Joint Classification of Hyperspectral Images and LiDAR Data Based on Candidate Pseudo Labels Pruning and Dual Mixture of Experts

Yi Kong, *Member, IEEE*, Shaocai Yu, Yuhu Cheng, *Senior Member, IEEE*, C. L. Philip Chen, *Life Fellow, IEEE*, and Xuesong Wang, *Member, IEEE*

***Abstract*— Hyperspectral images (HSIs) contain rich spatial and spectral information, while light detection and ranging (LiDAR) data can provide elevation details. Effectively fusing HSI and LiDAR data can help achieve more accurate classification results. However, the joint classification of HSI and LiDAR data still faces several challenges, such as the redundancy of HSI spectral bands, the limitations of singular multimodal data fusion strategy, and the high cost of pixelwise labeling in remote sensing images. To tackle these challenges, we propose a classification method based on candidate pseudo labels pruning and dual mixture of experts (CPLP-DMoEs). First, we employ the multihead mixture of bands (MMoBs) to perform diverse, dense mixing of spectral bands, thereby alleviating the issue of high similarity between adjacent bands. Then, to overcome the limitations of single fusion strategies, we design a mixture of multimodal fusion expert (MoMFE) mechanism, which selects and mixes multiple fusion experts (FEs) to achieve diverse feature fusion of HSI and LiDAR data. Next, we introduce information entropy to balance the selection of FEs. Finally, facing the challenge of limited labeled samples, we propose a candidate pseudo labels pruning (CPLP)-based semi-supervised learning method. CPLP can prune the candidate pseudo label set from both intrasample and intersample perspectives to obtain more reliable pseudo labels, thereby facilitating the learning of a more accurate classification model. The experimental results on three datasets, including Houston 2013, MUUFL, and Augsburg, validate the effectiveness of the proposed method.**

***Index Terms*— Classification, hyperspectral image (HSI), light detection and ranging (LiDAR) data, mixture of experts, pseudo labels.**

I. Introduction

MULTIMODAL remote sensing images are captured by different sensors and can reflect various characteristics of ground objects, making them widely used in fields such as mineral exploration [1], environmental monitoring [2], forest management [3], precision agriculture [4], oil spill monitoring [5], visual grounding [6], and hyperspectral image (HSI) super-resolution [7], [8], [9]. HSIs are high-dimensional images with numerous spectral bands, capable of representing both the spatial distribution and spectral reflectance information of ground objects simultaneously. HSIs are often used as primary data for land cover classification tasks [10], [11]. Light detection and ranging (LiDAR) data can record the elevation information of observed objects and are less affected by environmental factors, such as weather [12]. Due to differences in imaging mechanisms, HSI and LiDAR data capture different aspects of ground object features. Effectively utilizing their complementary information can enhance the performance of land cover classification [13], [14]. Therefore, studying joint classification methods based on HSI and LiDAR data for remote sensing image classification is a hot topic.

In the early stages, remote sensing image classification methods primarily relied on single-modal data. With the development of deep learning, known for its powerful feature learning and pattern recognition capabilities, it has been widely applied to classification tasks [15], [16]. For instance, convolutional neural networks (CNNs) utilize weight sharing and local connectivity mechanisms, enabling the extraction of deep, nonlinear abstract features with fewer parameters. Consequently, various classification methods have been developed based on CNNs. For example, Chakraborty and Trehan [17] proposed a wavelet transform layer to extract spectral features. Zheng et al. [18] introduced a fast, patch-free global learning framework based on fully CNN, which integrates the global spatial information and semantic information of HSI. Typically, CNNs focus more on extracting the local features and struggle to process global spectral information, making it difficult to capture significant spectral changes. To this end, the vision transformer (ViT) [19], capable of capturing global information, has been used for feature extraction and classification in remote sensing images. Moreover, to leverage both the ViT's ability to extract global spectral information and the CNN's advantage in capturing spatial positional information and pixel interrelationships, numerous methods combining CNNs and ViTs have been proposed. These methods facilitate the effective transmission and interaction of spatial information. For instance, Sun et al. [20] introduced a spectral–spatial feature tokenization transformer to capture spectral–spatial features and high-level semantic

Received 30 October 2024; revised 19 January 2025; accepted 12 February 2025. Date of publication 19 February 2025; date of current version 28 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62176259 and Grant 62373364 and in part by the Key Research and Development Program of Jiangsu Province under Grant BE2022095. *(Corresponding author: Xuesong Wang.)*

Yi Kong, Shaocai Yu, Yuhu Cheng, and Xuesong Wang are with the Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, and the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China (e-mail: kongyicumt@163.com; yshaocai@163.com; chengyuhu@163.com; wangxuesongcumt@163.com).

C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: philip.chen@ieee.org).

Digital Object Identifier 10.1109/TGRS.2025.3543498

features. Yang et al. [21] combined convolution operations with transformers to annotate and utilize spectral differential information and local spatial context information. He et al. [22] used CNN to extract spatial features of HSI and proposed a densely connected transformer to capture spectral sequence relationships.

Despite the outstanding classification performance achieved by the aforementioned works, relying solely on single-modal data has certain limitations. For instance, different types of ground objects may exhibit similar spectral curves, thus leading to misclassification. Fortunately, LiDAR data can provide additional elevation information. Therefore, effectively utilizing the complementary information from HSI and LiDAR data can significantly enhance the model's feature representative ability. For example, Lin et al. [23] proposed a difference transformer to simulate the differences between dual-temporal images, using a token-swapping difference assessment module to highlight inconsistencies between the change areas and their surrounding environments, thereby emphasizing the differences between the dual-temporal images. Hang et al. [24] introduced an unsupervised feature learning model that first extracts features from HSI and LiDAR data, and then employs a dual fine-tuning strategy to transfer the extracted features. Ghamisi et al. [25] utilized features extracted from HSI and LiDAR data as inputs to a CNN for deeper feature extraction. Xu et al. [26] used two independent CNN branches to separately extract deep features from HSI and LiDAR data and designed a decision-level fusion method to achieve multimodal feature fusion. Ding et al. [27] first utilized CNN to extract multiscale local features from HSI and LiDAR data separately. Then, a feature-level fusion strategy was designed to integrate features from HSI and LiDAR data. Finally, a transformer branch was used to extract global features from the fused features. Zhao et al. [28] proposed a hierarchical CNN and transformer network, which first extracted features from HSI and LiDAR data separately. Then, a cross-attention module was designed to fuse features from HSI and LiDAR data.

Given the outstanding performance of CNN and ViT, this article also employs CNN and ViT for feature extraction. Moreover, the aforementioned multimodal classification methods typically adopt a single fusion strategy, which limits the model to a one-sided or even suboptimal fusion approach, such as over-relying on features from a particular modality. To address this, inspired by the mixtures of experts (MoEs) [29], this article employs multiple fusion expert networks to achieve diversified and dynamic multimodal fusion. Furthermore, while HSI provides rich spectral features, it also exhibits strong similarities between adjacent bands. To tackle this issue, we utilize the selection and mixing strategy of MoE to obtain a subset from the redundant bands. However, the commonly used sparse selection strategy [30] inevitably causes the loss of spectral information. To mitigate this, a soft mixing strategy [31] is introduced to achieve weighted aggregation across all bands. Additionally, to achieve diversified band mixing, a multihead mixture of band (MMoB) method is further developed.

Almost all supervised deep learning methods rely on a large number of labeled samples; otherwise, they suffer from severe overfitting. However, pixelwise labeling in remote sensing images is costly. Consequently, numerous semi-supervised and self-supervised learning strategies have been proposed. For instance, Duan et al. [32] proposed an innovative work for HSI classification in marine scenes, in which a novel self-supervised learning paradigm was designed to solve the small sample issue. Ding et al. [33] proposed a novel graph neural network based on the distribution of labeled superpixel locations, which enhances the influence of labeled superpixels located at the class centers. Zhao et al. [34] proposed a semi-supervised image registration framework based on multimodal cross-attention, which maps cross-modal features to the decoder for keypoints detection and expands the keypoints set of unlabeled samples by continuously adding newly extracted reliable keypoints. Ding et al. [35] introduced a novel uncertainty-aware contrastive learning method. Initially, it analyzes label uncertainty based on multilevel probability estimation, dividing the samples into reliable and unreliable classes, and then employs a well-designed hybrid contrastive learning strategy to jointly learn from these two types of samples. Wu and Prasad [36] obtained the pseudo labels of unlabeled samples by a nonparametric Bayesian clustering method and applied them to pretrain a deep CNN, thereby improving the initialization of the classification model and enhancing classification performance. The performance of the aforementioned pseudo-label-based semi-supervised learning methods largely depends on the accuracy of the pseudo labels. To obtain more reliable pseudo labels, this article proposes a candidate pseudo labels pruning (CPLP) method for unlabeled data. Compared to directly predicting a single pseudo labels, CPLP generates multiple pseudo labels for the unlabeled data to form a candidate label set. Then, two rounds of screening on the candidate label set from the perspectives of model prediction confidence and feature similarity are, respectively, performed to obtain more reliable pseudo labels. In summary, this article proposes a method for the joint classification of HSI and LiDAR data based on CPLP and dual mixture of experts (CPLP-DMoEs). The main contributions of this article are summarized as follows.

1) An MoE-inspired MMoB method is designed. On the one hand, the gating mechanism assigns higher weights to important bands, thereby avoiding the issue of band information loss caused by sparse MoE-based methods. On the other hand, to avoid over-reliance on a single aggregation result and to obtain more diversified aggregated features, the proposed method is extended to a multihead version.
2) To address the issue of single multimodal fusion strategies leading to one-sided and suboptimal fusion models, a mixture of multimodal fusion expert (MoMFE) strategy is proposed. Different experts correspond to different multimodal fusion strategies, and the diverse fusion strategies are weighted and mixed through the importance scores from a gating network. Additionally, to alleviate the load imbalance problem, more balanced expert selection is achieved by minimizing information entropy.

3) To obtain more reliable pseudo labels that support accurate semi-supervised model learning, the CPLP technique is designed to screen the candidate label set from both intrasample and intersample perspectives.

The rest of this article is organized as follows. In Section II, the details of the proposed network are described. Section III presents the experimental data, results, and analysis of different models. Finally, the conclusions of this work are drawn in Section IV.

II. CPLP-DMoE-Based Joint Classification of HSI and LiDAR Data

As shown in Fig. 1, the joint classification method of HSI and LiDAR data-based CPLP-DMoE mainly consists of three steps.

1) By selecting several pixels around a specific pixel, the neighbor representation of HSI is constructed. Subsequently, MMoB is utilized to achieve a diverse mixing of HSI spectral features. Then, spatial feature extraction is performed using 2D-CNN.
2) To obtain the same number of channels as HSI, a 2D-CNN is used for spatial feature extraction of LiDAR data. Then, MoMFE is employed for the dynamic and diverse fusion of HSI and LiDAR data features, resulting in the corresponding fused features.
3) ViT is used for global feature extraction of the fused features. Additionally, to obtain more reliable pseudo labels and facilitate effective semi-supervised learning, the CPLP is employed to screen the candidate label set from both intrasample and intersample perspectives based on model prediction confidence and feature similarity for unlabeled samples.

A. HSI Feature Extraction Based on MMoB

Given the HSI $\boldsymbol{X}^{(\mathrm{H})} \in \mathbb{R}^{m\times n\times d^{(\mathrm{H})}}$ and LiDAR data $\boldsymbol{X}^{(\mathrm{L})} \in \mathbb{R}^{m\times n\times d^{(\mathrm{L})}}$ of the same region on the Earth's surface, where m and n represent the width and height of this region, $d^{(\mathrm{H})}$ and $d^{(\mathrm{L})}$ represent the number of channels in the HSI and LiDAR data, respectively. For each pixel, by selecting several surrounding pixels, we can construct the neighbor representation of the HSI and LiDAR data, $\mathcal{X}^{(\mathrm{H})} \in \mathbb{R}^{s\times s\times d^{(\mathrm{H})}}$ and $\mathcal{X}^{(\mathrm{L})} \in \mathbb{R}^{s\times s\times d^{(\mathrm{L})}}$, where s is the spatial dimension of the neighborhood representation. The labeled and unlabeled sample sets are denoted as $\mathcal{D}^{(\mathrm{S})} = \left\{\left(\mathcal{X}_i^{(\mathrm{H})}, \mathcal{X}_i^{(\mathrm{L})}, \boldsymbol{y}_i\right) | i = 1, \ldots, n^{(\mathrm{S})}\right\}$ and $\mathcal{D}^{(\mathrm{U})} = \left\{\left(\mathcal{X}_j^{(\mathrm{H})}, \mathcal{X}_j^{(\mathrm{L})}\right) | j = 1, \ldots, n^{(\mathrm{U})}\right\}$, respectively. Here, $n^{(\mathrm{S})}$ and $n^{(\mathrm{U})}$ are the number of samples in the labeled and unlabeled sample sets, respectively, and $\boldsymbol{y}^{(i)}$ is the corresponding class vector.

While HSI provides rich spectral features, it also exhibits strong similarity between neighboring bands. To reduce the interference caused by redundant bands, this article introduces MoE to fuse the spectral bands. MoE typically consists of a group of expert networks and a gating network. The expert networks focus on different tasks or aspects of the data, while the gating network is responsible for assigning inputs to the appropriate experts. Generally, MoE is based on a sparse strategy, where the gating network only activates a subset of experts for forward computation at each step. Therefore, MoE can enhance the capability of the model without incurring additional forward computation costs. However, when directly applying MoE to the spectral band mixing of HSI, the sparse strategy results in only a subset of bands being activated, which leads to the loss of spectral information. Inspired by [30], this article proposes an MMoB method based on the dense strategy. Given the vectorized HSI neighborhood representation $\hat{\mathcal{X}}^{(\mathrm{H})} \in \mathbb{R}^{s^2\times d^{(\mathrm{H})}}$, a dense feature aggregation is realized by multiple band-mixing experts

$$\boldsymbol{e}_i = \hat{\mathcal{X}}^{(\mathrm{H})} \cdot \boldsymbol{w}_i^{(\mathrm{B})} \tag{1}$$

where $\boldsymbol{w}_i^{(\mathrm{B})} \in \mathbb{R}^{d^{(\mathrm{H})}\times 1}$ represents the ith spectral mixture expert (SME). The more important spectral bands will be assigned larger weights for downstream tasks. $\boldsymbol{e}_i \in \mathbb{R}^{s^2\times 1}$ represents the band-mixing features. Additionally, to obtain more diverse band-mixing results, $n_{\mathrm{E}^{(\mathrm{B})}}$ spectral mixture matrices are set as $\boldsymbol{W}_i^{(\mathrm{B})} = [\boldsymbol{w}_{i,1}^{(\mathrm{B})}, \ldots, \boldsymbol{w}_{i,n_{\mathrm{E}^{(\mathrm{B})}}}^{(\mathrm{B})}] \in \mathbb{R}^{d^{(\mathrm{H})}\times n_{\mathrm{E}^{(\mathrm{B})}}}$. Then, the multihead spectral mixing features of the ith SME are $\boldsymbol{E}_i = [\boldsymbol{e}_{i,1}, \ldots, \boldsymbol{e}_{i,n_{\mathrm{E}^{(\mathrm{B})}}}] \in \mathbb{R}^{s^2\times n_{\mathrm{E}^{(\mathrm{B})}}}$. Furthermore, the features of pixels from different spatial locations and classes should be obtained from different SMEs or by mixing multiple SMEs. Therefore, MMoB sets $n^{(\mathrm{B})}$ SMEs and utilizes a sparse gating mechanism to control the selection of different SMEs. The gating mechanism initially utilizes a gating network to obtain the contribution scores $\boldsymbol{g} \in \mathbb{R}^{s^2\times n^{(\mathrm{B})}}$ corresponding to different experts

$$\boldsymbol{g} = \hat{\mathcal{X}}^{(\mathrm{H})} \cdot \boldsymbol{W}_{\mathrm{G_B}} \tag{2}$$

where $\boldsymbol{W}_{\mathrm{G_B}} \in \mathbb{R}^{d^{(\mathrm{H})}\times n^{(\mathrm{B})}}$ represents the weights of the gating network in MMoB, and $n^{(\mathrm{B})}$ represents the number of SMEs. Subsequently, based on the contribution scores $\boldsymbol{g}$, K_{B} SMEs are selected to characterize a specific pixel. The selection strategy can be represented as

$$\boldsymbol{G}_{\mathrm{B}} = \sigma\left(\mathrm{Top}K_{\mathrm{B}}[\boldsymbol{g}]\right) \tag{3}$$

where $\sigma(\cdot)$ denotes the softmax function, and $\mathrm{Top}K_{\mathrm{B}}[\cdot]$ denotes selecting the top K_{B} maximum values and replacing the remaining values with $-\infty$. $\boldsymbol{G}_{\mathrm{B}} \in \mathbb{R}^{s^2\times n^{(\mathrm{B})}}$ denotes the activation values of SMEs. Through this mechanism, SMEs with higher contribution scores will be assigned larger weights, while SMEs with lower contribution scores will have their weights set to 0. Furthermore, a weighted aggregation of the selected SMEs is implemented using $\boldsymbol{G}_{\mathrm{B}}$

$$\boldsymbol{F}^{(\mathrm{H})} = \sum_{i=1}^{n^{(\mathrm{B})}} \left(\boldsymbol{G}_{\mathrm{B},i} \cdot \boldsymbol{E}_i\right) \tag{4}$$

where $\boldsymbol{F}^{(\mathrm{H})} \in \mathbb{R}^{s^2\times n_{\mathrm{E}^{(\mathrm{B})}}}$ represents the output features of MMoB.

B. Feature Fusion Based on MoMFE

HSI and LiDAR data can reflect various aspects of ground object features. By effectively fusing images from these two modalities, more discriminative features can be extracted, thereby enhancing classification accuracy. To overcome the

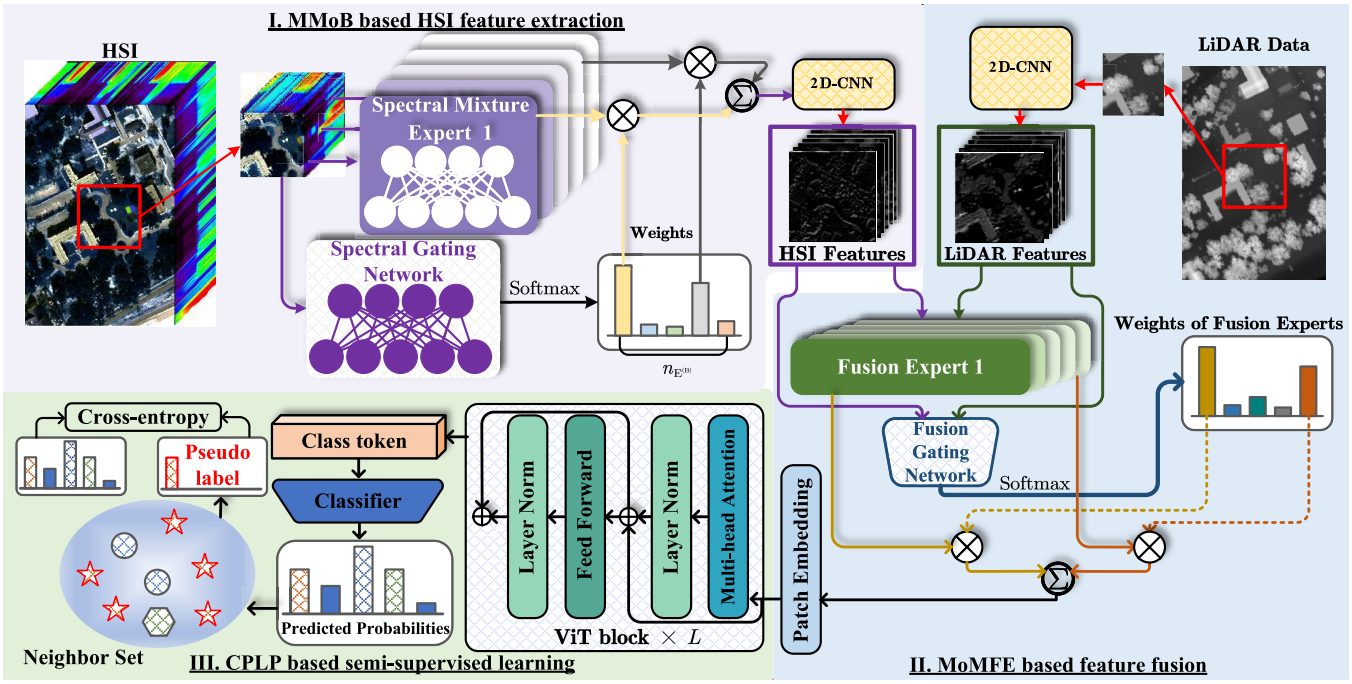

Fig. 1. Flowchart of joint classification of HSI and LiDAR based on the CPLP-DMoE.

limitations associated with individual HSI and LiDAR data fusion strategies, we propose the MoMFE to achieve diversified integration of HSI and LiDAR data. MoMFE consists of multiple fusion experts (FEs), and the computational process for each FE is as follows:

$$\boldsymbol{F}_i = \boldsymbol{\alpha}_i \odot \tilde{\boldsymbol{F}}^{(\mathrm{H})} + (1 - \boldsymbol{\alpha}_i) \odot \tilde{\boldsymbol{F}}^{(\mathrm{L})} \tag{5}$$

where $\tilde{\boldsymbol{F}}^{(\mathrm{H})} \in \mathbb{R}^{s\times s\times d_{\mathrm{M}}}$ represents the HSI features obtained after 2D-CNN processing, and $\tilde{\boldsymbol{F}}^{(\mathrm{L})} \in \mathbb{R}^{s\times s\times d_{\mathrm{M}}}$ represents the LiDAR features obtained after 2D-CNN processing. $\boldsymbol{\alpha}_i \in \mathbb{R}^{s\times s\times d_{\mathrm{M}}}$ denotes the ith FE, $\odot$ denotes the dot product operation, and $\boldsymbol{F}_i$ denotes the fusion result of the ith FE. By setting multiple FEs, different neighborhood representations correspond to various fusion strategies. For each neighborhood representation, its corresponding fusion strategy can be determined by selecting K_{F} FEs from n_{M} FEs based on the contribution scores provided by the gating network. The calculation method for the contribution score $\boldsymbol{h}$ corresponding to each FE is as follows:

$$\boldsymbol{h} = \left(\tilde{\boldsymbol{F}}^{(\mathrm{H})} + \tilde{\boldsymbol{F}}^{(\mathrm{L})}\right) \cdot \boldsymbol{W}_{\mathrm{G_M}} \tag{6}$$

where $\boldsymbol{W}_{\mathrm{G_M}} \in \mathbb{R}^{d_{\mathrm{M}}\times n_{\mathrm{M}}}$ represents the weights of the gating network in MoMFE. Then, K_{F} FEs can be selected using the following strategy:

$$\boldsymbol{G}_{\mathrm{M}} = \sigma(\mathrm{Top}K_{\mathrm{F}}[\boldsymbol{h}]) \tag{7}$$

where $\boldsymbol{G}_{\mathrm{M}} \in \mathbb{R}^{s^2\times n_{\mathrm{M}}}$ represents the activation values corresponding to the FEs. Then, FEs with higher contribution scores will be assigned larger weights, while FEs with lower contribution scores will be assigned a weight of 0. Finally, the fusion result $\boldsymbol{F}^{(\mathrm{M})}$ of multiple FEs can be obtained by performing a weighted sum of the selected FEs

$$\boldsymbol{F}^{(\mathrm{M})} = \sum_{i=1}^{n_{\mathrm{M}}} (\boldsymbol{G}_{\mathrm{M}} \cdot \boldsymbol{F}_i) \ . \tag{8}$$

Furthermore, the gating network often converges to assign large weights only to a few identical experts, leading to an over-reliance on a small subset of FEs while neglecting a large number of the other FEs. To mitigate this issue, we introduce information entropy to measure the selection balance of the FEs

$$H^{(\mathrm{M})} = -\sum_{i=1}^{n_{\mathrm{M}}} (\boldsymbol{P}_i \cdot \log \boldsymbol{P}_i) \tag{9}$$

where $\boldsymbol{P} = \sigma(\boldsymbol{h})$ denotes the probability distribution of the selected FEs. From (9), it can be observed that a small $H^{(\mathrm{M})}$ indicates that only a few FEs are associated with a high probability, while a large $H^{(\mathrm{M})}$ suggests that the selection probabilities across feature extractors are relatively balanced. Therefore, the gating network of MoMFE is encouraged to make more balanced selections through minimizing (9). The load balancing loss is defined as follows:

$$\mathcal{L}^{(\mathrm{M})} = \frac{1}{n^{(\mathrm{S})} + n^{(\mathrm{U})}} \sum_{j=1}^{n^{(\mathrm{S})}+n^{(\mathrm{U})}} H_j^{(\mathrm{M})} \ . \tag{10}$$

C. Feature Extraction Based on ViT

After obtaining the fusion features, we further utilize a ViT for deeper feature extraction. First, $\boldsymbol{F}^{(\mathrm{M})}$ is concatenated with a classification token $\boldsymbol{t}_{\mathrm{cls}} \in \mathbb{R}^{1\times d_{\mathrm{M}}}$ to obtain $\boldsymbol{h}_0 = [\boldsymbol{t}_{\mathrm{cls}}, \boldsymbol{F}^{(\mathrm{M})}] \in \mathbb{R}^{(s^2+1)\times d_{\mathrm{M}}}$. To preserve the spatial positional information between pixels [19], positional encoding is then

applied to $\boldsymbol{h}_0$, resulting in $\tilde{\boldsymbol{h}}_0 = \boldsymbol{h}_0 + \boldsymbol{t}_{\text{pos}}$. Then, the entire computation process of the ViT is as follows:

$$\tilde{\boldsymbol{h}}_l = \text{LN}\big(\tilde{\boldsymbol{h}}_{l-1} + \text{MHA}\big(\tilde{\boldsymbol{h}}_{l-1}\big)\big) \quad l \in \{1, 2, \ldots, L\} \tag{11}$$

where LN$(\cdot)$ and MHA$(\cdot)$ represent the layer normalization and multihead attention mechanisms, respectively. L denotes the number of self-attention modules in the ViT. The multihead attention is then obtained by concatenating multiple attention heads

$$\text{MHA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_h)\boldsymbol{W}_{\text{M}} \tag{12}$$

where Concat$(\cdot)$ represents the feature concatenation operation. $\boldsymbol{A}_i = \sigma(\boldsymbol{Q}_i \boldsymbol{K}_i^{\text{T}} / (d_{\text{k}})^{1/2})\boldsymbol{V}_i$ represents the output of the ith self-attention head, where $\boldsymbol{A}_i \in \mathbb{R}^{(s^2+1)\times d_{\text{v}}}$. $\boldsymbol{Q}_i$, $\boldsymbol{K}_i$, and $\boldsymbol{V}_i$ represent the query, key, and value of the ith self-attention head, respectively, which can be obtained through three mapping matrices $\boldsymbol{W}_i^{\text{q}} \in \mathbb{R}^{(d_{\text{M}}/h)\times d_{\text{q}}}$, $\boldsymbol{W}_i^{\text{k}} \in \mathbb{R}^{(d_{\text{M}}/h)\times d_{\text{k}}}$, and $\boldsymbol{W}_i^{\text{v}} \in \mathbb{R}^{(d_{\text{M}}/h)\times d_{\text{v}}}$, respectively. $(d_{\text{k}})^{1/2}$ is the scaling factor, h is the number of attention heads, and $\boldsymbol{W}_{\text{M}} \in \mathbb{R}^{(h\times d_{\text{k}})\times d_{\text{M}}}$ is the linear mapping matrix used to restore the output dimension of the multihead attention to d_{M}. After passing through multiple attention blocks, the output $\boldsymbol{F}^{(\text{T})} \in \mathbb{R}^{(s^2+1)\times d_{\text{M}}}$ of the ViT can be expressed as

$$\boldsymbol{F}^{(\text{T})} = \text{LN}\big(\tilde{\boldsymbol{h}}_L + \text{FFN}\big(\tilde{\boldsymbol{h}}_L\big)\big) \tag{13}$$

where FFN$(\cdot)$ represents the feedforward networks.

D. Semi-Supervised Learning Based on CPLP

To leverage the limited labeled samples and the abundant unlabeled samples to improve the performance of DMoE, we propose a semi-supervised learning method based on CPLP. First, pseudo labels for the unlabeled samples are generated through DMoE

$$\boldsymbol{p}_i = \sigma\Big(f_\Theta\Big(\mathcal{X}_i^{(\text{H})}, \mathcal{X}_i^{(\text{L})}\Big)\Big) = (p_{i1}, p_{i2}, \cdots, p_{iC}) \tag{14}$$

where $f_\Theta(\cdot)$ represents the feature transformation implemented by DMoE, $\boldsymbol{p}_i \in \mathbb{R}^C$ is the class probability vector predicted for the ith unlabeled sample, and C is the total number of classes. The pseudo labels cannot be fully trusted, as they may contain incorrect labels. Therefore, inspired by [37], we design a CPLP method to obtain more reliable pseudo labels. First, based on the class probability vector, the label set is filtered from the perspective of the sample by selecting the top K_{L} classes with the highest confidence. This process can be expressed as

$$\boldsymbol{p}_i' = \text{Top}K_{\text{L}}\big(\boldsymbol{p}_i\big) \tag{15}$$

where Top$K_{\text{L}}(\cdot)$ retains the largest K_{L} values and sets the remaining values to 0. According to (15), the candidate label set $\mathcal{C}_i = \{c | p_{i,c}' \neq 0\}$ for the ith sample can be obtained. However, the intrasample filtering not only relies heavily on the prediction accuracy of DMoE but also overlooks the correlations between different pixels. Generally, pixels with high feature similarity are more likely to belong to the same class. To this end, the K_{N} nearest neighbors with the highest similarity to the features of the ith sample are selected to form the auxiliary set $\mathcal{A}_i = \{\mathcal{C}_{i,1}, \ldots, \mathcal{C}_{i,K_{\text{L}}}\}$. Subsequently, the occurrence frequency of candidate labels is counted based on the auxiliary set. The process procedure can be expressed as

$$\text{count}(c) = \sum_{j=1}^{K_{\text{L}}} \mathbb{I}\big(c \in \mathcal{C}_{i,j}\big) \tag{16}$$

where $c \in \mathcal{C}_i$. A larger count(c) indicates that the candidate label appears more frequently in the auxiliary set, suggesting that this candidate label has a higher class consistency within the local neighbor of the sample, thereby exhibiting greater reliability. Furthermore, a threshold clipping strategy is employed to obtain the final pseudo labels set

$$\gamma_i = \tau\big(\big|\boldsymbol{p}_i'\big| - 1\big) \tag{17}$$

where τ is a hyperparameter, and $|\cdot|$ represents the number of elements to be selected from it. When count$(c) < \gamma_i$ is met, the class c will be removed from $\mathcal{C}_i$, resulting in the final pseudo label and the corresponding one-hot vector $\tilde{\boldsymbol{\mathcal{C}}}_i$. Thus, after applying CPLP, the filtered pseudo labels not only exhibit high confidence but also demonstrate greater class consistency within the local neighbor of the feature space, thereby achieving higher reliability.

As mentioned earlier, the loss function $\mathcal{L}$ of the proposed CPLP-DMoE consists of the classification loss $\mathcal{L}^{(\text{S})}$ on the labeled sample set, the classification loss $\mathcal{L}^{(\text{U})}$ on the unlabeled sample set, and the load balancing loss $\mathcal{L}^{(\text{M})}$

$$\mathcal{L} = \mathcal{L}^{(\text{S})} + \mathcal{L}^{(\text{U})} - \lambda\mathcal{L}^{(\text{M})} \tag{18}$$

where $\mathcal{L}^{(\text{S})} = (1/n^{(\text{S})}) \sum_{i=1}^{n^{(\text{S})}} \boldsymbol{y}_i \log(\tilde{\boldsymbol{y}}_i)$, $\mathcal{L}^{(\text{U})} = (1/n^{(\text{U})}) \sum_{i=1}^{n^{(\text{U})}} \boldsymbol{\mathcal{C}}_i \log(\tilde{\boldsymbol{y}}_i)$, and λ is the balancing coefficient.

III. Experiments and Analysis

A. Data Description

Experiments were conducted on three widely used HSI and LiDAR datasets to evaluate the performance of the CPLP-DMoE. The Houston 2013 dataset was collected by the National Center for Airborne Laser Mapping using the Compact Airborne Spectrographic Imager sensor over the University of Houston campus and its surrounding urban areas. It consists of an HSI and a LiDAR-based digital surface model (DSM), both with a spatial resolution of 2.5 m and a data size of 349 × 1905 pixels. The HSI contains 144 spectral bands, covering a wavelength range from 0.38 to 1.05 μm. The dataset includes a total of 15 029 ground-truth samples, spanning 15 classes [38]. The MUUFL dataset was collected in November 2010 at the Gulfport campus of the University of Southern Mississippi, Long Beach, MS, USA. This dataset includes campus HSI-based and LiDAR-based DSMs, with a size of 325 × 220 pixels. The HSI was collected using the ITRES Compact Airborne Spectrographic Imager (CASI-1500) sensor, which provides 64 spectral channels ranging from 0.38 to 1.05 μm, with a spatial resolution of 0.54 × 1.0 m. The LiDAR data were captured by an ALTM sensor using a laser with a wavelength of 1064 nm, offering a spatial resolution of 0.60 × 0.78 m. The dataset comprises a total of 53 687 ground-truth samples and studies 11 distinguishable

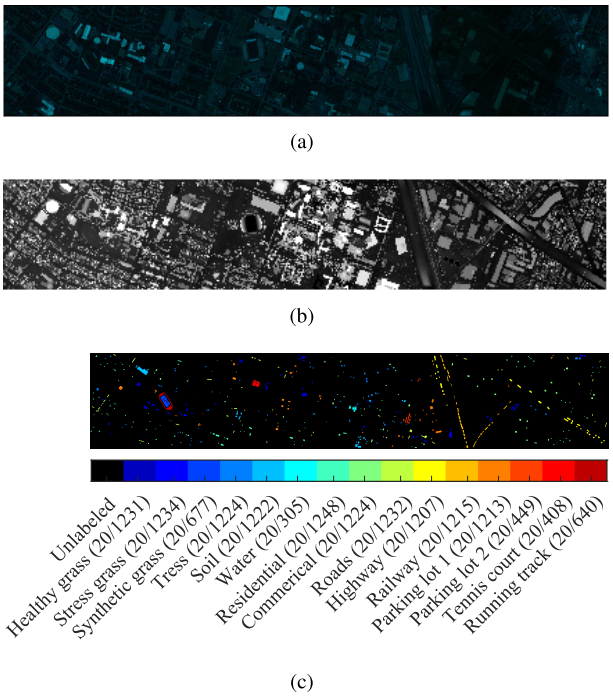

Fig. 2. Houston 2013 dataset. (a) Pseudo-color image for HSI. (b) Grayscale image for LiDAR-based DSM. (c) Ground-truth map.

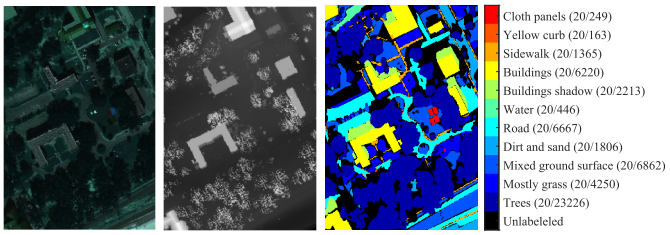

Fig. 3. MUUFL dataset. (a) Pseudo-color image for HSI. (b) Grayscale image for LiDAR-based DSM. (c) Ground-truth map.

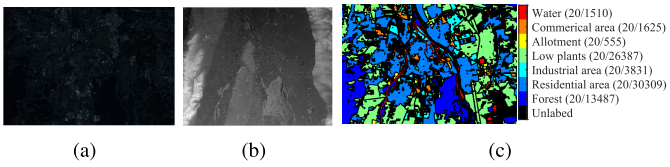

Fig. 4. Augsburg dataset. (a) Pseudo-color image for HSI. (b) Grayscale image for LiDAR-based DSM. (c) Ground-truth map.

class labels [39]. The Augsburg dataset was captured over the city of Augsburg, Germany. The HSI data were acquired using the DAS-EOC HySpex sensor [40], while the LiDAR-based DSM data were collected by the DLR-3K system. Both types of imagery were downsampled to a uniform spatial resolution of 30 m to facilitate effective multimodal fusion. The dataset size is 332 × 485 pixels. It contains a total of 78 294 ground-truth samples and describes seven distinct land cover classes. Figs. 2–4 present the pseudo-color image for HSI, the grayscale image for LiDAR-based DSM, and the ground-truth map, along with sample information from the three datasets, respectively.

B. Experimental Setting

Five evaluation metrics were selected to evaluate the performance of different methods, including accuracy for each class, overall accuracy (OA), average accuracy (AA), the Kappa coefficient, and consumed time. To minimize the impact of random factors, all the experimental results were averaged over five classification runs. All classification methods were implemented on the PyTorch platform and executed on a 3.80-GHz Intel Core i7-10700KF CPU, 32-GB RAM, and an RTX 2080Ti GPU. The Adam optimizer was employed with a learning rate of 0.001, and the batch size and the number of training epochs were set to 32 and 500, respectively. To facilitate fusion, the number of convolutional kernels, kernel size, stride, and padding for the two CNNs were set to be the same, specifically 80, 3 × 3, 1, and 1, respectively. The input dimensions for the CNN that extracts features from HSI for the Houston 2013, MUUFL, and Augsburg datasets were 11 × 11 × 144, 11 × 11 × 64, and 11 × 11 × 180, respectively. For the CNN that extracts features from LiDAR data, the input dimensions for the Houston 2013, MUUFL, and Augsburg datasets were 11 × 11 × 1, 11 × 11 × 2, and 11 × 11 × 1, respectively. Then, the dimensions of the output for both modalities on the three datasets are $11 \times 11 \times 80$. The hyperparameters of the MMoB, MoMFE, and CPLP network structures are presented in Table I. On all three datasets, the number of SMEs was set to 60, K_{B} was set to 20, the number of FEs and K_{F} were set to 40 and 2, respectively, the number and size of the convolutional kernels were set to 80 and 3 × 3, respectively, and L and λ were set to 3 and 0.001, respectively. The sizes of the SMEs were configured as 60 × 144, 60 × 64, and 60 × 180 on the Houston 2013, MUUFL, Augsburg datasets, respectively. On the Houston 2013, MUUFL, and Augsburg datasets, K_{L} was set to 10, 7, and 5, respectively.

C. Comparative Experiments

To validate the effectiveness of CPLP-DMoE, the following methods were selected for the comparative experiments, including coupled CNN [41], extended ViT (ExViT) [42], nearest neighbor-based contrastive learning network (NNCNet) [43], dynamic scale hierarchical fusion network (DSHFNet) [44], hierarchical CNN and transformer (HCT) [28], multimodal fusion transformer (MFT) [45], and multimodal fusion network (M2FNet) [10]. The hyperparameters for these comparative methods were set according to their respective references. Tables II–IV present the classification performance of different methods on the Houston 2013, MUUFL, and Augsburg datasets. The best results are highlighted in bold for emphasis. From Figs. 5–7 and Tables II–IV, the following can be observed.

1) Compared to methods based on a single fusion strategy (all other comparative methods except CPLP-DMoE), CPLP-DMoE, which is equipped with multiple fusion strategies, can achieve higher OA. This is because CPLP-DMoE can offer diverse and differentiated fusion strategies for different samples, thereby avoiding over-reliance on a single fusion strategy.

TABLE I
HYPERPARAMETERS IN MMOB, MOMFE, AND CPLP

Structure of type	Number of experts	Experts size	Value of Top$K$	Value of K_{N}	τ
MMoB	60	$60 \times 144/64/180$	20	-	-
MoMFE	40	-	2	-	-
CPLP	-	-	10 / 7 / 5	2	0.2

TABLE II
COMPARISON OF CLASSIFICATION PERFORMANCE (HOUSTON 2013 DATASET)

	Coupled CNN [41]	ExViT [42]	NNCNet [43]	DSHFNet [44]	HCT [28]	MFT [45]	M2FNet [10]	CPLP-DMoE
Healthy grass (%)	94.04	91.53	93.49	89.79	**95.89**	91.94	93.18	93.31
Stress grass (%)	92.41	95.25	95.27	96.16	**97.62**	89.69	92.49	94.70
Synthetic grass (%)	98.85	99.20	**99.88**	99.53	99.38	99.64	98.94	99.79
Tress (%)	**99.31**	97.79	96.21	96.69	98.09	96.93	95.68	96.99
Soil (%)	99.66	99.51	99.89	99.53	**99.98**	99.44	96.67	99.93
Water (%)	97.84	94.69	**98.76**	91.87	95.74	93.70	94.69	94.23
Residential (%)	89.92	**94.36**	89.54	94.02	90.77	89.14	91.18	93.88
Commercial (%)	84.85	78.37	75.60	76.88	82.19	74.18	77.71	**87.42**
Road (%)	77.06	85.67	83.31	54.80	78.79	83.82	84.06	**90.91**
Highway (%)	85.34	88.18	91.17	88.35	89.74	90.61	83.96	**93.09**
Railway (%)	95.60	97.48	**97.51**	94.35	97.09	95.46	95.69	96.13
Parking lot 1 (%)	92.05	92.05	91.77	93.74	**94.64**	91.87	89.73	93.88
Parking lot 2 (%)	97.37	97.24	**97.99**	91.76	97.15	94.29	93.58	95.15
Tennis court (%)	**100**	99.31	**100**	99.51	99.55	**100**	96.81	**100**
Running track (%)	**100**	99.75	99.97	**100**	**100**	99.81	99.50	**100**
OA (%)	92.35	93.11	92.95	89.30	93.53	91.58	91.25	**94.75**
AA (%)	93.62	94.02	94.21	90.63	94.47	92.64	92.26	**95.29**
Kappa (%)	91.73	92.54	92.38	88.14	92.98	90.90	90.54	**94.32**
Time (s)	**25.23**	85.32	605.82	338.47	285.56	72.15	31.08	46.10

TABLE III
COMPARISON OF CLASSIFICATION PERFORMANCE (MUUFL DATASET)

	Coupled CNN [41]	ExViT [42]	NNCNet [43]	DSHFNet [44]	HCT [28]	MFT [45]	M2FNet [10]	CPLP-DMoE
Tress (%)	82.69	85.13	78.11	78.98	83.20	81.69	87.36	**93.63**
Mostly grass (%)	81.64	75.88	**85.73**	78.74	75.37	72.51	76.05	83.08
Mixed ground surface (%)	57.54	59.93	46.79	**71.73**	61.60	62.03	51.79	64.97
Dirt and sand (%)	91.38	90.10	**91.66**	90.64	87.43	84.65	78.69	86.05
Road (%)	**80.61**	76.30	77.01	71.13	72.89	76.29	75.63	80.46
Water (%)	95.16	98.75	96.28	**99.73**	99.42	99.19	98.56	98.97
Buildings shadow (%)	82.27	81.71	74.55	**91.67**	85.60	79.93	83.22	78.05
Bulidings (%)	92.14	**93.47**	90.54	88.52	90.48	88.06	89.56	91.51
Sidewalk (%)	**67.48**	56.02	58.05	48.45	55.24	53.69	57.29	46.89
Yellow curb (%)	78.16	71.90	**84.42**	40.24	82.33	76.93	75.58	70.55
Cloth panels (%)	96.06	95.98	**98.63**	95.90	96.78	95.09	95.58	95.50
OA (%)	80.11	80.51	76.60	78.44	79.09	78.01	79.55	**85.10**
AA (%)	**81.74**	80.48	78.84	77.79	80.94	79.10	79.03	80.88
Kappa (%)	74.78	75.06	70.17	72.90	73.38	72.14	73.79	**80.49**
Time (s)	**23.51**	198.67	483.14	596.46	32.11	36.47	216.24	85.10

TABLE IV
COMPARISON OF CLASSIFICATION PERFORMANCE (AUGSBURG DATASET)

	Coupled CNN [41]	ExViT [42]	NNCNet [43]	DSHFNet [44]	HCT [28]	MFT [45]	M2FNet [10]	CPLP-DMoE
Forest (%)	90.43	94.11	93.85	79.05	**95.47**	95.04	94.43	95.35
Residential area (%)	81.36	79.27	77.75	59.09	74.59	76.19	70.74	**92.19**
Industrial area (%)	56.15	52.79	59.23	25.07	50.51	**78.42**	61.41	55.73
Low plants (%)	85.84	75.47	84.79	81.61	82.98	**95.04**	77.69	94.87
Allotment (%)	80.24	82.71	**91.89**	86.27	88.97	81.33	82.38	80.33
Commercial area (%)	65.11	63.95	55.74	**77.78**	57.51	58.25	58.49	54.18
Water (%)	71.52	71.14	70.65	69.22	66.87	**88.43**	63.64	63.42
OA (%)	84.77	78.78	81.54	69.38	79.49	84.23	76.49	**90.44**
AA (%)	75.80	74.21	76.27	68.31	73.84	76.19	72.67	**76.58**
Kappa (%)	79.13	71.21	75.09	60.19	72.85	78.42	68.49	**86.49**
Time (s)	**42.93**	193.88	367.80	736.08	165.38	61.71	691.76	312.17

2) Compared to the Houston 2013 and Augsburg datasets, all methods achieve lower OA on the MUUFL dataset. This is due to the more complex data distribution and higher interclass similarity in the MUUFL dataset. However, even on the MUUFL dataset, the CPLP-DMoE method can still achieve the highest OA.

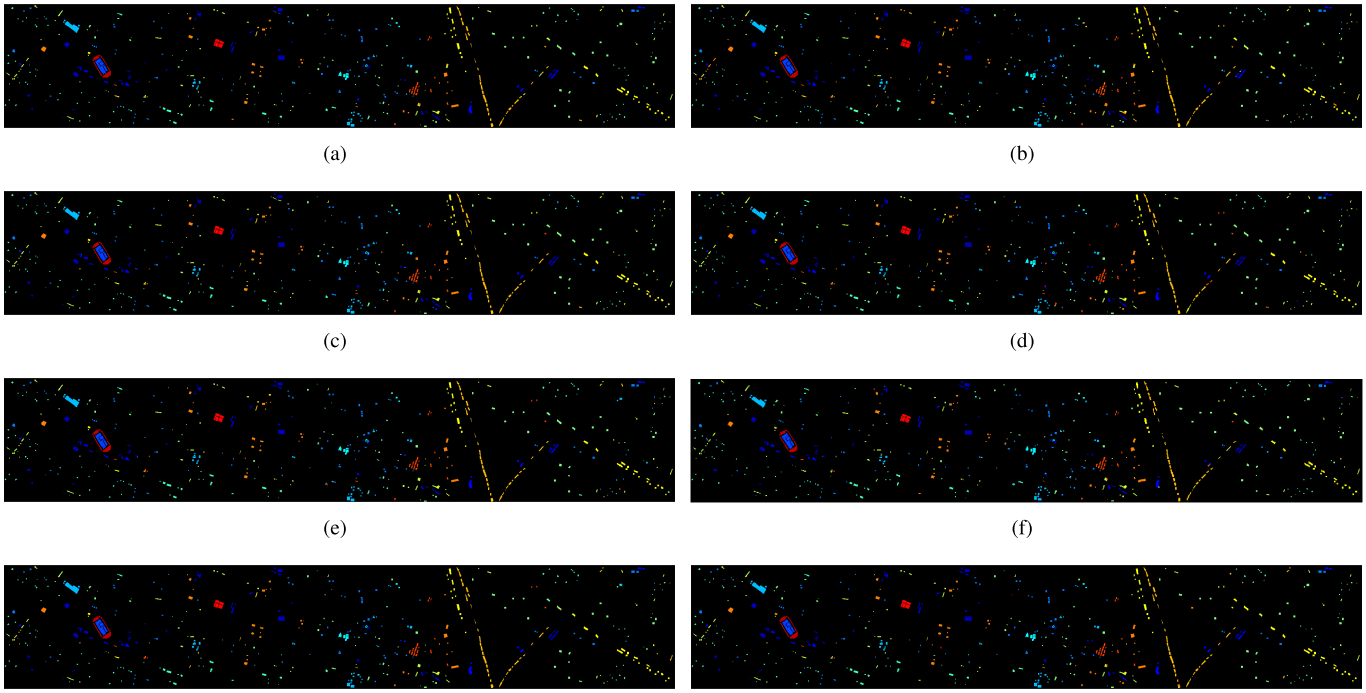

Fig. 5. Classification maps of different methods on the Houston 2013 dataset. (a) Coupled CNN. (b) DSHFNet. (c) ExViT. (d) MFT. (e) HCT. (f) M2FNet. (g) NNCNet. (h) CPLP-DMoE.

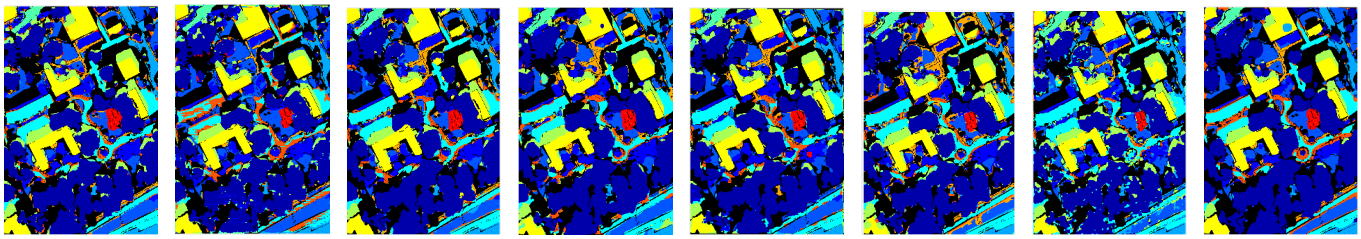

Fig. 6. Classification maps of different methods on the MUUFL dataset. (a) Coupled CNN. (b) DSHFNet. (c) ExViT. (d) MFT. (e) HCT. (f) M2FNet. (g) NNCNet. (h) CPLP-DMoE.

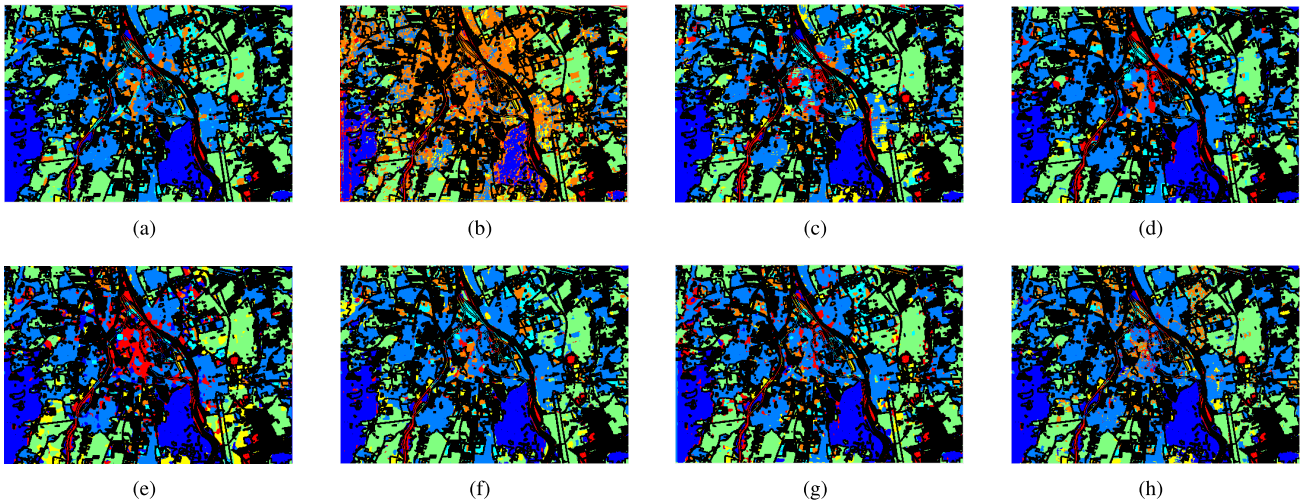

Fig. 7. Classification maps of different methods on the Augsburg dataset. (a) Coupled CNN. (b) DSHFNet. (c) ExViT. (d) MFT. (e) HCT. (f) M2FNet. (g) NNCNet. (h) CPLP-DMoE.

3) On all three datasets, CPLP-DMoE achieves the highest OA compared to other methods. Taking the MUUFL dataset as an example, CPLP-DMoE attains the highest or near-highest OA in most classes. Its advantage is particularly notable in the tree and grassland classes. This is because the spectral signatures of trees and grasslands

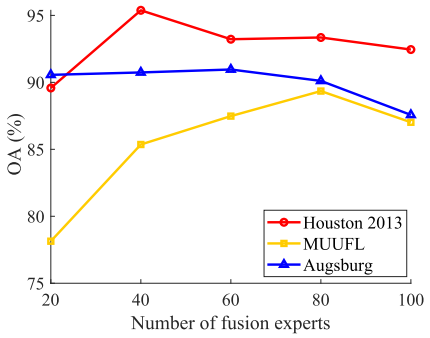

Fig. 8. OA versus number of FEs.

are quite similar, making classification challenging. Relying solely on HSI would result in the mixing of these two classes. Fortunately, these classes have distinct elevation information. Therefore, CPLP-DMoE, with its multiple fusion strategies, can extract more discriminative fusion features compared to single fusion strategy-based methods.

4) Figs. 5–7 show the classification maps obtained by all the methods mentioned above. From Figs. 5–7, it is evident that the classification maps obtained by CPLP-DMoE are smoother, more detailed, and exhibit clearer boundaries. For example, on the MUUFL dataset, the boundary regions between the tree and grassland classes are very distinct with CPLP-DMoE, whereas other methods show noticeable confusion in these areas. Similarly, on the Houston 2013 and Augsburg datasets, the classification maps demonstrate that boundary regions are clearer and less affected by neighboring areas. This leads to the conclusion that CPLP-DMoE produces the most precise boundary regions, further validating the effectiveness of this method.
5) Among all the comparison methods, the approach that uses only CNN (couple CNN) consumes less time than the other methods that combine transformers with CNNs. This is mainly due to the quadratic computational complexity of the attention mechanism in transformers. Among all the classification methods, CPLP-DMoE does not have the longest or the shortest time consumption, and it falls within an acceptable range.

D. Parameter Analysis

First, we analyze the impact of the number of FEs on the OA of CPLP-DMoE. As shown in Fig. 8, where the range for the number of FEs is {20, 40, 60, 80, 100}. Different numbers of FEs represent different numbers of fusion strategies, which are crucial for the subsequent fusion of HSI and LiDAR data features. On the one hand, when the number of selected FEs is too small, it fails to achieve a diverse fusion of HSI and LiDAR data features. On the other hand, selecting too many FEs increases the risk of overfitting. In this article, the number of FEs is set to 40 on the Houston 2013 dataset, 80 on the MUUFL dataset, and 60 on the Augsburg dataset.

Second, we analyze the relationship between OA and the number of labeled samples per class, as shown in Fig. 9. The range for the number of labeled samples per class

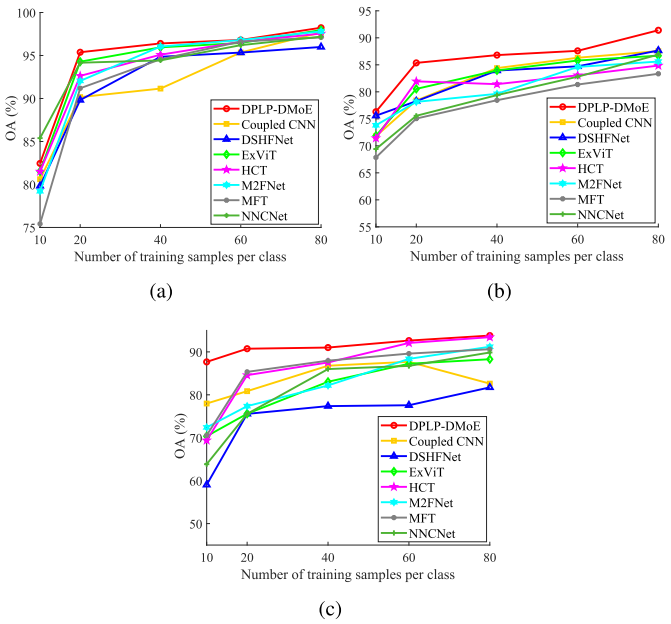

Fig. 9. Relationship between OA and the number of training samples on different datasets. (a) Houston 2013. (b) MUUFL. (c) Augsburg.

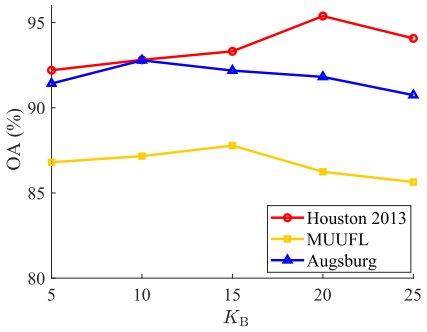

Fig. 10. OA versus K_B.

is {10, 20, 40, 60, 80}. As the number of training samples increases, the OA generally exhibits an increasing trend. This is because more labeled samples provide additional supervised information, enabling the semi-supervised CPLP-DMoE to achieve higher OA.

Third, we analyze the impact of different values of K_B on the OA of CPLP-DMoE. As shown in Fig. 10, the range of selected values for K_B is {5, 10, 15, 20, 25}. Different values of K_B correspond to the selection of different numbers of SMEs. On the one hand, when the value of K_B is too small, the model may fail to sufficiently learn the features of pixels from different spatial locations and classes. On the other hand, when the value of K_B is too large, it increases the risk of overfitting and increases computational costs. As shown in Fig. 10, when the value of K_B is 20, 15, and 10, CPLP-DMoE achieves the highest OA on the Houston 2013, MUUFL, and Augsburg datasets, respectively.

Fourth, we analyze the impact of different values for K_F on the OA of CPLP-DMoE. As shown in Fig. 11, the range of selected values of K_F is {2, 4, 6, 8, 10}. Different values of K_F correspond to the selection of different numbers of FEs. On the one hand, when the value of K_F is too small, there are fewer types of feature fusion strategies for HSI

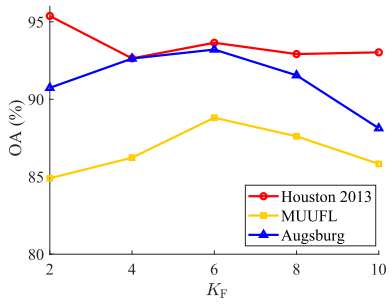

Fig. 11. OA versus K_F.

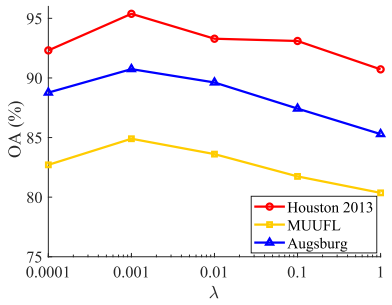

Fig. 12. OA versus λ.

and LiDAR data, which leads to a one-sided and suboptimal fusion model. On the other hand, when the value of K_F is too large, it similarly increases the risk of overfitting and raises computational costs. As shown in Fig. 11, when the value of K_F is 2, 6, and 6, CPLP-DMoE achieves the highest OA on the Houston 2013, MUUFL, and Augsburg datasets, respectively.

Finally, we analyze the impact of different values of λ on the OA of CPLP-DMoE. As shown in Fig. 12, the range of selected values for λ is $\{0.0001, 0.001, 0.01, 0.1, 1\}$. λ represents the weight assigned to the load balancing loss. On the one hand, when the value of λ is too large, the model focuses excessively on the balanced use of experts, which may prevent the model from fully leveraging the most effective experts. On the other hand, when the value of λ is too small, the model struggles to ensure balanced usage of FEs, leading to the overuse of a few experts while neglecting usable information from others. It can be observed that on the Houston 2013, MUUFL, and Augsburg datasets, the highest OA occurs when the value of λ is 0.001.

E. Ablation Studies

To validate the impact of each component of CPLP-DMoE on OA, ablation experiments were conducted on the three datasets, as shown in Fig. 13. When the MoMFE is not included in the experiments, the fusion of HSI and LiDAR data features is performed directly using (5). It can be observed that CPLP-DMoE can achieve the highest OA. This is because MMoB can assign larger weights to bands that are more important for the classification task. After adding the MoMFE, the fusion of HSI and LiDAR data transitions from a single strategy to a multi-strategy method, thereby facilitating a more diverse fusion of features from both modalities. Moreover, after adding the CPLP, a large number of unlabeled samples are utilized, ensuring sufficient training of network parameters.

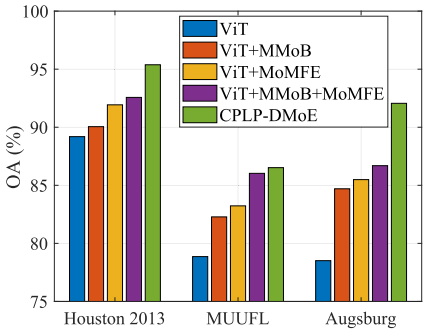

Fig. 13. Ablation study.

IV. Conclusion

To tackle the challenges of redundancy of HSI bands, the single fusion strategies for multimodal data, and the high cost of pixelwise labeling in remote sensing images, we propose CPLP-DMoE for the joint classification of HSI and LiDAR data. First, the dense mixture of spectral bands in HSI is achieved by assigning greater weight coefficients to bands that are more crucial for downstream tasks. This not only mitigates the impact of redundant bands but also prevents the loss of spectral information that often occurs with conventional sparse MoE. Subsequently, multiple FEs are configured in conjunction with a gating-based selection and mixing mechanism to achieve diverse feature fusion of HSI and LiDAR data. Additionally, the introduction of information entropy balances the selection of FEs. Finally, by introducing CPLP, more reliable pseudo labels are assigned to a large number of unlabeled samples, thereby facilitating the learning of a more accurate class discrimination model. The experimental results on the Houston 2013, MUUFL, and Augsburg datasets demonstrate the effectiveness of CPLP-DMoE. In future work, it is valuable to take into account the positional relationships of neighboring pixels during intersample selection.

References

[1] C. Laukamp, "Geological mapping using mineral absorption feature-guided band-ratios applied to prisma satellite hyperspectral level 2D imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 5981–5984.

[2] J. Li and Z. Ou, "Construction of urban ecological environment detection system based on GIS and RS image processing algorithm," in *Proc. Int. Conf. Netw., Informat. Comput. (ICNETIC)*, May 2023, pp. 684–688.

[3] S. Li, Q. Liu, Z. Li, Z. Qi, L. Si, and N. Wang, "Forest canopy gap dynamics based on time-series of airborne LiDAR data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 6119–6121.

[4] A. N. J. Kukunuri and D. Singh, "Efficient application of drone with satellite data for early-stage wheat detection: For precision agriculture monitoring," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 4388–4391.

[5] P. Duan, X. Kang, P. Ghamisi, and S. Li, "Hyperspectral remote sensing benchmark database for oil spill detection with an isolation forest-guided unsupervised detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5509711.

[6] R. Hang, S. Xu, and Q. Liu, "A regionally indicated visual grounding network for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5647411.

[7] J. Li, K. Zheng, W. Liu, Z. Li, H. Yu, and L. Ni, "Model-guided coarse-to-fine fusion network for unsupervised hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[8] P. Duan, T. Shan, X. Kang, and S. Li, "Spectral super-resolution in frequency domain," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 29, 2024, doi: 10.1109/TNNLS.2024.3481060.
[9] J. Li, K. Zheng, Z. Li, L. Gao, and X. Jia, "X-shaped interactive autoencoders with cross-modality mutual learning for unsupervised hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5518317.
[10] L. Sun, X. Wang, Y. Zheng, Z. Wu, and L. Fu, "Multiscale 3-D–2-D mixed CNN and lightweight attention-free transformer for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 2100116.
[11] 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.
[12] J. Jung, E. Pasolli, S. Prasad, J. C. Tilton, and M. M. Crawford, "A framework for land cover classification using discrete return LiDAR data: Adopting pseudo-waveform and hierarchical segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 2, pp. 491–502, Jan. 2014.
[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] 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.
[15] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct. 2021.
[16] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.
[17] T. Chakraborty and U. Trehan, "SpectralNET: Exploring spatial–spectral WaveletCNN for hyperspectral image classification," 2021, *arXiv:2104.00341*.
[18] Z. Zheng, Y. Zhong, A. Ma, and L. Zhang, "FPGA: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5612–5626, Aug. 2020.
[19] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021, pp. 1–21.
[20] 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. 3144158.
[21] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3171551.
[22] X. He, Y. Chen, and Z. Lin, "Spatial–spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, p. 498, Jan. 2021.
[23] H. Lin, R. Hang, S. Wang, and Q. Liu, "DiFormer: A difference transformer network for remote sensing change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024, Art. no. 6003905.
[24] R. Hang, X. Qian, and Q. Liu, "Cross-modality contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532812.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.
[30] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2017, pp. 1–19.
[31] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2023, pp. 1–11.
[32] P. Duan, Z. Xie, X. Kang, and S. Li, "Self-supervised learning-based oil spill detection of hyperspectral images," *Sci. China Technolog. Sci.*, vol. 65, no. 4, pp. 793–801, Apr. 2022.
[33] Y. Ding, M. Hou, Y. Ding, C.-H. Zheng, Q. Yan, and D.-S. Huang, "Exploring positional distributions of labeled superpixels within graph convolutional networks for hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5535115.
[34] M. Zhao, J. Liu, and Y. Wu, "A semi-supervised image registration framework based on multimodal cross-attention," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024, Art. no. 5002905.
[35] K. Ding, T. Lu, and S. Li, "Uncertainty-aware contrastive learning for semi-supervised classification of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5519413.
[36] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.
[37] S. He, C. Wang, G. Yang, and L. Feng, "Candidate label set pruning: A data-centric perspective for deep partial-label learning," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2024, pp. 1–12.
[38] 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.
[39] 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.
[40] A. Baumgartner, P. Gege, C. Köhler, K. Lenhard, and T. Schwarzmaier, "Characterisation methods for the hyperspectral sensor HySpex at DLR's calibration home base," *Proc. SPIE*, vol. 8533, pp. 371–378, Nov. 2012.
[41] 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.
[42] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.
[43] M. Wang, F. Gao, J. Dong, H.-C. Li, and Q. Du, "Nearest neighbor-based contrastive learning for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501816.
[44] 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.
[45] 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.

Yi Kong (Member, IEEE) received the Ph.D. degree from China University of Mining and Technology, Xuzhou, China, in 2019.

He is currently an Associate Professor with the School of Information and Control Engineering, China University of Mining and Technology. His main research interests include remote sensing image analysis.

Shaocai Yu received the B.E. degree in building electricity and intelligence from Anhui Jianzhu University, Hefei, China, in 2022. He is currently pursuing the M.E. degree with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China.

His main research interests include remote sensing image analysis.

Yuhu Cheng (Senior Member, IEEE) received the Ph.D. degree in control science and technology from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently a Professor with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China. His main research interests include machine learning and intelligent systems.

C. L. Philip Chen (Life Fellow, IEEE) received the M.E. degree from the University of Michigan, Ann Arbor, MI, USA, in 1985, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 1988.

He is currently the Chair Professor and the Dean of the College of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Being a Program Evaluator of the Accreditation Board of Engineering and Technology Education (ABET), Baltimore, MD, USA, for computer engineering, electrical engineering, and software engineering programs, he successfully architects the University of Macau's engineering and computer science programs receiving accreditations from Washington/Seoul Accord through Hong Kong Institute of Engineers (HKIE), Hong Kong, of which is considered as his utmost contribution in engineering/computer science education for Macau as the former Dean of the Faculty of Science and Technology. His research interests include cybernetics, systems, and computational intelligence.

Dr. Chen is a fellow of AAAS, IAPR, CAA, and HKIE, and a member of the Academia Europaea (AE), European Academy of Sciences and Arts (EASA), and the International Academy of Systems and Cybernetics Science (IASCYS). He was a recipient of the 2016 Outstanding Electrical and Computer Engineers Award from his alma mater, Purdue University, in 1988. He received IEEE Norbert Wiener Award in 2018 for his contribution to systems and cybernetics, and machine learning. He is also a highly cited researcher by Clarivate Analytics in 2018 and 2019. He was the Chair of TC 9.1 Economic and Business Systems of International Federation of Automatic Control from 2015 to 2017. He was the IEEE Systems, Man, and Cybernetics Society President from 2012 to 2013, the Editor-in-Chief of IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS from 2014 to 2019 and IEEE TRANSACTIONS ON CYBERNETICS from 2020 to 2021, and an Associate Editor of IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE and IEEE TRANSACTIONS ON FUZZY SYSTEMS.

Xuesong Wang (Member, IEEE) received the Ph.D. degree in control science and technology from China University of Mining and Technology, Xuzhou, China, in 2002.

She is currently the Dean of the Engineering Research Center of Intelligent Control for Underground Space, Ministry of Education, China University of Mining and Technology, and also a Professor with the School of Information and Control Engineering. Her main research interests include machine learning and image processing.

Dr. Wang is a Senior Associate Editor of IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS and an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE SIGNAL PROCESSING LETTERS, *International Journal of Machine Learning and Cybernetics*, *Acta Automatica Sinica*, and *Acta Electronica Sinica*.