_{1\times3}$ represents a 2-D convolution block with a kernel shape of 1×3 and σ represents the sigmoid activation function.

C. PIP Module

The self-attention mechanism in transformer makes the model with a global receptive field, enabling it to capture long-range dependencies between any two tokens in the sequence. However, this global perspective can sometimes overshadow the importance of local information, limiting the model's ability to extract fine-grained features. To mitigate this

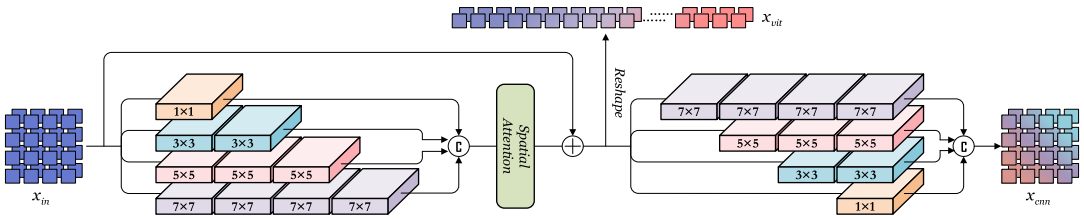

Fig. 2. Structure of PIP. The features are initially processed using pyramid convolution and position attention. Then, the processed features are divided into two branches: one for global features and the other for local features using inverted pyramid convolution.

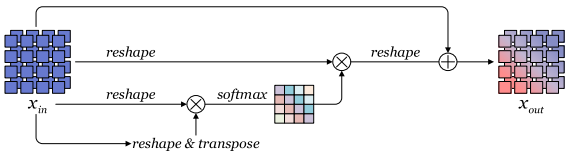

Fig. 3. Structure of positional attention.

issue, a PIP is proposed, with its detailed structure illustrated in Fig. 2.

Let $\boldsymbol{E}$ be the input of this section. First, the pyramid convolution block is employed to extract multiscale features from the $\boldsymbol{E}$. This block consists of four convolutional layers, each layer using grouped convolution, and the size of the convolution kernel is positively correlated with the number of groups. This design enables the model to efficiently extract rich features at different scales. The output channels of each branch are 1/4 of the input channels, and the output of multiscale feature fusion is finally obtained through feature concatenation. The output is denoted as $\boldsymbol{E}' \in \mathbb{R}^{s\times s\times d}$. This process can be represented as

$$\boldsymbol{E}' = \text{Concat}\big(F^*_{1\times1}(\boldsymbol{E}), F^*_{3\times3}(\boldsymbol{E}), F^*_{5\times5}(\boldsymbol{E}), F^*_{7\times7}(\boldsymbol{E})\big) \quad (12)$$

where $F^*_{1\times1}$ represents a convolution layer with a kernel size of 1, similar to others.

Subsequently, a positional attention module is employed to weigh the local features in the feature map to distinguish the importance of different positions. Its detailed structure is shown in Fig. 3. Specifically, convolution is first employed to perform a nonlinear transformation on the output $\boldsymbol{E}'$, resulting in two outputs $\boldsymbol{Q} \in \mathbb{R}^{s\times s\times d}$, $\boldsymbol{K} \in \mathbb{R}^{s\times s\times d}$, and $\boldsymbol{V} \in \mathbb{R}^{s\times s\times d}$. Afterward, the shapes of these outputs are reshaped into $\mathbb{R}^{n\times d}$, where n is the number of elements in the spatial dimension. Next, the positional attention map is calculated by multiplying $\boldsymbol{Q}$ and the transposition of $\boldsymbol{K}$, with its score values ranging from (0, 1). Afterward, perform matrix multiplication on the attention maps and $\boldsymbol{V}$ and fuse the features from the original input. This process can be formulated as

$$\mathbf{M}_{i,j} = \frac{\exp\big(\boldsymbol{Q}_i \times \boldsymbol{K}_j^{\mathrm{T}}\big)}{\sum_{i=1}^{N}\exp\big(\boldsymbol{Q}_i \times \boldsymbol{K}_j^{\mathrm{T}}\big)} \quad (13)$$

$$\boldsymbol{E}''_j = a \cdot \sum_{i=1}^{N}\big(\mathbf{M}_{ij} \times V_j\big) + E'_j \quad (14)$$

where a represents the dynamically adjusted weighting coefficient. The results are denoted as $\boldsymbol{E}'' \in \mathbb{R}^{s\times s\times d}$.

Then, the weighted feature $\boldsymbol{E}''$ is fused with the original feature $\boldsymbol{E}$ and the result is divided into two branches.

One branch reshapes the result into sequence for subsequent global feature extraction. The sequence is denoted as $\boldsymbol{T} \in \mathbb{R}^{n\times d}$.

The other branch sends the feature map into an inverted pyramid convolution block. This block adopts the reverse design of pyramid convolution, integrates the previously extracted fine-grained features through large-size convolution kernels, and further refines the coarse-grained features using small-size convolution kernels. This design can effectively fuse multiscale features and enrich feature representation. The result is denoted as $\boldsymbol{U} \in \mathbb{R}^{s\times s\times d}$. This process can be represented as

$$\boldsymbol{T}, \boldsymbol{E}^o \leftarrow \text{Reshape}\big(\boldsymbol{E}'' + \boldsymbol{E}\big), \boldsymbol{E}'' + \boldsymbol{E} \quad (15)$$

$$\boldsymbol{U} = \text{Concat}\big(F^*_{7\times7}(\boldsymbol{E}^o), F^*_{5\times5}(\boldsymbol{E}^o), F^*_{3\times3}(\boldsymbol{E}^o), F^*_{1\times1}(\boldsymbol{E}^o)\big) \quad (16)$$

where $\boldsymbol{E}^o$ represents the fusion result of the weighted feature $\boldsymbol{E}''$ and the original feature $\boldsymbol{E}$.

D. MHAA Mechanism

Although MHSA performs well in capturing sequential dependencies, the lack of effective interaction between its attention heads limits the model's ability to model multiscale features. To mitigate this issue, an MHAA mechanism is designed to replace the MHSA in transformer, and its detailed structure is shown in Fig. 4.

This section takes the tensor $\boldsymbol{T}$ in PIP as input, permutes its dimensions, and fuses the learnable class label and position encoding to generate the final input sequence. The result is denoted as $\boldsymbol{T}_{\text{in}} \in \mathbb{R}^{(d+1)\times n}$. Then, through the linear transformation with weight sharing, the row vector and column vector are projected to different subspaces. The two vectors are denoted as $\boldsymbol{T}_r, \boldsymbol{T}_c \in \mathbb{R}^{(d+1)\times s\times s}$. This design aims to promote the correlation between row and column vectors. Meanwhile, the value vectors are processed through a separate linear layer to preserve more original information, denoted as $\boldsymbol{T}_v \in \mathbb{R}^{(d+1)\times s\times s}$. This process can be expressed as

$$\boldsymbol{T}_{\text{in}} = \text{Concat}\big(\boldsymbol{T}_0^{\text{cls}}, \boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_{n+1}\big) + \text{PE} \quad (17)$$

$$\begin{cases} \boldsymbol{T}_r, \boldsymbol{T}_c & \leftarrow \text{Reshape}(\text{Linear}(\boldsymbol{T}_{\text{in}})) \\ \boldsymbol{T}_v & \leftarrow \text{Reshape}(\text{Linear}(\boldsymbol{T}_{\text{in}})) \end{cases} \quad (18)$$

where PE represents the positional encode and $\boldsymbol{T}_0^{\text{cls}}$ represents the class token.

Then, a row pooling operation is implemented on $\boldsymbol{T}_r$, aiming to capture the key information of the vector in the

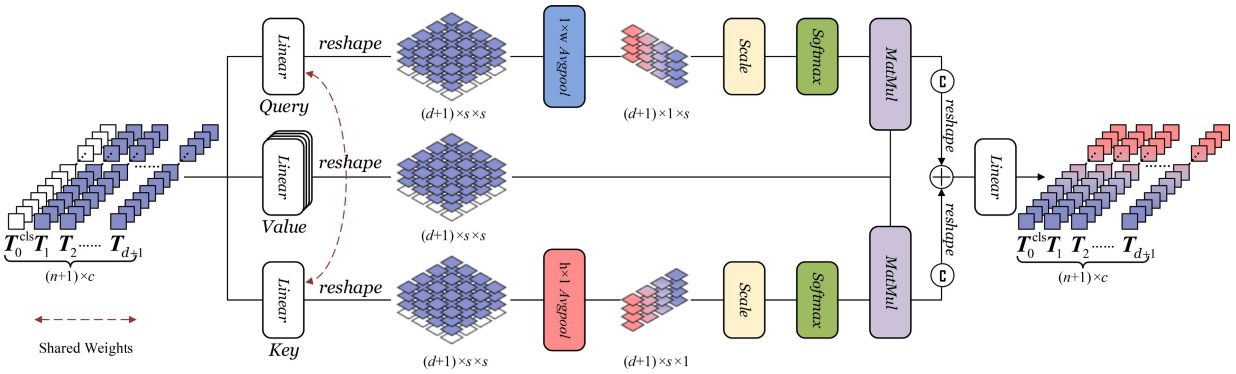

Fig. 4. Structure of MHAA. $1 \times w$ represents horizontal pooling and $h \times 1$ represents vertical pooling.

Y-direction. Correspondingly, on $\boldsymbol{T}_c$, column pooling operation is employed to extract important features in the X-direction. Next, the pooled results are utilized to construct an attention map and convert it into normalized attention scores through scaling operation and the softmax function. The two attention maps of different scales are multiplied with the value vector, resulting in the weighted feature representations of each branch. This process can be represented as

$$\begin{aligned} \mathrm{AA}^r &= \mathrm{Attention}(\boldsymbol{T}_r, \boldsymbol{T}_v) \\ &= \mathrm{Softmax}\left(\frac{F'_{\mathrm{avg}}(\mathrm{Reshape}(\boldsymbol{T}_r))}{\sqrt{d_r}}\right) \times \boldsymbol{T}_v \qquad (19) \\ \mathrm{AA}^c &= \mathrm{Attention}(\boldsymbol{T}_c, \boldsymbol{T}_v) \\ &= \mathrm{Softmax}\left(\frac{F''_{\mathrm{avg}}(\mathrm{Reshape}(\boldsymbol{T}_c))}{\sqrt{d_c}}\right) \times \boldsymbol{T}_v \qquad (20) \end{aligned}$$

where F'_{avg} represents the average pooling in the Y-direction, d_r represents the dimension of $\boldsymbol{T}_r$, F''_{avg} represents the average pooling in the X-direction, d_k represents the dimension of $\boldsymbol{T}_c$, and AA^r and AA^c represent the weighted result of a single head. Subsequently, the multihead attention results on these two branches are merged and fused into a new feature representation, denoted as $\boldsymbol{T}'$

$$\begin{aligned} \boldsymbol{T}'_r &= \mathrm{MultiHead}(\boldsymbol{T}_r, \boldsymbol{T}_v) \\ &= \mathrm{Concat}\left(\mathrm{AA}^r_1, \mathrm{AA}^r_2, \ldots, \mathrm{AA}^r_{n+1}\right) \cdot \mathbf{W} \qquad (21) \\ \boldsymbol{T}'_c &= \mathrm{MultiHead}(\boldsymbol{T}_c, \boldsymbol{T}_v) \\ &= \mathrm{Concat}\left(\mathrm{AA}^c_1, \mathrm{AA}^c_2, \ldots, \mathrm{AA}^c_{n+1}\right) \cdot \mathbf{W} \qquad (22) \\ \boldsymbol{T}' &= \mathrm{Reshape}\left(\boldsymbol{T}'_r + \boldsymbol{T}'_c\right) \qquad (23) \end{aligned}$$

where $\mathbf{W}$ represents the weight matrix. $\boldsymbol{T}'$ undergoes processing through modules such as LN, MLP, and residual connections to complete a global feature processing, as shown in Fig. 1. After repeating this encoder block multiple times, we extract $\boldsymbol{T}^{\mathrm{cls}}$ for final classification.

III. Experimental Results and Analysis

A. Dataset Description

1) Houston 2013 [64]: This dataset comes from a mapping project carried out by the National Airborne Laser Mapping Center of USA in June 2012. This project used a compact airborne spectral imager (CASI) sensor to collect data in the University of Houston campus and its surrounding urban areas. Among them, HSI contains 144 bands, with a wavelength range from 0.38 to 1.05 μm, and has a spatial resolution of 2.5 m, with dimensions of 349 × 1905 pixels. The LiDAR data have the same size and resolution as the HSI data but contain only one band, providing elevation information of the ground structure. The entire dataset contains 15 029 field-verified samples, which are classified into 15 different categories, aiming to promote the research on object recognition and classification in complex urban environments. Fig. 5(a)–(c), respectively, shows the pseudo-color image of HSI, the digital surface model (DSM) image of LiDAR, and the ground truth map on this dataset.

2) Trento [33]: This dataset comes from a rural area in southern Trento, Italy. HSI data are collected by the airborne hyperspectral imaging system (AISA) Eagle sensor, which has 63 spectral bands with a wavelength range of 0.42–0.99 μm. Correspondingly, LiDAR data are obtained by the Optech airborne laser terrain mapping (ALTM) 3100EA sensor. The size of the dataset is 166 × 600 pixels, with a spatial resolution of 1 m. The entire dataset contains 30 214 samples, which are classified into six categories. Fig. 6(a)–(c) shows the pseudo-color image of HSI, the DSM image of LiDAR, and the ground truth map on the dataset, respectively.

3) MUUFL [65], [66]: This dataset was collected in November 2010 using CASI-1500 sensors at the University of Southern Mississippi's Gulfport campus in Long Beach, MS, USA. Among them, HSI data cover 72 spectral bands from 0.38 to 1.05 μm, but eight bands were removed due to noise issues. LiDAR data consist of two bands: one is a digital elevation model (DEM) image and the other is an intensity image. The dataset has a size of 325 × 220 pixels and contains a total of 53 687 samples, covering 11 different types of land cover. Fig. 7(a)–(c) shows the pseudo-color image of HSI, the DSM image of LiDAR, and the ground truth map on the dataset, respectively.

The land cover category names, the number of training set samples, and the number of test set samples for the above three datasets are detailed in Table I.

B. Experimental Configuration

1) Hardware Configuration: The proposed HMAT is implemented in the PyTorch deep learning framework under version 1.10.1. To ensure stability and efficiency, all experiments are

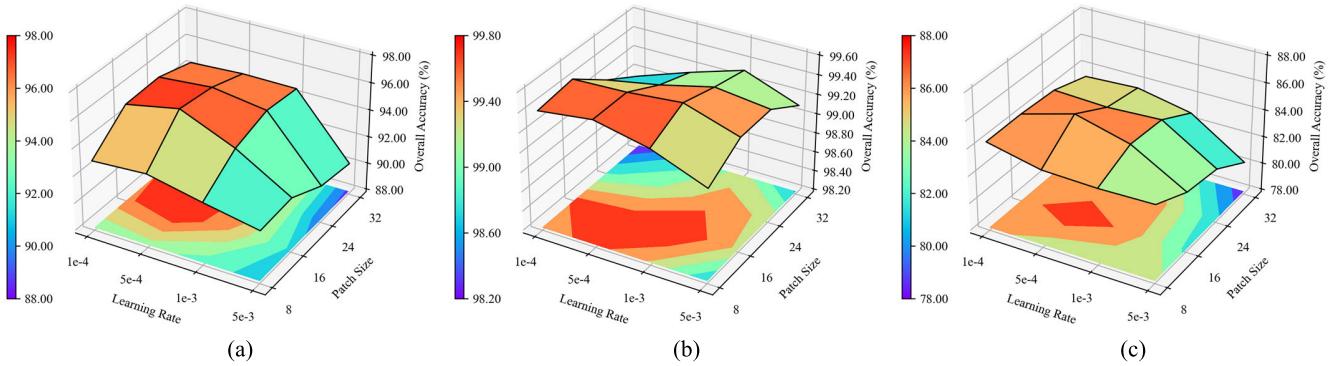

Fig. 5. Parameter analysis for HMFA. (a)–(c) Houston 2013, Trento, and MUUFL, respectively.

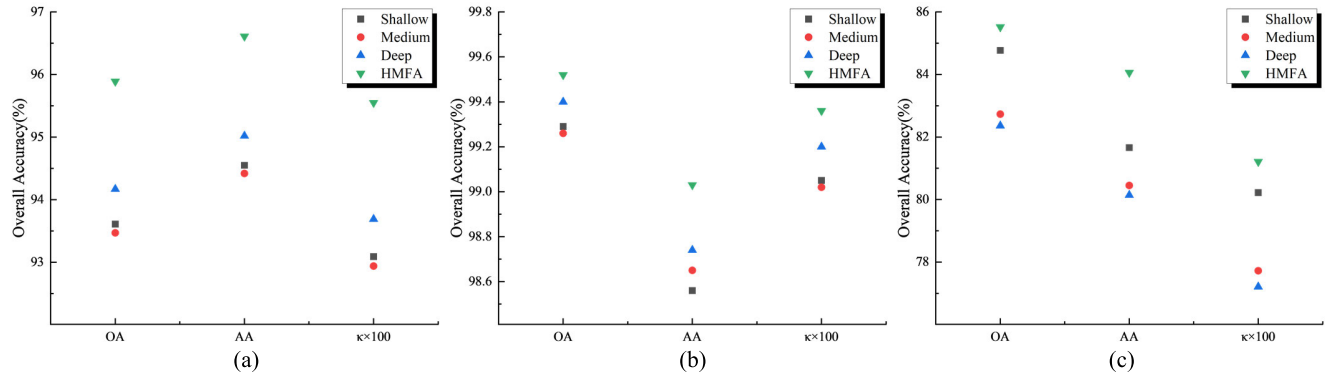

Fig. 6. Ablation for HMFA. (a)–(c) Houston 2013, Trento, and MUUFL, respectively.

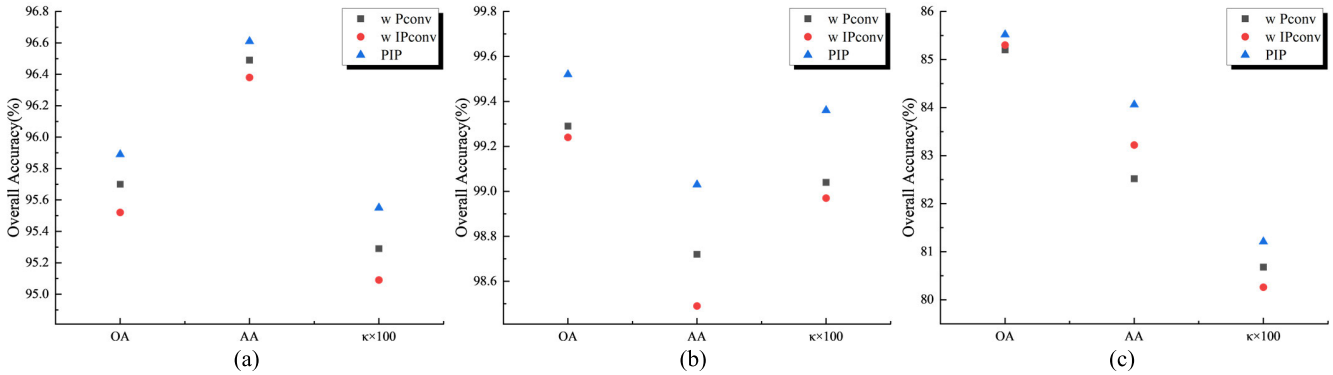

Fig. 7. Ablation for PIP. (a)–(c) Houston 2013, Trento, and MUUFL, respectively.

verified on an NVIDIA GeForce RTX 3090 with 24 GB of video memory.

2) Evaluation Indicator: To comprehensively evaluate the performance of all methods, some evaluation indicators were selected, including overall accuracy (OA), average accuracy (AA), kappa coefficient ($\kappa \times 100$), parameters (Params), and floating-point operations (FLOPs).

3) Parameter Analysis: To ensure the reliability of experimental results, all comparative methods were rigorously evaluated using the parameter settings provided in their original articles. Furthermore, to mitigate the impact of randomness, all experimental results were averaged over ten independent experiments. The proposed method employs the Adam optimizer for network training, with a batch size of 64 and a training epoch of 300. Learning rate and patch size are crucial factors influencing model performance. As shown in Fig. 5, the learning rate decay coefficient is set to 0.9. The proposed method achieves optimal classification performance on the Houston and muufl gulfport (MUUFL) datasets with a patch size of 16 and a learning rate of $5e^{-4}$. Also, a larger learning rate of $1e^{-3}$ yields better results on the Trento dataset. Table II presents the optimal kernel sizes for each convolution layer within the PIP module across the three datasets. The transformer model uses a five-layer encoder, and each self-attention layer contains four attention heads.

C. Ablation Experiments for HMFA

To validate the effectiveness of the multilevel feature extraction and hierarchical multimodal feature fusion strategies of

TABLE I
TRAINING AND TEST SAMPLE NUMBERS FOR HOUSTON 2013, TRENTO, AND MUUFL

No	Houston 2013			Trento			MUUFL		
	Class Name	Training	Test	Class Name	Training	Test	Class Name	Training	Test
1	Healthy Grass	20	1231	Apple Trees	20	4014	Trees	20	23226
2	Stressed Grass	20	1234	Buildings	20	2883	Mostly Grass	20	4250
3	Synthetic Grass	20	677	Ground	20	459	Mixed Ground Surface	20	6862
4	Trees	20	1224	Woods	20	9103	Dirt and Sand	20	1806
5	Soil	20	1222	Vineyard	20	10481	Road	20	6667
6	Water	20	305	Roads	20	3154	Water	20	446
7	Residential	20	1248				Buildings Shadow	20	2213
8	Commercial	20	1224				Buildings	20	6220
9	Road	20	1232				Sidewalk	20	1365
10	Highway	20	1207				Yellow Curb	20	163
11	Railway	20	1215				Cloth Panels	20	249
12	Parking Lot 1	20	1213						
13	Parking Lot 2	20	449						
14	Tennis Court	20	408						
15	Running Track	20	640						
-	Total	300	14729	Total	120	30094	Total	220	53467

TABLE II
COMPARISON OF CONVOLUTION KERNELS OF DIFFERENT SIZES IN PIP

Case	Houston 2013			Trento			MUUFL		
	OA	AA	κ×100	OA	AA	κ×100	OA	AA	κ×100
1,3,3,7	95.07	95.89	94.67	99.19	98.73	98.92	84.77	84.32	80.41
1,3,5,7	94.75	95.58	94.32	**99.52**	**99.03**	**99.36**	84.46	81.36	79.82
3,5,5,7	95.02	95.74	94.68	99.20	98.74	98.94	**85.52**	**84.06**	**81.21**
3,5,7,9	**95.89**	**96.42**	**95.55**	98.67	97.94	98.23	85.13	82.22	80.70
3,7,7,9	94.72	95.64	94.29	98.97	98.38	98.63	83.52	83.51	78.85

the proposed HMFA module, in this section, HMFA is compared with baseline models that utilize single-level features. These baseline models fuse shallow, middle-level, or deep features of different modalities through concatenation and linear mapping.

The experimental results (Fig. 6) indicate that while deep features exhibit superior representational ability on the Houston and Trento datasets, shallow features prove more advantageous on the MUUFL dataset. Nonetheless, HMFA consistently outperforms the baseline models on all datasets, which proves the effectiveness of the proposed multilevel feature extraction and hierarchical multimodal feature fusion strategies in enhancing model representational ability.

D. Ablation Experiments for PIP and MHAA

To evaluate the effectiveness of the PIP module and the MHAA module, in this section, HMFA is adopted as the baseline model. By introducing the PIP module and the MHAA module one by one, a series of ablation experiments are constructed.

1) Ablation Experiments for PIP: The proposed PIP module employs a multiscale and complementary convolution structure to effectively extract local features. As shown in Case 1 of Table III, the model's performance is lowest when only the HMFA module is used. The introduction of the PIP module leads to a significant performance improvement compared to Case 1. Specifically, the OA on the Houston, Trento, and MUUFL datasets increases by 1.52%, 0.16%, and 2.74%, respectively. In addition, as illustrated in Fig. 7, independent experiments were conducted on various components of the PIP module. The experimental results consistently demonstrate that the complete PIP module outperforms only utilizing either pyramid convolution or inverted pyramid convolution alone. These experimental results fully prove the effectiveness of the PIP module.

2) Ablation Experiments for MHAA: By introducing the axial pooling attention mechanism, the MHAA module significantly enhances the model's ability to capture global contextual information. The experimental results show that after introducing the MHAA module based on the baseline model, the model has achieved significant performance

TABLE III
ABLATION EXPERIMENTS

Case.	Component.			Dataset.		
	HMFA	PIP	MHAA	Houston 2013	Trento	MUUFL
1	✔	✘	✘	93.50	98.99	82.10
2	✔	✔	✘	95.02	99.15	84.84
3	✔	✘	✔	95.40	99.22	85.28
4	✔	✔	✔	**95.89**	**99.52**	**85.52**

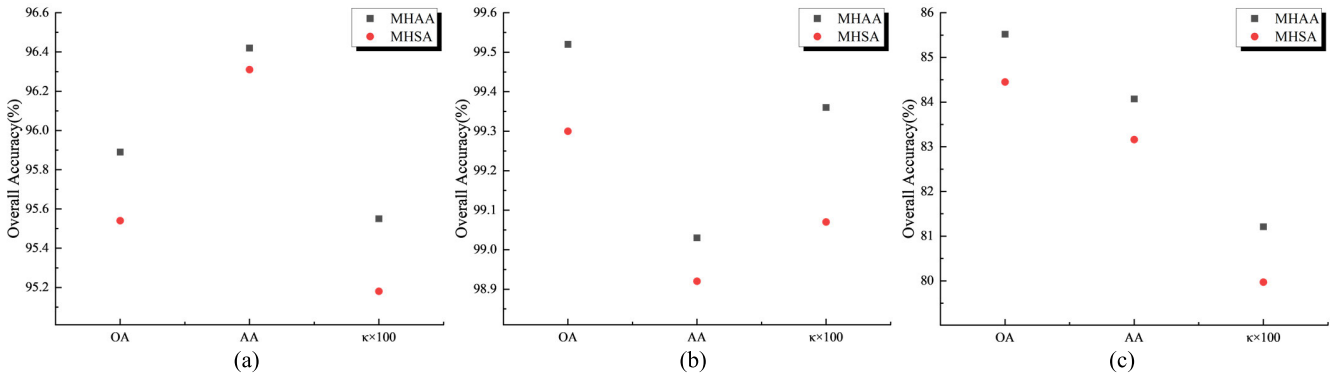

Fig. 8. Comparison of MHAA and MHSA. (a)–(c) Houston 2013, Trento, and MUUFL, respectively.

improvements on multiple datasets. As shown in Case 3 of Table III, on the Houston, Trento, and MUUFL datasets, the classification accuracy of the model increased by 1.90%, 0.23%, and 3.18%, respectively. Furthermore, when the HMFA, PIP, and MHAA modules are combined into a complete model (Case 4), the model performance reaches the best, indicating the existence of collaboration between the three modules. Moreover, compared with the traditional MHSA module, the MHAA module shows better performance on all datasets (Fig. 8). This result further demonstrates the effectiveness of the MHAA module in improving model performance, as well as its unique advantages in global context modeling.

E. Quantitative Experiments

To evaluate the performance of the proposed method, this section selects nine state-of-the-art joint classification methods such as CoupledCNN [31], CALC [40], S2ENet [34], GLT [49], MFT [50], HCT [51], S2EFT [52], GAMF [43], and CrossHL [53] for comparison. These methods cover a variety of deep learning models such as CNN, adversarial learning, GAT, and transformer fusion. Notably, CALC, GLT, and the proposed method all adopt an MLF strategy. Experiments were conducted on three datasets: Houston, Trento, and MUUFL. The classification performance of each method was evaluated using three metrics: OA, AA, and $\kappa \times 100$. The experimental results are presented in Tables IV–VI and Figs. 9–11.

While CoupledCNN fuses multimodal data at both the feature level and the decision level, its relatively simple architecture, consisting solely of convolutional and fully connected layers, limits its ability to capture and exploit global feature dependencies, thereby hindering overall classification performance. S2EFT and MFT directly fuse multimodal data at the pixel level, without adequately addressing the intrinsic differences between modalities. Although MFT mitigates the issue of data imbalance through differential handling of different modalities, both methods still exhibit limitations in classification performance due to their relatively simple network architectures. S2ENet, while enhancing multimodal interaction through different modal processing modules, still suffers from limitations in classification performance due to its relatively basic network architecture. HCT achieves relatively good classification performance by converting multimodal data into unified low-dimensional representations through a feature labeling module. GAMF encodes multimodal data into graph data and utilizes a graph attention mechanism for classification.

The remaining methods, including CALC, GLT, and others, all leverage a multilevel feature processing strategy. CALC introduces coupled adversarial learning to effectively extract high-level semantic information from multimodal data, leading to superior performance compared to most baseline methods. GLT enhances the model's predictive ability through multiscale spatial feature processing and unsupervised loss reconstruction.

Based on the proposed multilevel feature processing framework, HMAT further incorporates the PIP module to refine local features and the MHAA module to enhance global context understanding. As a result, HMAT achieves optimal performance on all three datasets, demonstrating the effectiveness of the proposed approach. The standard deviations in Tables IV–VI also reflect that the proposed method has stronger robustness.

TABLE IV
CLASSIFICATION PERFORMANCE OBTAINED BY DIFFERENT METHODS FOR THE HOUSTON 2013 DATASET (OPTIMAL RESULTS ARE BOLDED)

No.	CoupledCNN	CALC	S2ENet	GLT	MFT	HCT	S2EFT	GAMF	CrossHL	Proposed
1	89.89±3.56	88.89±5.16	96.18±5.53	87.24±4.49	91.43±6.55	96.50±5.16	94.39±0.20	95.36±5.73	**97.80±5.35**	91.06±2.23
2	91.76±4.08	99.09±2.75	92.23±6.19	99.02±0.66	93.16±5.44	98.49±1.63	95.94±1.53	98.46±1.41	97.48±1.51	**99.59±0.14**
3	99.37±0.22	99.55±0.96	99.29±0.74	**100.00**	99.61±0.27	99.40±0.57	99.55±0.57	98.22±0.84	99.26±0.82	99.70±0.12
4	92.82±3.13	92.38±1.95	96.76±1.94	97.79±0.78	96.33±2.07	94.28±2.71	92.40±3.63	94.28±2.68	93.70±1.09	**98.85±0.91**
5	96.66±2.42	**100.00**	99.75±0.45	99.83±0.14	96.84±1.81	99.34±0.21	98.03±0.21	**100.00**	99.83±0.71	99.91±0.12
6	90.55±4.21	95.86±4.99	98.49±3.02	94.42±2.04	98.22±3.22	**100.00**	89.18±1.39	98.03±2.59	95.73±2.06	98.68±2.72
7	85.03±5.27	95.11±0.94	94.48±3.71	87.25±2.09	92.06±2.51	**99.03±1.93**	89.02±1.51	93.50±4.53	94.87±1.45	95.91±2.66
8	**90.49±4.03**	82.28±3.57	87.85±3.13	92.56±3.03	77.79±5.74	79.90±5.79	73.77±2.93	69.60±7.06	77.04±4.68	88.31±3.54
9	77.30±6.80	82.17±2.97	88.94±5.48	**96.99±3.19**	86.02±4.52	79.87±3.87	72.07±2.91	83.03±5.04	79.62±3.67	87.34±3.11
10	78.91±10.96	**98.82±2.61**	90.70±5.57	98.01±1.59	89.01±5.01	97.26±2.36	85.41±5.43	97.26±4.95	94.11±2.95	96.43±4.51
11	90.88±3.49	93.33±3.07	92.88±4.64	96.04±4.63	89.69±4.14	86.83±3.11	80.41±0.66	91.44±6.65	95.14±6.57	**97.44±3.59**
12	82.22±4.89	90.43±3.21	89.69±4.44	92.41±3.39	87.22±2.59	92.91±4.35	87.30±2.18	94.64±7.13	96.04±4.21	**96.37±2.20**
13	94.69±3.15	94.65±1.78	96.03±1.77	98.88±1.79	94.83±4.16	93.98±1.99	65.03±2.91	95.76±2.25	95.32±1.89	**99.55±0.68**
14	99.31±0.78	**100.00**	**100.00**	99.97±0.06	99.70±0.39	**100.00**	97.79±1.98	**100.00**	**100.00**	**100.00**
15	99.06±0.63	98.78±0.43	99.87±0.25	**100.00**	99.75±0.35	**100.00**	98.12±1.75	**100.00**	99.68±0.24	**100.00**
OA(%)	89.24±0.86	93.22±0.49	93.95±1.00	95.44±0.61	91.42±0.99	93.55±0.87	87.64±1.01	92.89±1.57	93.54±0.63	**95.89±0.57**
AA(%)	90.60±0.74	94.09±0.48	94.88±0.98	96.02±0.51	92.78±0.94	94.55±0.84	87.90±0.88	93.98±1.38	94.38±0.56	**96.61±0.58**
$\kappa\times100$	88.37±0.93	92.67±0.53	93.46±1.09	95.07±0.66	90.73±1.07	93.03±0.95	86.64±1.09	92.31±1.70	93.01±0.69	**95.55±0.62**

TABLE V
CLASSIFICATION PERFORMANCE OBTAINED BY DIFFERENT METHODS FOR THE TRENTO DATASET (OPTIMAL RESULTS ARE BOLDED)

No.	CoupledCNN	CALC	S2ENet	GLT	MFT	HCT	S2EFT	GAMF	CrossHL	Proposed
1	98.65±0.86	**99.74±0.05**	99.22±0.34	98.87±0.89	97.85±0.63	98.57±0.37	92.22±5.58	97.90±0.69	99.21±0.41	99.64±0.39
2	98.15±0.48	97.42±0.52	97.70±1.09	98.95±1.74	97.19±1.09	96.68±2.54	96.14±3.26	**99.23±1.17**	98.36±0.77	97.74±1.29
3	96.33±0.97	98.51±2.76	99.73±0.52	99.78±0.31	99.08±0.47	99.65±0.33	97.12±1.84	**100.00**	99.91±0.17	98.60±0.81
4	99.81±0.32	**100.00**	99.93±0.07	**100.00**	**100.00**	99.99±0.04	99.99±0.08	99.96±0.11	**100.00**	**100.00**
5	99.84±0.22	99.99±0.02	**100.00**	**100.00**	99.93±0.05	99.62±0.28	98.33±0.77	99.98±0.43	99.94±0.04	**100.00**
6	95.47±1.51	95.46±2.16	94.08±0.91	95.14±2.52	92.99±1.08	94.12±0.62	90.22±3.12	95.27±3.13	88.37±2.39	**98.21±0.87**
OA(%)	99.00±0.85	99.21±0.23	99.03±0.19	99.24±0.17	98.42±0.14	98.74±0.19	96.94±1.22	99.14±0.44	98.50±0.26	**99.52±0.06**
AA(%)	98.04±0.07	98.52±0.75	98.44±0.27	98.79±0.31	97.37±0.26	98.11±0.37	95.67±2.15	98.73±0.57	97.63±0.39	**99.03±0.07**
$\kappa\times100$	98.66±0.22	98.95±0.31	98.70±0.26	98.98±0.23	97.89±0.18	98.31±0.27	95.92±1.63	98.85±0.59	97.99±0.34	**99.36±0.08**

Specifically, on the Houston 2013 dataset, HMAT demonstrated a relative improvement of approximately 0.5% in terms of OA, AA, and $\kappa \times 100$ compared to the suboptimal methods. Moreover, HMAT achieved the best performance in seven out of 15 categories. As depicted in Fig. 9, the classification map generated by HMAT exhibits the least amount of salt-and-pepper noise. Notably, the magnified local details highlight the preservation of the most complete edge structure in the "railway" category.

Despite the relatively easier classification task of the Trento dataset, our proposed method still achieved the best performance across all evaluation metrics and most categories. The classification map in Fig. 10 further corroborates these experimental findings.

On the MUUFL dataset, HMAT continued to demonstrate superior classification performance, particularly in terms of AA, achieving a 1.72% improvement over the suboptimal method. The magnified local image in Fig. 11 reveals that HMAT preserves the most complete edge structure in the "building" category.

In contrast, the classification maps generated by the comparative methods on the three datasets exhibited varying degrees of category confusion, particularly for the relatively simple architectures of CoupledCNN and S2EFT. These methods struggled to capture complex textures and subtle differences, leading to increased misclassification errors in the resulting maps. This highlights the superior robustness of the HMAT method in handling complex scenes.

TABLE VI
CLASSIFICATION PERFORMANCE OBTAINED BY DIFFERENT METHODS FOR THE MUUFL DATASET (OPTIMAL RESULTS ARE BOLDED)

No.	CoupledCNN	CALC	S2ENet	GLT	MFT	HCT	S2EFT	GAMF	CrossHL	Proposed
1	85.21±4.37	88.72±1.34	91.65±1.17	90.58±3.48	85.59±1.77	87.84±1.72	87.09±1.62	90.72±3.06	88.61±0.37	**92.84±0.77**
2	75.43±9.95	72.59±4.45	76.63±7.02	73.51±4.16	71.71±9.95	77.81±1.94	72.40±5.43	75.83±3.93	75.83±3.22	**81.48±1.49**
3	55.32±8.24	63.15±4.55	66.18±3.09	68.11±1.55	59.91±8.53	60.27±7.66	57.44±3.57	59.60±5.41	**69.96±6.46**	63.90±2.88
4	86.82±5.51	86.97±8.87	84.24±8.99	85.81±8.45	93.10±2.24	84.31±14.34	75.58±2.26	**94.96±4.77**	90.58±1.93	91.86±3.74
5	77.86±5.03	77.36±7.81	83.27±6.22	80.35±4.12	70.76±2.71	76.41±6.78	**83.93±1.09**	72.97±5.76	82.87±5.53	80.39±2.90
6	99.10±1.09	99.86±0.27	99.46±0.86	99.59±0.39	99.55±0.69	99.14±1.05	98.87±3.92	**100.00**	94.84±1.87	98.20±1.70
7	78.34±6.46	83.84±10.28	84.70±6.96	84.04±8.05	89.70±2.69	78.74±5.64	84.22±2.76	63.62±4.80	87.98±3.24	**91.77±1.99**
8	92.39±3.27	94.21±1.83	93.23±2.78	94.45±1.37	89.05±2.41	93.24±3.34	91.14±1.61	94.88±2.26	**96.80±2.36**	92.23±2.49
9	38.65±7.09	41.11±19.96	38.27±18.22	**58.72±6.61**	53.46±5.21	44.79±14.97	50.10±6.50	26.01±10.54	56.84±6.46	51.94±6.95
10	55.82±20.07	65.03±5.63	60.12±25.97	71.16±9.36	66.62±6.19	76.19±9.37	65.03±5.05	57.66±4.96	62.57±8.94	**80.36±9.09**
11	93.89±3.08	95.50±3.10	95.66±2.18	95.26±1.17	99.51±0.16	94.21±4.61	96.38±1.56	96.38±4.35	98.79±6.56	**99.59±2.18**
OA(%)	79.16±1.57	81.95±1.04	84.41±1.15	84.30±1.29	79.47±1.56	81.21±1.64	80.82±0.51	81.19±0.81	84.69±1.16	**85.52±0.49**
AA(%)	76.26±2.03	76.70±1.26	79.40±2.78	81.96±0.46	79.90±1.02	79.37±3.69	78.39±0.93	75.70±1.04	82.34±0.75	**84.06±1.09**
κ×100	73.26±1.79	78.94±1.82	79.73±1.38	79.67±1.54	73.87±1.86	75.78±2.69	75.40±0.60	75.64±0.94	80.22±1.49	**81.21±0.62**

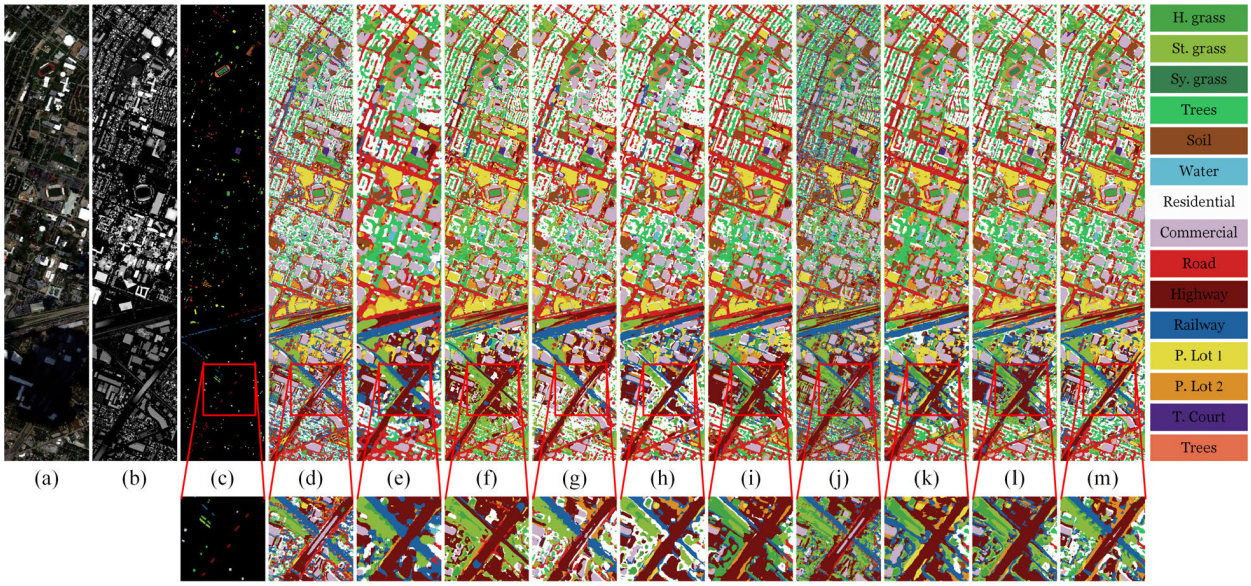

Fig. 9. Classification maps of all methods on the Houston 2013 dataset. (a)–(m) Pseudo-color image for HSI, LiDAR-based DSM, ground truth, CoupledCNN, CALC, S2ENet, GLT, MFT, HCT, S2EFT, GAMF, CrossHL, and proposed, respectively.

F. Comparison of Computational Costs

To objectively evaluate the computational efficiency of the proposed method, two metrics, Params and FLOPs, were employed to compare different models. As shown in Table VII, while CoupledCNN exhibited the lowest computational cost across all datasets, its relatively simple architecture limited its ability to effectively capture complex features, ultimately hindering classification performance. S2ENet and S2EFT suffered from similar limitations.

Although CALC introduced coupled adversarial learning to improve the performance, it also incurred additional computational costs due to unsupervised learning, resulting in increased Params and FLOPs. GAMF, while powerful, exhibited the highest computational cost, particularly on the Houston 2013 dataset, due to the integration of GNNs within a batch training framework, which necessitates the construction of numerous adjacency matrices.

In contrast, the proposed HMAT demonstrated a balance between performance and computational efficiency. By carefully designing the network architecture and employing efficient training strategies, HMAT achieved superior classification performance while maintaining reasonable computational costs.

G. Comparison of Different Training Sample Sizes

To further demonstrate the robustness of the proposed method, some experiments were conducted on the Houston 2013, Trento, and MUUFL datasets using varying numbers of

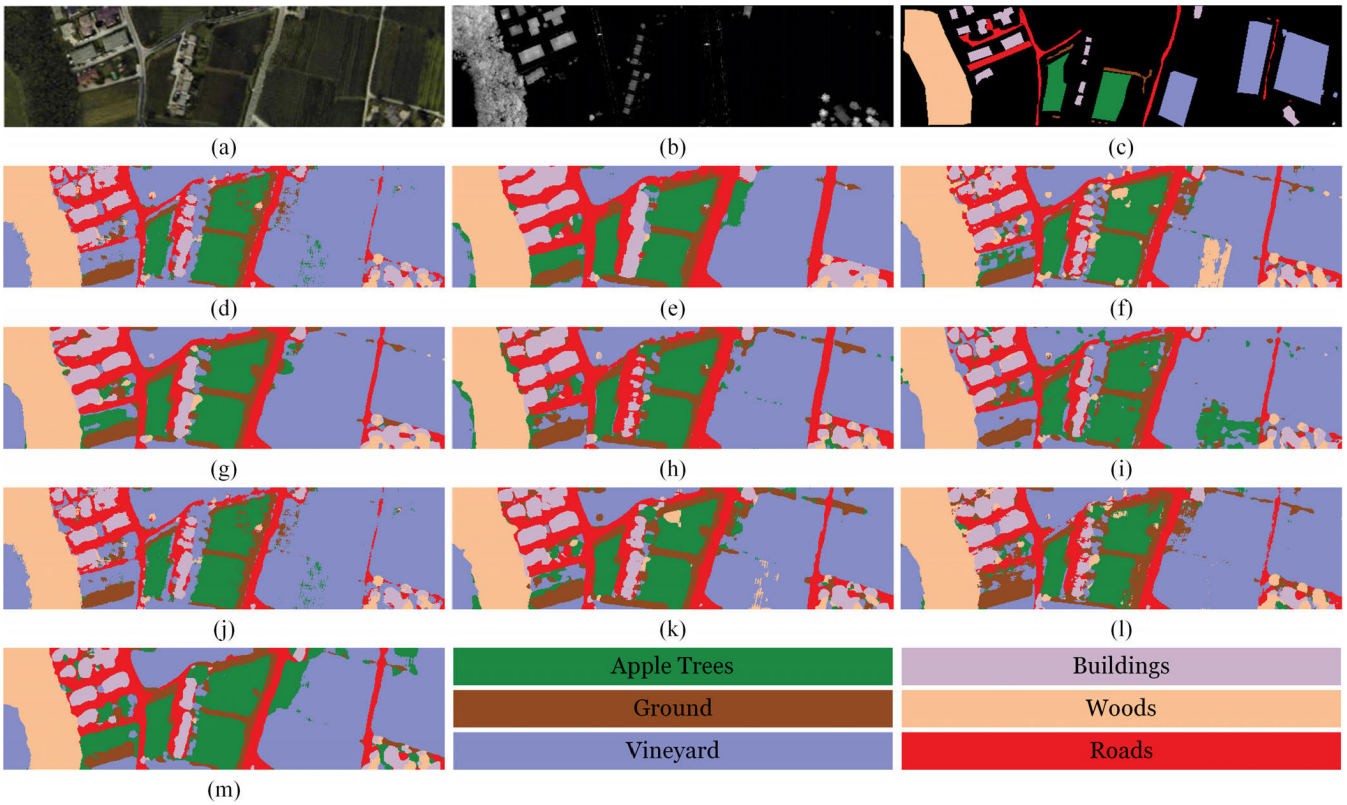

Fig. 10. Classification maps of all methods on the Trento dataset. (a)–(m) Pseudo-color image for HSI, LiDAR-based DSM, ground truth, CoupledCNN, CALC, S2ENet, GLT, MFT, HCT, S2EFT, GAMF, CrossHL, and proposed, respectively.

TABLE VII
COMPARISON OF PARAMS AND FLOPS FOR DIFFERENT METHODS

Methods	Houston 2013		Trento		MUUFL	
	Params	FLOPs	Params	FLOPs	Params	FLOPs
CoupledCNN	107.66K	2.65M	106.40K	2.68M	104.18K	2.64M
CALC	903.82K	47.83M	834.29K	30.36M	830.93K	29.92M
S2ENet	270.87K	32.77M	178.65K	21.62M	177.26K	21.48M
GLT	560.77K	156.54M	400.88K	95.38M	396.35K	93.83M
MFT	313.07K	25.30M	249.39K	11.05M	263.53K	12.72M
HCT	429.84K	79.33M	430.16K	7.98M	429.25K	79.32M
S2EFT	229.69K	22.85M	126.47K	8.35M	125.34K	8.20M
GAMF	7.31M	3.09G	3.21M	1.23G	4.56M	1.31G
CrossHL	456.94K	50.65M	392.68K	36.33M	407.40K	38.06M
Proposed	303.79K	30.29M	227.41K	21.06M	207.72K	16.13M

training samples (20, 40, 60, 80, and 100) per class. As shown in Fig. 12, the performance of all methods generally improves with an increasing number of training samples. However, CoupledCNN exhibits unexpected fluctuations on the Houston dataset, and GAMF displays similar behavior on the Trento and MUUFL datasets, suggesting limitations in their ability to effectively extract and utilize feature information.

In contrast, the proposed method consistently outperforms other methods across various training sample sizes, demonstrating its robustness to data scarcity. This superior performance highlights the effectiveness of the proposed method in capturing complex features and its generalization to unseen data.

H. t-SNE Visualization

To visually analyze the learned feature representations, t-distributed stochastic neighbor embedding (t-SNE) [67] was employed to project high-dimensional features into a 2-D space. Figs. 13–15 illustrate the t-SNE visualizations of

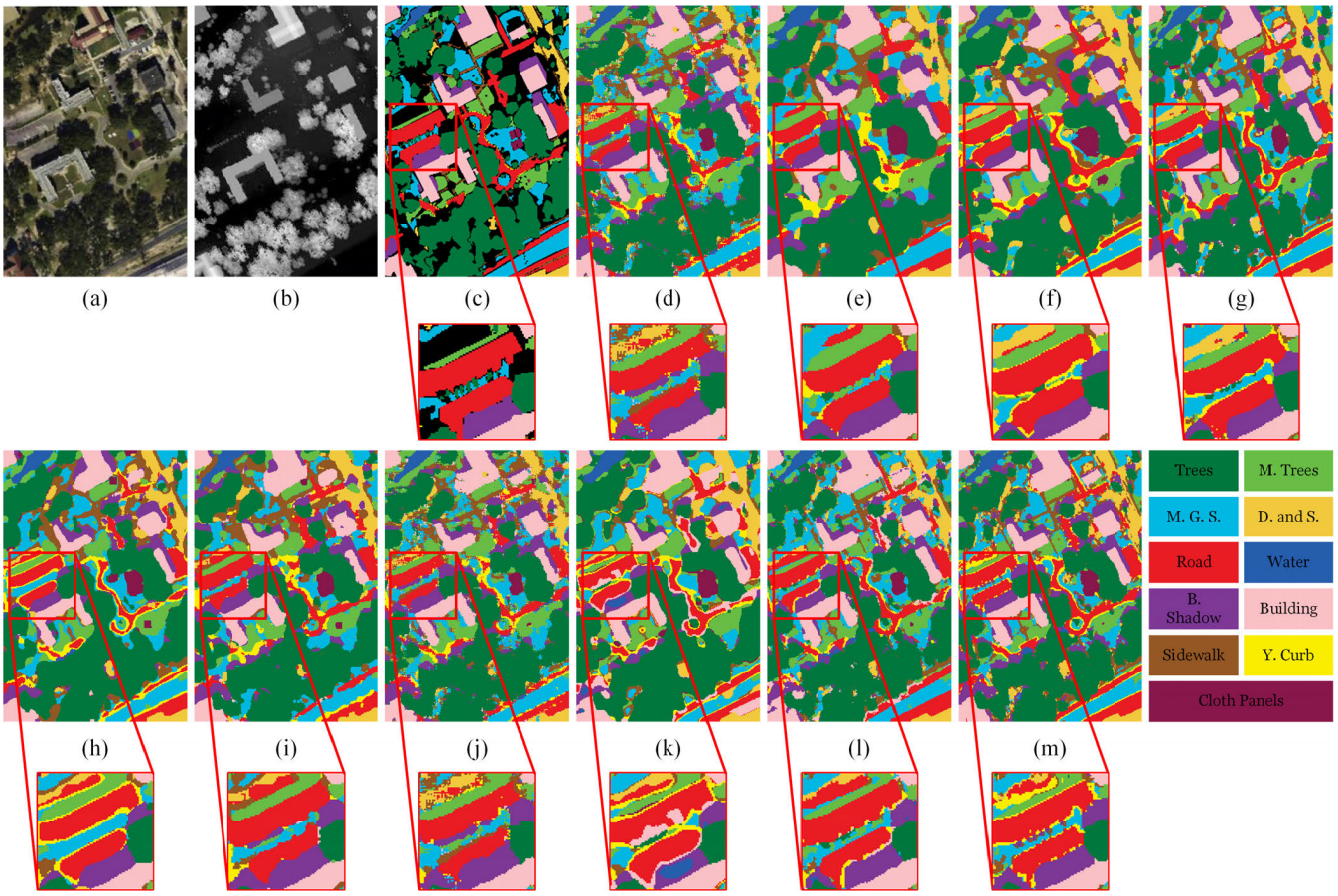

Fig. 11. Classification maps of all methods on the MUUFL dataset. (a)–(m) Pseudo-color image for HSI, LiDAR-based DSM, ground truth, CoupledCNN, CALC, S2ENet, GLT, MFT, HCT, S2EFT, GAMF, CrossHL, and proposed, respectively.

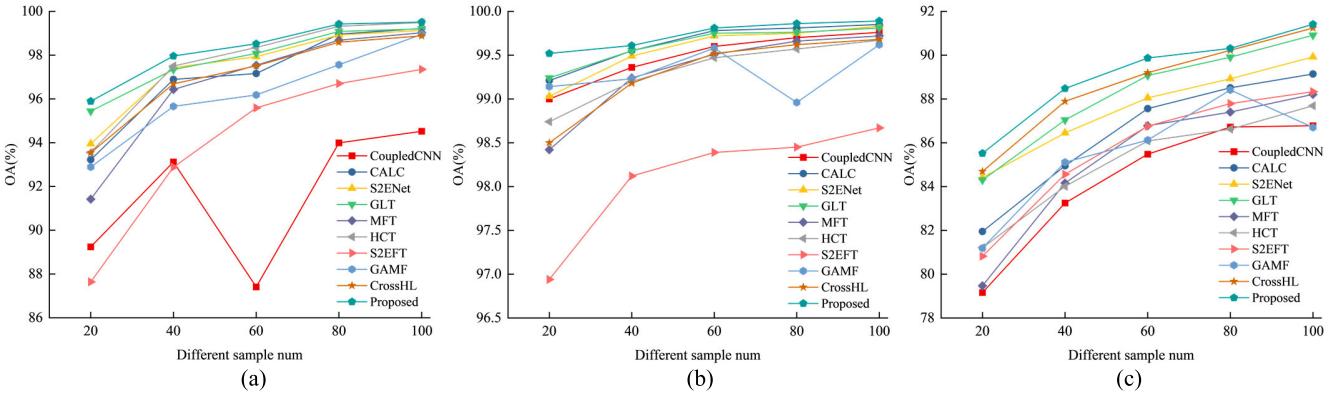

Fig. 12. Comparison of different training sample sizes. (a)–(c) Houston 2013, Trento, and MUUFL, respectively.

features extracted by the proposed method and three state-of-the-art comparison methods on the Houston, Trento, and MUUFL datasets, respectively.

The proposed method consistently exhibits superior clustering performance across all datasets. Specifically, on the Houston 2013 dataset, HMAT demonstrates the most compact intraclass clusters, the largest interclass margins, and the fewest misclassifications, particularly for the eighth class (commercial). Similarly, on the Trento dataset, HMAT achieves the best clustering performance, especially for the fifth class (vineyard). On the MUUFL dataset, HMAT exhibits significantly better clustering performance for the eighth class (buildings) compared to other methods. In contrast, the comparison methods suffer from varying degrees of intraclass dispersion and interclass confusion.

I. Comparison of Confusion Matrices

Figs. 16–18 present the confusion matrix visualizations of the proposed method and three state-of-the-art comparison methods on the Houston, Trento, and MUUFL datasets. The proposed method demonstrates superior performance in minimizing category confusion. While all methods exhibit

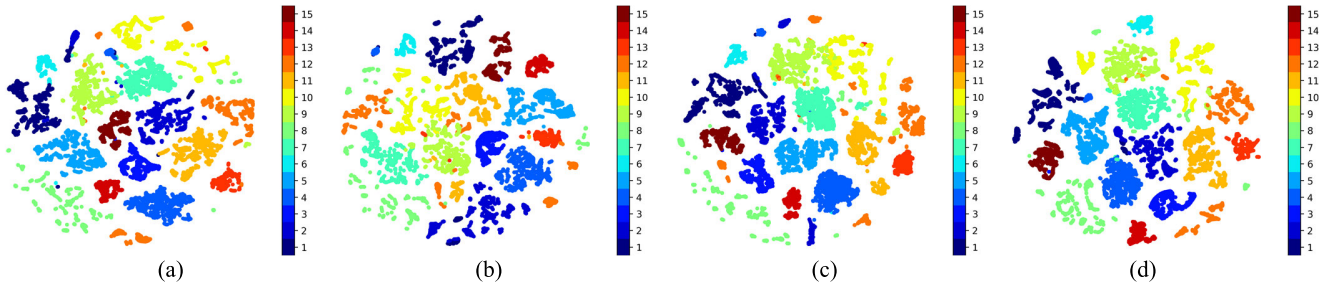

Fig. 13. Feature visualization via t-SNE on the Houston 2013 dataset. (a)–(d) GLT, HCT, CrossHL, and proposed, respectively.

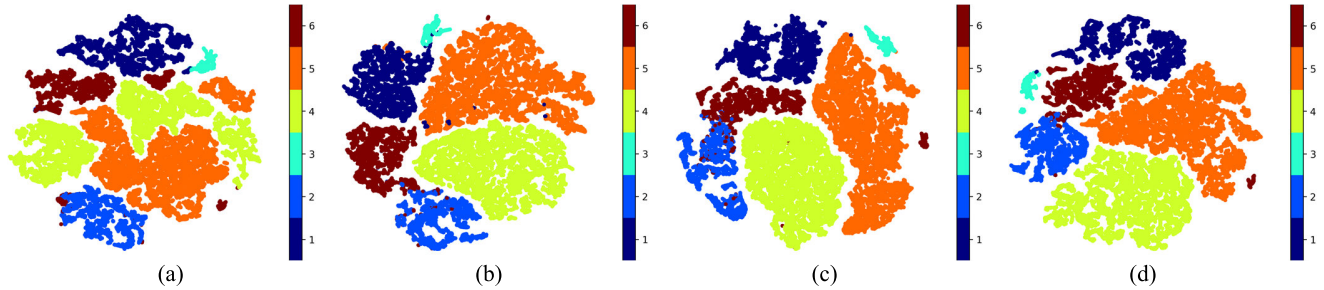

Fig. 14. Feature visualization via t-SNE on the Trento dataset. (a)–(d) GLT, HCT, CrossHL, and proposed, respectively.

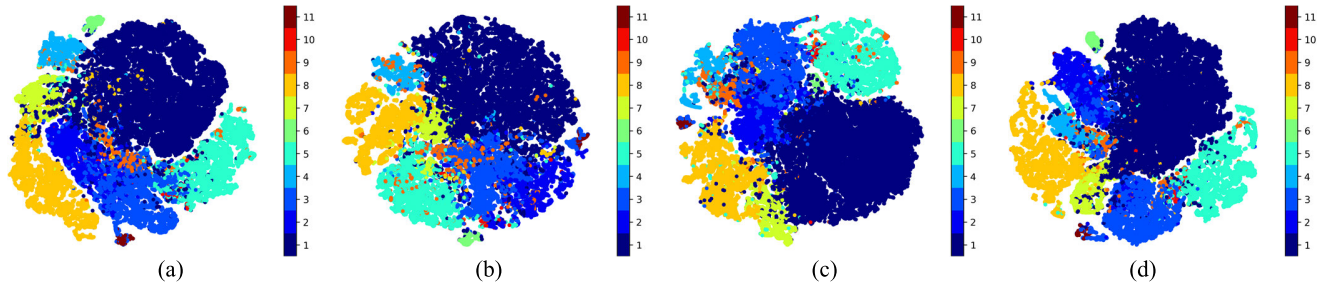

Fig. 15. Feature visualization via t-SNE on the MUUFL dataset. (a)–(d) GLT, HCT, CrossHL, and proposed, respectively.

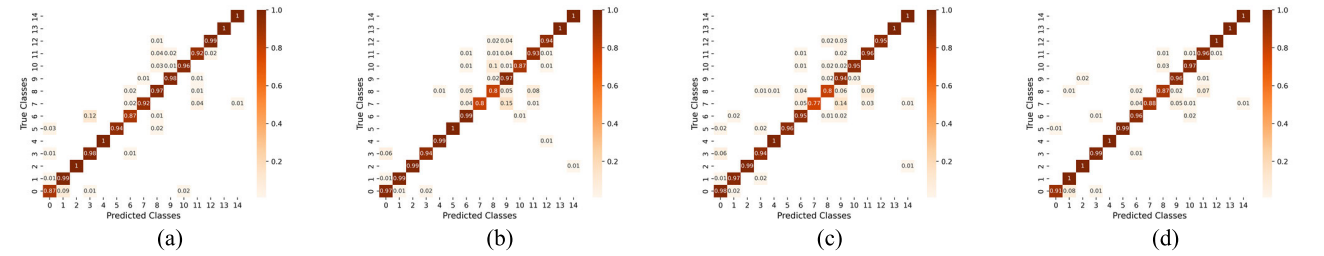

Fig. 16. Confusion matrices on the Houston 2013 dataset. (a)–(d) GLT, HCT, CrossHL, and proposed, respectively.

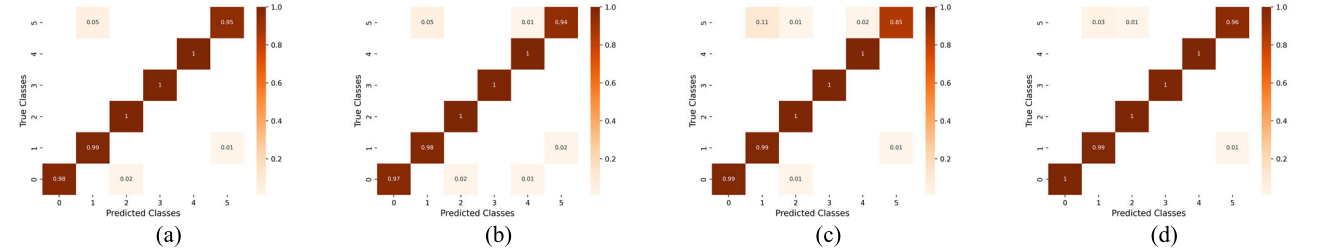

Fig. 17. Confusion matrices on the Trento dataset. (a)–(d) GLT, HCT, CrossHL, and proposed, respectively.

some degree of confusion, particularly in the ninth category of the Houston dataset, the HMAT method exhibits stronger discriminative power, significantly reducing misclassification rates. Similarly, the HMAT method demonstrates lower lev-

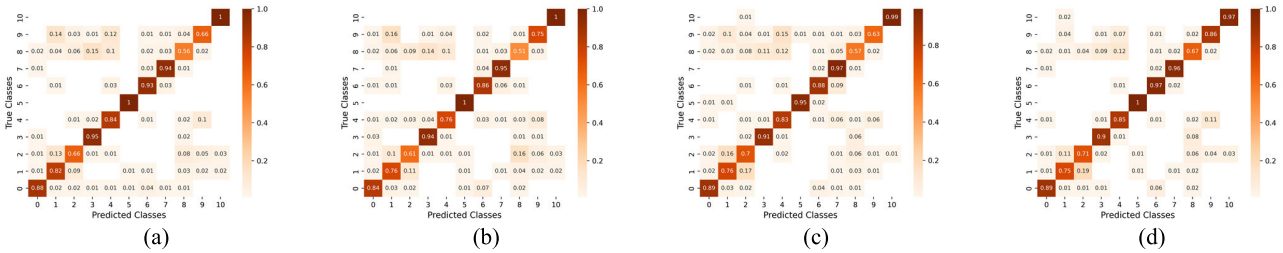

Fig. 18. Confusion matrices on the MUUFL dataset. (a)–(d) GLT, HCT, CrossHL, and proposed, respectively.

els of class confusion on the Trento and MUUFL datasets. These results fully demonstrate the superiority of the proposed method.

IV. Conclusion

In this article, a novel HMAT method is proposed for the joint classification of hyperspectral and LiDAR data. To effectively fuse multimodal data, an HMFA module is designed that can extract multilevel features from different modalities and enables cross-level feature fusion to realize the interaction and complementarity of different modalities. In addition, a PIP module is constructed to enhance local neighborhood information and preserve detailed features. To fully utilize multimodal fusion features at various scales, we employ a transformer encoder based on the proposed MHAA to process these features and extract more discriminative global features. The experimental results demonstrate that the HMAT method outperforms existing state-of-the-art methods in the joint classification task of hyperspectral and LiDAR data.

Nevertheless, how to further leverage the intrinsic features of different modalities to enhance feature extraction within the model remains a challenging and crucial question. In our future work, building upon the proposed method, we will further concentrate on the unique characteristics of hyperspectral and LiDAR data to design deep learning models that are more tailored to multimodal remote sensing data analysis.

References

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.

[2] L. Gedminas and S. Martin, "Soil organic matter mapping using hyperspectral imagery and elevation data," in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, USA, Mar. 2019, pp. 1–8.

[3] L. Chen, K. Tan, X. Wang, and C. Pan, "Estimation soil organic matter using airborne hyperspectral imagery," in *Proc. 13th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Athens, Greece, Oct. 2023, pp. 1–5.

[4] A. Nisha and A. Anitha, "Current advances in hyperspectral remote sensing in urban planning," in *Proc. 3rd Int. Conf. Intell. Comput. Instrum. Control Technol. (ICICICT)*, Aug. 2022, pp. 94–98.

[5] C. Weber et al., "Hyperspectral imagery for environmental urban planning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 1628–1631.

[6] M. Wang, Y. Xu, Z. Wang, and C. Xing, "Deep margin cosine autoencoder-based medical hyperspectral image classification for tumor diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–12, 2023.

[7] Y. Li, R. Wu, Q. Tan, Z. Yang, and H. Huang, "Masked spectral bands modeling with shifted windows: An excellent self-supervised learner for classification of medical hyperspectral images," *IEEE Signal Process. Lett.*, vol. 30, pp. 543–547, 2023.

[8] C. M. Gevaert, J. Suomalainen, J. Tang, and L. Kooistra, "Generation of spectral–temporal response surfaces by combining multispectral satellite and hyperspectral UAV imagery for precision agriculture applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3140–3146, Jun. 2015.

[9] A. U. G. Sankararao, K. Saikiran, and P. Rajalakshmi, "Hyperspectral image denoising: A comparative study on UAV based vegetation data," in *Proc. 13th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Athens, Greece, Oct. 2023, pp. 1–5.

[10] S. Pattem and S. Thatavarti, "Hyperspectral image classification using machine learning techniques—A survey," in *Proc. IEEE Int. Students' Conf. Electr., Electron. Comput. Sci. (SCEECS)*, Bhopal, India, Feb. 2023, pp. 1–14.

[11] M. Ahmad et al., "Hyperspectral image classification-traditional to deep models: A survey for future prospects," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 968–999, 2022.

[12] B. Rasti, P. Ghamisi, and R. Gloaguen, "Hyperspectral and LiDAR fusion using extinction profiles and total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3997–4007, Jul. 2017.

[13] R. Hänsch and O. Hellwich, "Fusion of multispectral LiDAR, hyperspectral, and RGB data for urban land cover classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 2, pp. 366–370, Feb. 2021.

[14] 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.

[15] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal low-rank tensor completion for visual data," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 673–685, Feb. 2021.

[16] C. Ge, Q. Du, W. Li, Y. Li, and W. Sun, "Hyperspectral and LiDAR data classification using kernel collaborative representation based residual fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1963–1973, Jun. 2019.

[17] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza, "Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2971–2983, Jun. 2015.

[18] M. Brell, K. Segl, L. Guanter, and B. Bookhagen, "Hyperspectral and LiDAR intensity data fusion: A framework for the rigorous correction of illumination, anisotropic effects, and cross calibration," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2799–2810, May 2017.

[19] L. Fang, N. He, S. Li, P. Ghamisi, and J. A. Benediktsson, "Extinction profiles fusion for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1803–1815, Mar. 2018.

[20] W. Liao, A. Pižurica, R. Bellens, S. Gautama, and W. Philips, "Generalized graph-based fusion of hyperspectral and LiDAR data using morphological features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 552–556, Mar. 2015.

[21] M. Pedergnana, P. R. Marpu, M. D. Mura, J. A. Benediktsson, and L. Bruzzone, "Classification of remote sensing optical and LiDAR data using extended attribute profiles," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 7, pp. 856–865, Nov. 2012.

[22] M. Dalponte, L. Bruzzone, and D. Gianelle, "Fusion of hyperspectral and LiDAR remote sensing data for classification of complex forest areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1416–1427, May 2008.

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

[24] J. Li, Y. Liu, R. Song, Y. Li, K. Han, and Q. Du, "Sal^{2}RN: A spatial–spectral salient reinforcement network for hyperspectral and LiDAR data fusion classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500114.
[25] Y. Peng, Y. Zhang, B. Tu, C. Zhou, and Q. Li, "Multiview hierarchical network for hyperspectral and LiDAR data classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1454–1469, 2022.
[26] S. Xia, X. Zhang, H. Meng, and L. Jiao, "Ternary modality contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5522017.
[27] Y. Kong, Y. Cheng, Y. Chen, and X. Wang, "Joint classification of hyperspectral image and LiDAR data based on spectral prompt tuning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5521312.
[28] W.-S. Hu, W. Li, H.-C. Li, F.-H. Huang, and R. Tao, "Global clue-guided cross-memory quaternion transformer network for multisource remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 14, 2024, doi: 10.1109/TNNLS.2024.3406735.
[29] C. Shi, X. Zhao, and L. Wang, "A multi-branch feature fusion strategy based on an attention mechanism for remote sensing image scene classification," *Remote Sens.*, vol. 13, no. 10, p. 1950, May 2021.
[30] R. Xiao, C. Zhong, W. Zeng, M. Cheng, and C. Wang, "Novel convolutions for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5907313.
[31] C. Shi, D. Liao, T. Zhang, and L. Wang, "Hyperspectral image classification based on expansion convolution network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528316.
[32] C. Shi, T. Wang, and L. Wang, "Branch feature fusion convolution network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5194–5210, 2020.
[33] 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.
[34] 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.
[35] 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.
[36] S. Fang, K. Li, and Z. Li, "S^{2}ENet: 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] W. Han, W. Miao, J. Geng, and W. Jiang, "CMSE: Cross-modal semantic enhancement network for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5509814.
[38] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
[39] T. Zhang, S. Xiao, W. Dong, J. Qu, and Y. Yang, "A mutual guidance attention-based multi-level fusion network for hyperspectral and LiDAR classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[40] X. Wang, J. Zhu, Y. Feng, and L. Wang, "MS2CANet: Multiscale spatial–spectral cross-modal attention network for hyperspectral image and LiDAR classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
[41] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.
[42] S. Hao, Y. Xia, and Y. Ye, "Generative adversarial network with transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.
[43] T. Lu, K. Ding, W. Fu, S. Li, and A. Guo, "Coupled adversarial learning for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 93, pp. 118–131, May 2023.
[44] Y. Yang, J. Qu, W. Dong, T. Zhang, S. Xiao, and Y. Li, "TMCFN: Text-supervised multidimensional contrastive fusion network for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5511015.
[45] H. Hu, M. Yao, F. He, and F. Zhang, "Graph neural network via edge convolution for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[46] 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.
[47] 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.
[48] H. Wang, Y. Cheng, X. Liu, and X. Wang, "Reinforcement learning based Markov edge decoupled fusion network for fusion classification of hyperspectral and LiDAR," *IEEE Trans. Multimedia*, vol. 26, pp. 7174–7187, 2024.
[49] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.
[50] Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11999–12009.
[51] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12114–12124.
[52] B. Graham et al., "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12239–12249.
[53] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11916–11925.
[54] 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.
[55] 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.
[56] 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.
[57] Y. Feng, J. Zhu, R. Song, and X. Wang, "S2EFT: Spectral–spatial-elevation fusion transformer for hyperspectral image and LiDAR classification," *Knowl.-Based Syst.*, vol. 283, Jan. 2024, Art. no. 111190.
[58] 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.
[59] T. Song, Z. Zeng, C. Gao, H. Chen, and J. Li, "Joint classification of hyperspectral and LiDAR data using height information guided hierarchical fusion-and-separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5505315.
[60] Z. Zeng, T. Song, X. Ma, Y. Jiu, and H. Sun, "Joint classification of hyperspectral and LiDAR data using cross-modal hierarchical frequency fusion network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 5390–5394.
[61] X. Shi, J. Lin, Y. Rao, Y. Sun, and F. Gao, "Gated-cross aggregation network for hyperspectral and LiDAR data classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Pasadena, CA, USA, Jul. 2023, pp. 1265–1268.
[62] J. Qu, L. Zhang, W. Dong, N. Li, and Y. Li, "Shared-private decoupling-based multilevel feature alignment semisupervised learning for HSI and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5537314.
[63] W. Dong, J. Qu, T. Zhang, S. Xiao, and Y. Li, "Contrastive constrained cross-scene model-informed interpretable classification strategy for hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5527114.
[64] C. Debes et al., "Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.
[65] P. Gader, A. Zare, R. Close, J. Aitken, and G. Tuell, "MUUFL gulfport hyperspectral and LiDAR airborne data set," Univ. Florida, Gainesville, FL, USA, Tech. Rep. REP-2013-570, Oct. 2013.
[66] X. Du and A. Zare, "Technical report: Scene label ground truth map for MUUFL gulfport data set," Univ. Florida, Gainesville, FL, USA, Tech. Rep. 20170417, Apr. 2017.
[67] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, Nov. 2008.

Fei Zhu received the bachelor's degree from Luoyang Institute of Science and Technology, Luoyang, China, in 2021. He is currently pursuing the master's degree with Qiqihar University, Qiqihar, China.

His research interests include hyperspectral image processing and machine learning.

Cuiping Shi (Member, IEEE) received the M.S. degree from Yangzhou University, Yangzhou, China, in 2007, and the Ph.D. degree from Harbin Institute of Technology (HIT), Harbin, China, in 2016.

From 2017 to 2020, she was a Post-Doctoral Researcher with the College of Information and Communications Engineering, Harbin Engineering University, Harbin. Since 2024, she has been working with the College of Information Engineering, Huzhou University, Huzhou, China. She is a Professor with the Department of Communication Engineering, Qiqihar University, Qiqihar, China. She has published two academic books about remote sensing image processing and more than 90 papers in journals and conference proceedings. Her main research interests include remote sensing image processing, pattern recognition, and machine learning.

Dr. Shi's doctoral dissertation won the Nomination Award of Excellent Doctoral Dissertation of Harbin University of Technology (HIT) in 2016.

Kaijie Shi received the bachelor's degree from Heilongjiang University of Science and Technology, Harbin, China, in 2021. He is currently pursuing the master's degree with Qiqihar University, Qiqihar, China.

His research interests include remote sensing image compression and machine learning.

Liguo Wang (Member, IEEE) received the M.S. and Ph.D. degrees in signal and information processing from Harbin Institute of Technology, Harbin, China, in 2002 and 2005, respectively.

From 2006 to 2008, he held a post-doctoral position at Harbin Engineering University, Harbin. Since 2020, he has worked with the College of Information and Communication Engineering, Dalian Nationalities University, Dalian, China. He has published two books about hyperspectral image processing and more than 130 papers in journals and conference proceedings. His main research interests include remote sensing image processing.