





MHST: Multiscale Head Selection Transformer for Hyperspectral and LiDAR Classification

Kang Ni , Member, IEEE, Duo Wang , Zhizhong Zheng , and Peng Wang , Senior Member, IEEE

Abstract—The joint use of hyperspectral image (HSI) and light detection and ranging (LiDAR) data has gained significant performance on land-cover classification. Although spatial–spectral feature learning methods based on convolutional neural networks and transformer networks have achieved prominent advances, contextual information described by fixed convolutional kernels and all self-attention heads selected have limited ability to characterize the detailed information and nonredundant features of land-covers on multimodal data. In this article, a multiscale head selection transformer (MHST) network, is proposed to fully explore detailed and nonredundant features in spatial and spectral dimensions of HSI and LiDAR data. To better acquire detailed information of spatial and spectral features at different scales, a multiscale spectral–spatial feature extraction module, including cascaded multiscale 3-D and 2-D convolutional layers, is inserted into MHST. Simultaneously, an adaptive global feature extraction module based on head selection pooling transformer is given after transformer encoder module for alleviating token redundancy in an adaptive computation style. Finally, we develop a multimodal–multiscale feature fusion classification module with local features

and global class token, to exploit a powerful global–local fuse style. The extensive experiments on three popular datasets demonstrate that MHST significantly outperforms other related networks.

Index Terms—Classification, feature learning, global class token, hyperspectral image (HSI), light detection and ranging (LiDAR) data, transformer.

I. INTRODUCTION

HYPERSPECTRAL image (HSI) [1], [2], [3], [4], [5], one of remote sensing data, has been widely used in several applications [6], [7], [8], [9], [10] related to land-cover mapping, target detection, mineral exploration, etc., due to its rich spectral information, which can accurately reflect the spectral reflection characteristics of the surface of ground objects [11]. Nevertheless, with more types of land covers with complex structures, single remote sensing image has been unable to meet the requirements of high precision, such as ground objects with similar spectral characteristics and different elevation information. Light detection and ranging (LiDAR) data or digital surface model (DSM) could provide the object height information of Earth surface [12], [13], [14], then the integration of HSI and LiDAR data opens up the possibility to enhance the land-cover classification performance [15], [16], [17] by multimodal feature fusion and interaction.

Despite these advantages of HSI and LiDAR data, there remain some unique technical challenges in land-cover classification as follows that significantly constraining its applicability.

- 1) For data characteristics, the scale variation of land covers makes it difficult to accurately depict the local characteristics of land covers.
- 2) For feature learning, given the comprehensive consideration of feature redundancy, the global sequence properties of HSI spectral features and LiDAR data, limits the improvement in classification accuracy.

Due to rapid development of deep learning, a multitude of deep neural networks, e.g., convolutional neural networks (CNNs), have exhibited significant promise in joint hyperspectral and LiDAR classification [18], [19], [20], [21] due to their powerful ability to extract local features. Hang et al. [22] utilized coupled CNNs, feature-level fusion and decision-level fusion strategies for acquiring the distinguishable features from HSI and LiDAR data. Zhang et al. [23] proposed an interleaving perception CNN, which is an information fusion CNNs, for classifying land covers via hyperspectral and LiDAR data. Moreover, emphasizing the challenge of addressing weak boundaries and spatially fragmented classification, a dual-tunnel CNNs and

Manuscript received 12 December 2023; revised 22 January 2024; accepted 9 February 2024. Date of publication 16 February 2024; date of current version 4 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62101280, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20210588, in part by Project funded by China Postdoctoral Science Foundation under Grant 2023M731781, in part by the Key Laboratory of Radar Imaging and Microwave Photonics, Nanjing University of Aeronautics and Astronautics, Ministry of Education under Grant NJ20230005, in part by the Open Foundations of Jiangsu Province Engineering Research Center of Airborne Detecting and Intelligent Perceptive Technology under Grant JSECF2023-01 and Grant JSECF2023-05, in part by Jiangsu Geological Bureau Research Project under Grant 2021KY14 and Grant 2023KY11, and in part by the Nanjing University of Posts and Telecommunications Science Foundation (NUPTSF) under Grant NY222107. (Kang Ni and Duo Wang contributed equally to this work.) (Corresponding author: Zhizhong Zheng.)

Kang Ni is with the School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, also with the Jiangsu Province Engineering Research Center of Airborne Detecting and Intelligent Perceptive Technology, Nanjing 210049, China, and also with the Key Laboratory of Radar Imaging and Microwave Photonics (Nanjing University of Aeronautics and Astronautics), Ministry of Education, Nanjing 211106, China (e-mail: tznikang@163.com).

Duo Wang is with the College of Automation and College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: njwangduo@163.com).

Zhizhong Zheng is with the School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China, and also with the Jiangsu Province Engineering Research Center of Airborne Detecting and Intelligent Perceptive Technology, Nanjing 210049, China (e-mail: zhengzz_js@126.com).

Peng Wang is with the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China, and also with the Jiangsu Province Engineering Research Center of Airborne Detecting and Intelligent Perceptive Technology, Nanjing 210049, China (e-mail: Pengwang-B614080003@hotmail.com).

Code will be available online. <https://github.com/RSIP-NJUPT/MHST>
Digital Object Identifier 10.1109/JSTARS.2024.3366614

hierarchical random walk layer [24] were given and significantly enhanced classification performance of HSI and LiDAR data. Nevertheless, for land covers with different scales and complex terrain structure, multiple scale data blocks can be used as network inputs, e.g., global–local transformer network (GLT-Net) [25], or multiscale convolution operation is an effective method. This enables the simultaneous utilization of distinguishable local information at different levels, thus leading to a better understanding of high-level semantic features within HSI and LiDAR data.

For global sequence feature learning, transformer has a significant advantage in capturing long-term dependencies and global deep features. In the collaborative land cover classification using HSI and LiDAR, constructing long-range dependencies can effectively capture spectral information and global information of land cover, such as binary-tree transformer network [26] and parallel transformer [27]. Particularly, for multisource remote sensing feature fusion learning, feature redundancy affects the model discriminability. Then, a local information interaction transformer model was proposed by Zhang et al. [28] for mining the complementary information and data imbalance problem of HSI-LiDAR data, simultaneously, this proposed model reduces the redundant information via dynamically filtering source components. For addressing the limitations and gaps in the newly acquired Earth observation data from a single source data, Feng et al. [29] inserted two effective modules into spectral–spatial–elevation fusion transformer, it is worth noting that this proposed transformer network could reduce redundant spatial information. Currently, despite some existing transformer-based methods considering the redundancy of spatial features, most of them do not consider the redundancy of global sequence features for HSI and LiDAR data. Therefore, starting from the characteristics of transformer models, it is of great significance to construct an adaptive feature selection mechanism for extracting global sequence properties of HSI and LiDAR data.

To address the challenge of characterizing the detailed local information and nonredundant global sequence properties of land-covers on multimodal data, followed by CNNs-transformer feature learning style, a multiscale head selection transformer (MHST) network is proposed. For the spatial and spectral feature information of HSI images, cascading multiple multiscale 2-D convolutions and 3-D convolutions are utilized to fully capture the spatial–spectral detail information of HSI images. For LiDAR data, multiple multiscale 2-D convolutions are employed to fully capture the elevation spatial information of LiDAR. Furthermore, starting from the transformer structure, an adaptive global feature extraction module based on head selection pooling transformer after dual-branch fusion features, is introduced after the transformer encoder module to mitigate token redundancy in an adaptive computational style. Finally, the multiscale aggregated spatial–spectral features and nonredundant global sequence properties are fed into the classifier to accomplish land cover classification. Toward the end, the proposed MHST exhibits three main contributions, which include the following.

- 1) The embedding of multiscale spectral–spatial feature extraction (MSFE) module can simultaneously capture spatial and spectral features from HSI and LiDAR data at

different scales, effectively considering the global structure and local detailed characteristics of various-scale land-covers.

- 2) The head selection pooling transformer based on a decision network is proposed for learning global and nonredundant spectral features. This is achieved through the sequential stacking of multiple layers of conventional transformer and an adaptive head selection pooling transformer.
- 3) In three publicly available datasets, we validated the impact of different feature fusion weights on MHST and confirmed the effectiveness of MHST as proposed in this article. In addition, we publicly provide codes, training weights, and training log.

The rest of the article is organized as follows. Related works are given in Section II. In Section III, we introduce MHST and provide a comparison of its experimental results with other related methods in Section IV. Finally, Section V concludes this article.

II. RELATED WORK

A. CNNs-Based Methods

In CNNs-based methods, dual-branch or multibranch classification architecture can effectively classify land-cover [30], Xue et al. [31] inserted hierarchical residual structure, self-calibrated convolution, self-attention module, and nonlinear feature fusion style into multiscale deep learning network with self-calibrated convolution. Xu et al. [17] focused extensively on addressing the challenge of imbalanced multimodal learning and feature interaction, proposed a dual-branch dynamic modulation network. Roy et al. [32] incorporated morphology learning and convolutional features into dual-branch networks for exploring the powerful joint features. Fang et al. [33] utilized spatial and spectral enhancement modules to enhance the spatial and spectral features effectively. To enhance the collaborative utilization of multisource land cover classification, superpixel-guided kernel principal component analysis, 2-D and 3-D Gabor filters, and a weighted majority voting-based decision fusion strategy were incorporated to effectively enhance multisource land cover classification [34]. Other relevant CNNs-based methods include attention-based CNNs method [35], [36], [37], [38], [39], a triplet deep neural network [40], deep encoder–decoder network (EndNet) [41], MDL-cross method [42], and a feature fusion and extraction framework (FusAtNet) [43]. CNNs-based methods can effectively fuse the rich spectral information of HSI and the elevation information of LiDAR, fully leveraging their complementarity.

B. Transformer-Based Methods

Although CNNs excel in capturing local features and spatial structures in imagery, global spatial–spectral association is absolutely crucial for classifying HSI and LiDAR data. The GLT-Net [25] introduced multiscale convolutional and spectral feature learning (based on transformer network) modules, then

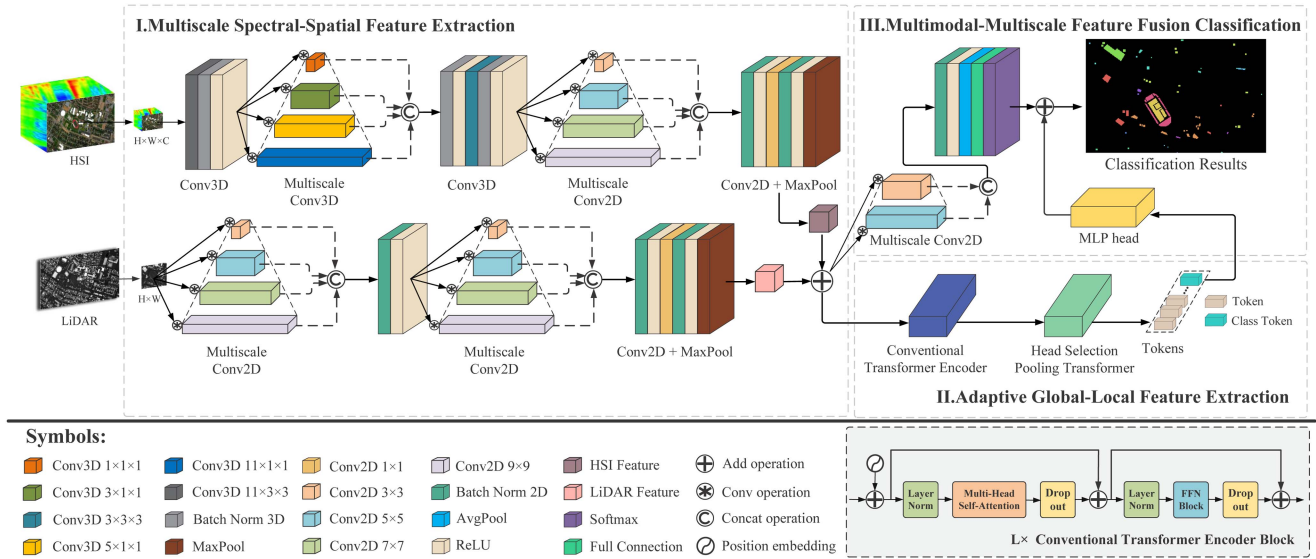


Fig. 1. Overall framework of MHST for HSI and LiDAR classification, including MSFE, AGFE, and MFFC.

the complete exploration and collaborative utilization of supplementary data in multiple modes, as well as the local and global spectral-spatial details, can be achieved. Meanwhile, a fusion encoder known as cross-token attention [44] was created to merge the spectral and spatial features of HSI and LiDAR data. Xue et al. [45] and Zhang et al. [46] designed spatial-spectral hierarchical transformer and multimodal transformer for exploring the effectiveness of transformer structures. A transformer and multiscale fusion network [47], including attention strategy and scale-based method, was performed on LiDAR and HSI classification. In addition, numbers of improved transformer networks were employed in HSI classification filed, such as [48], [49], [50], [51], [52], [53], [54]. Although the above-mentioned methods exhibit significant advantages in feature-level learning and fusion, they face limitations in effectively capturing the intricate details and distinctive characteristics of land cover in multimodal data.

III. METHODOLOGY

The proposed MHST primarily consists of three parts as follows. MSFE, adaptive global-local feature extraction (AGFE), and multimodal-multiscale feature fusion classification (MFFC), as illustrated in Fig. 1. The MSFE learns the spectral and spatial features of HSI by cascading multiscale 3-D CNNs and 2-D CNNs, and fuses them with elevation features extracted by multiscale 2-D CNNs to capture local spectral-spatial features. The AGFE aims to selectively learn global-local spectral features by cascading multiple layers of conventional transformer and adaptive head selection pooling transformer. MFFC integrates local spectral-spatial features and global-local spectral features, thereby improving the model's classification performance. The sections in the following provide detailed descriptions of MSFE, AGFE, and MFFC.

A. Multiscale Spectral-Spatial Feature Extraction

Since the conventional transformer has deficiencies in local feature expression, numbers of models applied on land-cover classification adopt CNNs to preliminarily extract local spatial features from input images. In previous work, smaller sized convolutional kernels have been widely used in CNNs due to their fewer parameters and high computational efficiency. These CNNs expand the receptive field by stacking multiple small-sized kernels to form a convolutional chain, and use down-sampling layers to gradually reduce the size of input. However, as each convolutional kernel only focuses on a local area, after stacking multiple layers, the receptive field may still be limited, leading to the filtering out of some subtle but important details in the feature map, thereby affecting the posterior transformer block's understanding of the global structure of HSI and LiDAR images based on the features extracted by CNNs. What is more, the strategy of using relatively smaller convolutional kernels and gradually increasing the receptive field may have limitations in handling objects of different sizes. To address these issues, we have designed the MSFE module, which consists of multiscale CNNs aimed at resolving the input image by parallel applying kernels of different sizes, expanding the receptive field, and capturing information at different levels to improve the model's ability to handle multiscale and complex scenes.

Moreover, there are physical interactions among spectral bands and correlations between spectral features in HSI data. Due to the distinct spectral reflection information of the same land cover in different bands, these bands provide unique and complementary information for land-cover classification. Building upon this, we propose a multiscale 3-D CNNs designed to simultaneously consider multiple correlated spectral bands during multiscale local spatial feature extraction, allowing the model to comprehensively capture the rich spectral information in HSI data, thereby improving classification accuracy. In particular, multiscale 3-D CNNs with four different levels of

convolutional kernels are inserted into MSFE to simultaneously capture spectral and spatial features from HSI data. The size of spatial dimensions of multiscale 3-D convolutional kernels are all set to 1, while the size of depth dimensions are sequentially set to a , $a \in \{1, 3, 5, 11\}$. For multiscale 2-D CNNs used in LiDAR data [55], [56], the sizes of four levels of convolutional kernels are empirically given as $b \times b$, $b \in \{3, 5, 7, 9\}$. Notably, spectral channels within each convolution layer are grouped for reducing model computation. Finally, batch normalization layer and ReLU activation function are given in MSFE.

MSFE captures features from single scale HSI and LiDAR image blocks by spectral–spatial feature encoder (SSFE) and spatial feature encoder (SFE). Concretely, SSFE is a 3-D convolutional sequence composed of single scale 3-D CNNs and multiscale 3-D CNNs, denoted as sequence operator E_{ssf} . An SFE can be characterized as a 2-D convolutional sequence, incorporating single scale 2-D CNNs alongside multiscale 2-D CNNs, denoted as sequence operator E_{sf} .

For HSI data $\mathbf{X}_H \in \mathbb{R}^{H \times W \times d_h}$, and LiDAR data $\mathbf{X}_L \in \mathbb{R}^{H \times W}$. $H \times W$ represents the original size of spatial dimensions of HSI and LiDAR data, and d_h is original size of spectral dimension of HSI data. After padding around data's edge pixels, patch extraction operations are performed on each pixel of HSI and LiDAR data separately, resulting in HSI cubes $\mathbf{X}_H^P \in \mathbb{R}^{m \times m \times d_h}$ and LiDAR cubes $\mathbf{X}_L^P \in \mathbb{R}^{m \times m}$, where $m \times m$ denotes spatial size.

The HSI cube \mathbf{X}_H^P of training set is employed as the input samples. Initially, it is passed through SSFE, generating a spectral–spatial signature cube. After flattening along the depth dimension, the cube is input into SFE to further extract spatial features from the spectral–spatial feature cube. Finally, a max-pooling operation is applied to reduce the spatial dimensions of the cube by half, resulting in the HSI spectral–spatial feature \mathbf{f}_{ss}^h . Similarly, for the LiDAR cube \mathbf{X}_L^P used for training, it goes through two identical layers of SFE and max-pooling operations, resulting in the LiDAR elevation feature \mathbf{f}_s^h . The features \mathbf{f}_{ss}^h and \mathbf{f}_s^h can be obtained via

$$\mathbf{f}_{\text{ss}}^h = \text{Maxpooling}(E_{\text{sf}}(E_{\text{ssf}}(\mathbf{X}_H^P))) \quad (1)$$

$$\mathbf{f}_s^h = \text{Maxpooling}(E_{\text{sf}}(E_{\text{sf}}(\mathbf{X}_L^P))). \quad (2)$$

Hence, multimodal local spectral–spatial feature \mathbf{f}_{cnn} based on CNNs extraction can be calculated via

$$\mathbf{f}_{\text{cnn}} = \omega \cdot \mathbf{f}_{\text{ss}}^h + (1 - \omega) \cdot \mathbf{f}_s^l \quad (3)$$

where ω is the weight coefficient, which can be manually adjusted. Herein, MSFE could capture spatial and spectral features from HSI and LiDAR data at different scales.

B. Adaptive Global–Local Feature Extraction

Multiple layers of conventional global transformers [57] are utilized to integrate multimodal features from local spectral space \mathbf{f}_{cnn} and are performed initial global spectral feature extraction of AGFE. Assuming that $\mathbf{z}_0 \in \mathbb{R}^{N \times d}$ represents the

features derived from \mathbf{f}_{cnn} after tokenization strategy, the processing of \mathbf{z}_0 through ViT encoder is as follows:

$$\mathbf{z}'_l = \text{MHSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \quad l = 1, 2, \dots, L_1 \quad (4)$$

$$\mathbf{z}_l = \text{FFN}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l, \quad l = 1, 2, \dots, L_1 \quad (5)$$

where L_1 represents the depth of conventional transformer encoder. Note \mathbf{f}'_{vit} as the spectral sequence attribute feature maps obtained after L_1 conventional transformer blocks. FFN, LN, and MHSA stand for feedforward networks, layer normalization, and multihead self-attention, respectively. The number of heads is empirically set to 4 and the attention calculation of each head is as follows:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} \quad (6)$$

where \mathbf{Q} , \mathbf{K} , and \mathbf{V} are the query, key and value matrices, respectively, and d_k represents the scaling factor. An MHSA employs the same computation process to obtain attention scores for each head, and it concatenates the attention scores from multiple heads and projects them into

$$\text{MHSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{Attn}_1, \dots, \text{Attn}_h) \mathbf{W}^O \quad (7)$$

where h represents the number of attention heads, \mathbf{W}^O denotes the parameter matrix. $\mathbf{W}^O \in \mathbb{R}^{h \times d_k \times N}$, N is the number of tokens.

While the MHSA mechanism in conventional transformer can map feature maps into different subspaces to extract global spectral sequence features, the potential for increased overlap in attention between heads becomes more pronounced as the number of layers in multilayered transformer deepens. This results in unnecessary information redundancy. Furthermore, because there are substantial variations between feature maps in different spectral bands, the way their long-range dependencies are managed differs. Therefore, we have designed an adaptive head selection decision network to learn the usage strategy of self-attention heads. This network chooses to selectively disable specific heads, reducing the model's computational cost and minimizing the processing of redundant information.

Concretely, the decision network consists of a linear layer, a sampling process, and a threshold selection layer ($\Theta(\cdot)$). The linear layer and sampling process are used to generate a policy probability matrix, and threshold selection layer sets a probability threshold (0.5) to filter out which self-attention heads to keep or discard. For the input \mathbf{z}_l at l th block, the self-attention head usage policy matrices for this block is

$$\mathbf{p}_l = \mathbf{W}_l \mathbf{z}_l, \quad \text{s.t. } \mathbf{p}_l \in \mathbb{R}^H \quad (8)$$

where $H = 16$ is the number of self-attention heads. $l \in [L_1 + 1, L_1 + L_2]$, and L_2 represents the depth of head selection pooling transformer. Subsequently, the binary decision k ($k = 1$ means the corresponding head is retained and $k = 0$, discarded) at i th head of l th block is derived in the following way:

$$P_{l,i,k} = \frac{\exp(p_{l,i,k} + G_{l,i,k}/\tau)}{\sum_{j=0}^1 \exp(p_{l,i,j} + G_{l,i,j}/\tau)}, \quad k \in \{0, 1\} \quad (9)$$

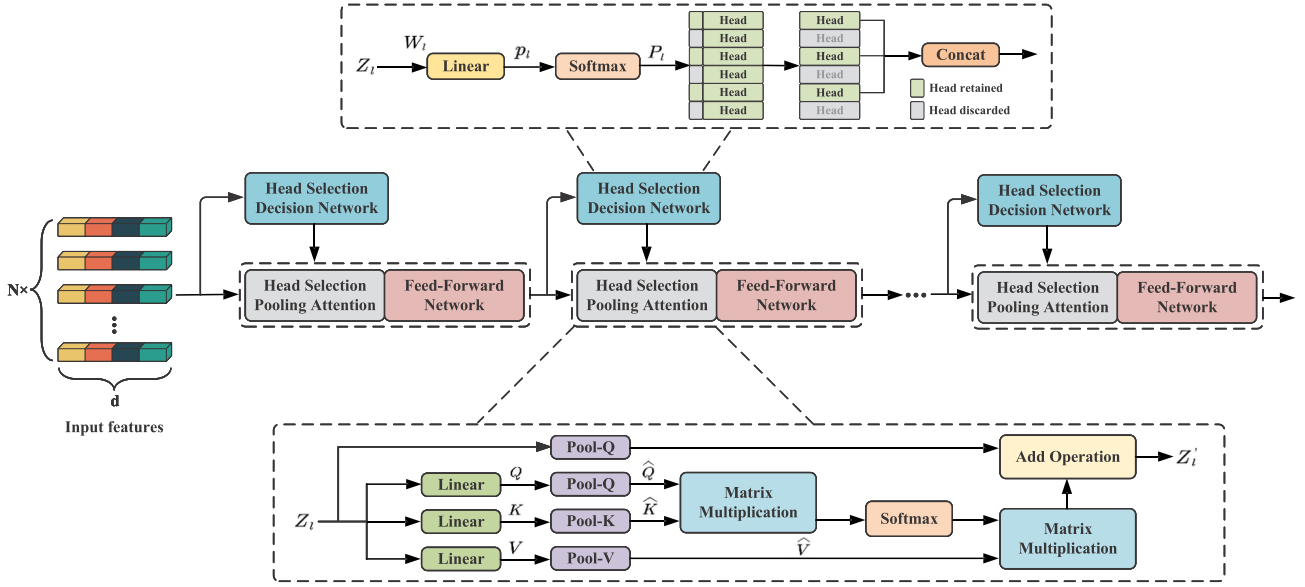


Fig. 2. Illustration of head selection pooling transformer block. We insert a head selection decision network before each vision transformer block. For the input initial features extracted by multiple layers of conventional transformer blocks, the decision network generates a usage strategy for self-attention heads. These instance-specific usage strategies aim to reduce processing of redundant information and lower computational costs. Simultaneously, we utilize the pooling operator $\mathcal{P}(\cdot; \phi)$ in the multihead self-attention mechanism to capture relationships between patches, aiming to consider both global spectral and local spatial information. See the texts for further details.

where $G_{l,i} = -\log(\text{Exp}_{l,i})$ in which $\text{Exp}_{l,i}$ is sampled from exponential distribution, and τ is used to control the distribution of output probability [58]. Afterward, the self-attention head selection strategy matrix in l th block \mathbf{P}_l is computed as

$$\mathbf{P}_l = [\Theta(\mathbf{P}_{l,1,k}); \dots; \Theta(\mathbf{P}_{l,H,k})], \quad k \in \{0, 1\} \quad (10)$$

while $\mathbf{P}_{l,i} = 1$, the i th head at l th block is retained; when $\mathbf{P}_{l,i} = 0$, the i th head at l th block is discarded.

Notably, before attending the input, the \mathbf{Q} , \mathbf{K} , and \mathbf{V} are achieved via the pooling operator $\mathcal{P}(\cdot; \phi)$ to capture spatial relationships between patches in MHSA, aiming to simultaneously consider global spectral and local spatial information

$$\hat{\mathbf{Q}} = \mathcal{P}(\mathbf{Q}; \phi_Q), \quad \hat{\mathbf{K}} = \mathcal{P}(\mathbf{K}; \phi_K), \quad \hat{\mathbf{V}} = \mathcal{P}(\mathbf{V}; \phi_V) \quad (11)$$

where $\phi = (\mathbf{k}, \mathbf{s}, \mathbf{p})$ in which \mathbf{k} , \mathbf{s} , and \mathbf{p} represent the pooling kernel, corresponding stride, and padding, respectively. Certainly, the operation imposes limitations represented by constraints $s_K \equiv s_V$ on the pooling operators applied to \mathbf{Q} , \mathbf{K} , and \mathbf{V} while utilizing the same padding strategy to preserve shape.

As shown in Fig. 2, head selection decision network and pooling operation are inserted into head selection pooling transformer block. In our model, two head selection strategies are adopted, namely partial discard and complete discard. For partial discard of attention heads, the attention matrix corresponding to that head is replaced by an identity matrix $\mathbf{1}$. Then the computation of attention in i th head of l th block [58] is as follows:

$$\text{Attn}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}})_{l,i} = \begin{cases} \text{Softmax}\left(\frac{\hat{\mathbf{Q}}\hat{\mathbf{K}}^T}{\sqrt{d_k}}\right) \cdot \hat{\mathbf{V}}, & \mathbf{P}_{l,i} = 1 \\ \mathbf{1} \cdot \hat{\mathbf{V}}, & \mathbf{P}_{l,i} = 0. \end{cases} \quad (12)$$

Regarding complete discard, the entire head is correspondingly removed from the MHSA mechanism and does not participate in the computation of self-attention for that layer

$$\text{MHSA}(\cdot)_{l,i} = \text{Concat}([\text{Attn}_{l,i:1 \rightarrow H} \text{ if } \mathbf{P}_{l,i} = 1])\mathbf{W}_l^{O'}. \quad (13)$$

In general, the forward propagation process takes spectral sequence feature \mathbf{f}'_{vit} obtained from conventional global transformer as input, and obtains the global-local spectral sequence feature \mathbf{f}_{vit} after L_2 head selection pooling transformer block. The cls token $\mathbf{f}_{\text{vit}}^{\text{cls}} \in \mathbb{R}^{1 \times d}$ is extracted from \mathbf{f}_{vit} for subsequent classification tasks

$$\mathbf{f}_{\text{vit}} = \mathbf{z}_{L_2}, \quad \mathbf{f}_{\text{vit}}^{\text{cls}} = \text{LN}(\mathbf{f}_{\text{vit}}^0). \quad (14)$$

C. Multimodal-Multiscale Feature Fusion Classification

Feature fusion decision classification is employed to fully capture the local spectral-spatial features and the global-local spectral features in MFFC. Specifically, \mathbf{f}_{cnn} is classified by a CNNs-based network for outputting classification probabilities, consisting of a multiscale 2-D CNNs, batch normalization, and ReLU activation layers. Afterward, adaptive global average pooling, fully connected layer, and softmax function are inserted. For $\mathbf{f}_{\text{vit}}^{\text{cls}}$, it is fed into multilayer perceptron and softmax layer for classification.

Finally, we make a decision classification based on classification probability vectors corresponding to CNNs and ViT, denoted as \mathbf{P}_{cnn} and \mathbf{P}_{vit} , to obtain the final probability vector $\mathbf{P}_f \in \mathbb{R}^{1 \times \text{cls}}$, where the label corresponding to the maximum probability is assigned as the class for that pixel. \mathbf{P}_f can be represented as $\mathbf{P}_f = \lambda \cdot \mathbf{P}_{\text{vit}} + (1 - \lambda) \cdot \mathbf{P}_{\text{cnn}}$ where λ is the weight coefficient for feature fusion decision classification.

IV. EXPERIMENT AND ANALYSIS

A. Data Description

In the experiments, three commonly HSI and LiDAR datasets are utilized to evaluate the effectiveness of MHST.

1) *Houston2013 Dataset*: The Houston2013 dataset [25] includes an HSI and a LiDAR-based DSM, collected by the National Center for Airborne Laser Mapping in June 2012 using the ITRES CASI-1500 imaging sensor over the campus of the University of Houston. The dataset was provided by the IEEE GRSS Data Fusion Competition. The HSI comprises 144 spectral bands covering a wavelength range from 0.38 to 1.05 μm while LiDAR data are provided for a single band. Both HSI and LiDAR data share dimensions of 349×1905 pixels with a spatial resolution of 2.5 m. The dataset contains 15 categories, with a total of 15 029 real samples available.

2) *Trento Dataset*: The Trento dataset comprises HSI and LiDAR data obtained from southern Trento, Italy. The HSI was obtained by the airborne hyperspectral imaging systems Eagle sensor, consisting of 63 spectral bands with a wavelength range from 0.42 to 0.99 μm [44]. LiDAR data were gathered using the Optech airborne laser topographic mapping (ALTM) 3100EA sensor with one raster. The scene consists of 166×600 pixels, with a spatial resolution of 1 m. This dataset contains six land cover types with a total of 30 214 real samples.

3) *MUUFLL Dataset*: The MUUFLL dataset was acquired in November 2010 over the area of the campus of University of Southern Mississippi Gulf Park, Long Beach Mississippi, USA. The HSI data were gathered via ITRES Research Limited (ITRES) compact airborne spectral imager (CASI-1500) sensor, initially comprising 72 bands. Due to excessive noise, the first and last eight spectral bands were removed, resulting in a total of 64 available spectral channels ranging from 0.38 to 1.05 μm . LiDAR data were captured by an ALTM sensor, containing two rasters with a wavelength of 1.06 μm . This dataset consists of 53 687 ground-truth pixels, encompassing 11 different land-cover classes.

B. Experimental Setting

The proposed MHST is implemented in PyTorch framework. The experiments are performed on Ubuntu 22.04 platform equipped with an I9-13 900 K CPU, a NVIDIA RTX 4090Ti GPU, and RAM: 32 GB. We use an AdamW optimizer with a learning rate decay parameter of 0.9 to optimize the network. In the training phase, the batch size, and the number of training epochs are set to 64, and 3000, respectively. Considering three datasets have different data scales and spatial resolutions, initial learning rates are set to $8e-4$ (Houston2013 dataset), $5e-4$ (Trento dataset), and $4e-4$ (MUUFLL dataset). In terms of model parameter configuration, the depths of conventional transformer encoder block and head selection pooling transformer block are set to 5 and 8, respectively. And initial values of feature fusion weight coefficient ω and feature fusion decision classification weight coefficient λ are set to 0.6 and 0.7, respectively. For head selection, a complete discard strategy is employed, unless otherwise specified. Standard cross-entropy is utilized as the loss function.

Moreover, three evaluation indicators are adopted to quantitatively reflect the classification performance of MHST: overall accuracy (OA), average accuracy (AA), and Kappa coefficient. Tables I, II, and III provide detailed information of training and testing samples.

C. Performance Comparison

To validate the effectiveness of our proposed framework, we selected several representative HSI and LiDAR joint classification models for comparison, including EndNet [41], FusAtNet [43], MDL-cross [42], S2E [33], HCT-Net [44], and GLT-Net [25]. Among these models, EndNet, FusAtNet, MDL-cross, and S2E are all based on deep CNNs architectures, while HCT-Net and GLT-Net utilize CNNs-transformer architectures. The parameters for these models were set according to their respective reference papers and optimized on the same server. Furthermore, same training and testing samples were used for fair comparison.

Several comparative methods are evaluated through visual comparisons (as shown in Figs. 3–5) and quantitative metrics, e.g., per-class accuracy, OA, AA, and Kappa coefficient. Tables I–III clearly present the objective classification results of our proposed method and each of the comparative methods on Houston2013, Trento, and MUUFLL datasets, with the best results in each row highlighted in bold.

1) *Quantitative Analysis*: Tables I–III display the quantitative classification results of different methods on three popular HSI and LiDAR datasets, respectively. Through the analysis, our MHST achieves the highest classification scores on three datasets, surpassing the second-best model by approximately 0.5%, 0.3%, and 3% on Houston2013, Trento, and MUUFLL datasets, respectively. By combining the analysis of features related to land-cover categories and their corresponding classification accuracy, we draw the following conclusions.

- 1) For the multimodal feature extraction fusion CNNs framework, feature extraction capability of multiscale convolution kernels is superior to single-scale kernels. For instance, even though MHST and HCT-Net both possess the ability to capture spectral and spatial features from HSI simultaneously, MHST outperforms HCT-Net and surpasses it by more than 2% on three objective metrics for two material-similar classes on Houston2013 dataset, “Park lot 1” and “Park lot 2.” This is due to the application of multiscale convolutions, allowing for the capture of more local spatial neighborhood features.
- 2) Concerning the extraction of features from HSI data with multiple spectral channels, 3-D CNNs are better suited for capturing spectral features as compared to 2-D CNNs. For example, with MHST and GLT-Net having multiscale HSI feature extraction modules, land-cover category “Railway” on Houston2013 dataset is challenging to classify accurately with single-modal data, our model’s classification accuracy exceeds GLT-Net by over 3%, thanks to the thorough exploration of spectral features by 3-D CNNs.
- 3) The global spectral features enhances the model’s classification performance. Specifically, although S2E method

TABLE I
COMPARISON OF CLASSIFICATION PERFORMANCES OBTAINED BY DIFFERENT METHODS FOR HOUSTON2013 DATA

No.	Class(Train/Test)	CNNs-based networks				CNNs-Transformer-based networks		
		EndNet	FusAtNet	MDL-Cross	S2E	HCT-Net	GLT-Net	MHST
1	Healthy grass (20/1231)	71.00	81.48	88.95	97.40	89.03	94.09	98.05
2	Stressed grass (20/1234)	96.84	93.84	98.06	98.78	96.52	96.74	98.22
3	Synthetic grass (20/677)	99.85	99.85	99.41	98.38	99.85	99.57	99.26
4	Trees (20/1224)	92.97	94.61	99.59	100	95.51	98.15	99.59
5	Soil (20/1222)	100	91.16	98.69	99.92	99.92	99.80	99.35
6	Water (20/305)	93.44	95.74	97.38	100	100	97.08	99.67
7	Residential (20/1248)	89.74	92.15	92.39	96.31	99.04	94.94	95.59
8	Commercial (20/1224)	78.27	77.12	93.95	89.71	89.38	91.59	90.77
9	Road (20/1232)	74.51	88.31	90.34	89.04	84.01	94.20	89.20
10	Highway (20/1207)	85.34	65.45	85.42	91.21	92.87	91.92	96.77
11	Railway (20/1215)	89.63	91.03	92.02	94.73	94.16	93.18	94.32
12	Park lot 1 (20/1213)	92.25	92.83	89.45	93.08	87.63	90.68	93.82
13	Park lot 2 (20/449)	61.91	90.87	95.99	95.32	100	98.40	97.33
14	Tennis court (20/408)	94.61	98.53	99.02	100	100	99.90	100
15	Running track (20/640)	99.38	100	99.53	100	100	99.98	100
	OA(%)	87.75	88.63	93.84	95.65	94.02	95.32	96.19
	AA(%)	87.98	90.20	94.68	96.26	95.19	96.01	96.80
	Kappa(%)	86.74	87.72	93.34	95.29	93.53	94.95	95.88

The bold entities indicate the highest classification accuracy obtained for a single land-cover category among all comparison algorithms and the proposed model.

TABLE II
COMPARISON OF CLASSIFICATION PERFORMANCES OBTAINED BY DIFFERENT METHODS FOR TRENTO DATA

No.	Class(Train/Test)	CNNs-based networks				CNNs-Transformer-based networks		
		EndNet	FusAtNet	MDL-Cross	S2E	HCT-Net	GLT-Net	MHST
1	Apple trees (20/4014)	85.97	98.85	99.10	99.58	98.28	99.34	99.60
2	Buildings (20/2883)	94.00	98.89	97.40	96.43	94.48	97.68	98.06
3	Ground (20/459)	94.34	97.60	97.39	97.82	100	97.89	99.78
4	Woods (20/9103)	98.76	99.93	100	100	100	99.93	100
5	Vineyard (20/10481)	62.71	99.55	99.90	99.99	99.55	99.80	99.99
6	Roads (20/3154)	90.71	97.40	95.97	98.10	92.61	97.17	97.08
	OA(%)	83.13	99.26	99.13	99.37	98.31	99.27	99.45
	AA(%)	87.75	98.71	98.29	98.65	97.49	98.63	99.09
	Kappa(%)	78.19	99.01	98.84	99.15	97.75	99.03	99.26

The bold entities indicate the highest classification accuracy obtained for a single land-cover category among all comparison algorithms and the proposed model.

exhibits high classification accuracy, surpassing all CNNs-based networks, our approach, which is based on a combination of conventional transformer and head selection pooling transformer in the global feature extraction module, makes full use of global spectral dependencies. As a result, our method still achieves higher classification accuracy and has the best classification performance in five categories of 15 land-cover categories as compared to S2E method.

- 4) The classification strategy after decision-level multimodal-multiscale feature fusion can further improve classification accuracy. For instance, the performance of MHST with MFFC module surpasses any single

feature-level fusion method on MUUFL dataset, such as MDL-cross. Similarly, GLT-Net, employing a similar decision-level feature fusion strategy, achieves the second-best results.

- 5) The selection of self-attention heads in the transformer contributes to reducing attention on redundant features. Specifically, for “Yellow curb” on MUUFL dataset with a smaller number of samples, despite the complexity and clutter in the scenes near the pixels of this category, benefiting from the embedding of the designed head selection decision network in transformer architecture, the proposed MHST achieves a classification accuracy of up to 92.68%, which is about 10% and 13% higher than HCT-Net and

TABLE III
COMPARISON OF CLASSIFICATION PERFORMANCES OBTAINED BY DIFFERENT METHODS FOR MUUFL DATA

No.	Class(Train/Test)	CNNs-based networks				CNNs-Transformer-based networks		
		EndNet	FusAtNet	MDL-Cross	S2E	HCT-Net	GLT-Net	MHST
1	Trees (60/23186)	82.04	90.75	87.14	89.64	87.32	92.20	91.81
2	Mostly grass (60/4210)	79.48	74.20	75.58	81.00	75.99	80.59	85.70
3	Mixed ground surface (60/6822)	64.00	64.45	74.04	67.40	69.29	70.57	72.78
4	Dirt and sand (60/1766)	89.47	87.49	87.77	94.73	95.13	80.41	89.81
5	Road (60/6627)	86.10	87.22	84.23	81.80	81.67	88.77	88.29
6	Water (60/406)	98.77	100	100	100	100	100	100
7	Buildings shadow (60/2173)	89.78	92.54	94.52	94.43	90.29	96.36	93.05
8	Buildings (60/6180)	87.61	93.06	94.16	92.51	94.66	93.41	95.52
9	Sidewalk (60/1325)	73.28	71.77	64.60	76.08	64.60	84.68	82.19
10	Yellow curb (60/123)	95.12	82.11	86.18	92.68	82.93	79.67	92.68
11	Cloth panels (60/209)	98.09	97.61	98.56	99.52	99.52	99.52	99.04
	OA(%)	81.24	85.45	84.89	85.60	84.20	87.86	88.71
	AA(%)	85.79	85.56	86.07	88.16	85.58	87.84	90.08
	Kappa(%)	76.08	81.15	80.52	81.42	79.67	84.20	85.32

The bold entities indicate the highest classification accuracy obtained for a single land-cover category among all comparison algorithms and the proposed model.

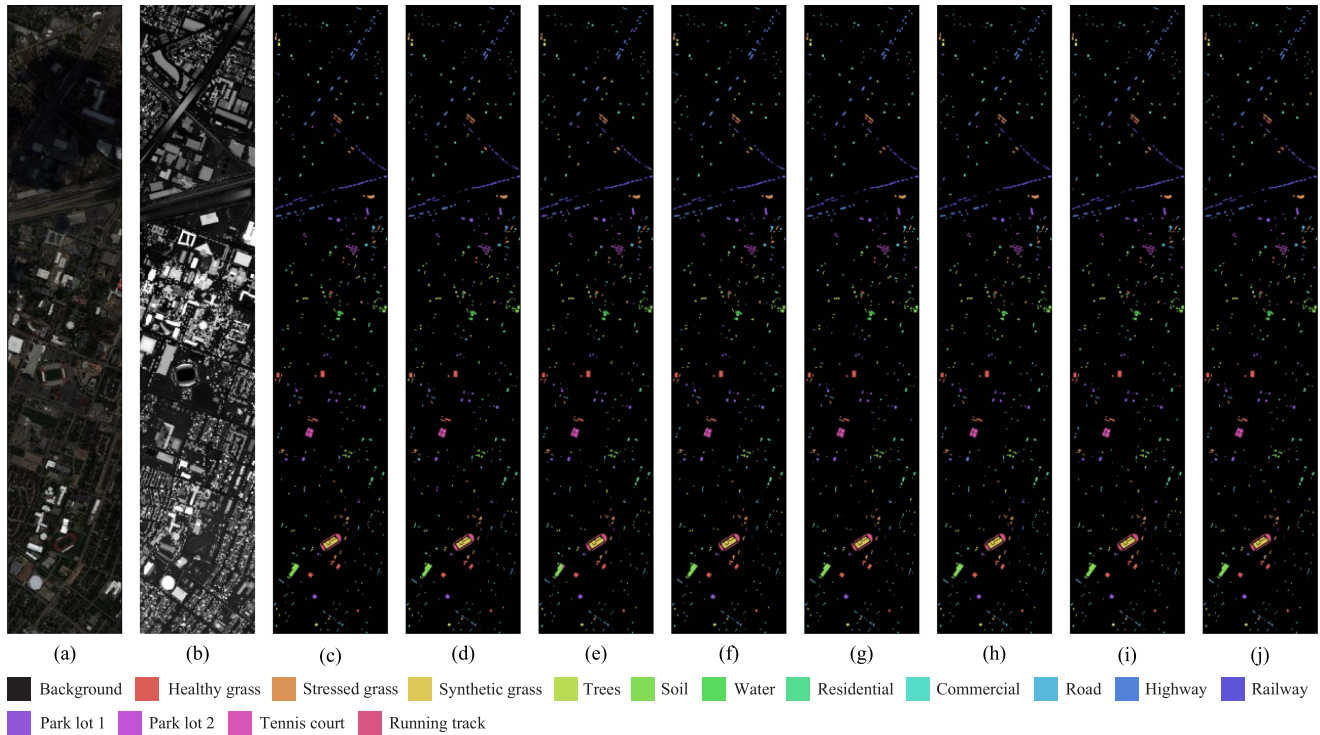


Fig. 3. Classification maps by different methods on Houston2013 dataset. (a) Pseudocolor image for HSI. (b) LiDAR-based DSM. (c) Ground-truth map. (d) EndNet (87.75%). (e) FusAtNet (88.63%). (f) MDL-cross (93.84%). (g) S2E (95.65%). (h) HCT-Net (94.02%). (i) GLT-Net (95.32%). (j) MHST (96.19%).

GLT-Net, respectively. Furthermore, our MHST achieves the highest classification scores in three out of 11 categories.

Based on these advantages, we seamlessly embed the proposed modules into an end-to-end framework, aiming to simultaneously consider multiscale, local spectral-spatial features. Simultaneously, we utilize MHSA mechanism with a combined head selection decision network to consider global spectral

nonredundant information. The decision-level feature fusion classification method is also adopted to improve the classification performance.

2) *Visual Comparison and Analysis*: Figs. 3–5, respectively, illustrate the classification maps of several methods on three datasets, where Trento and MUUFL utilize local zoom operations to more clearly display the performance differences between different methods. It can be seen that as compared to other

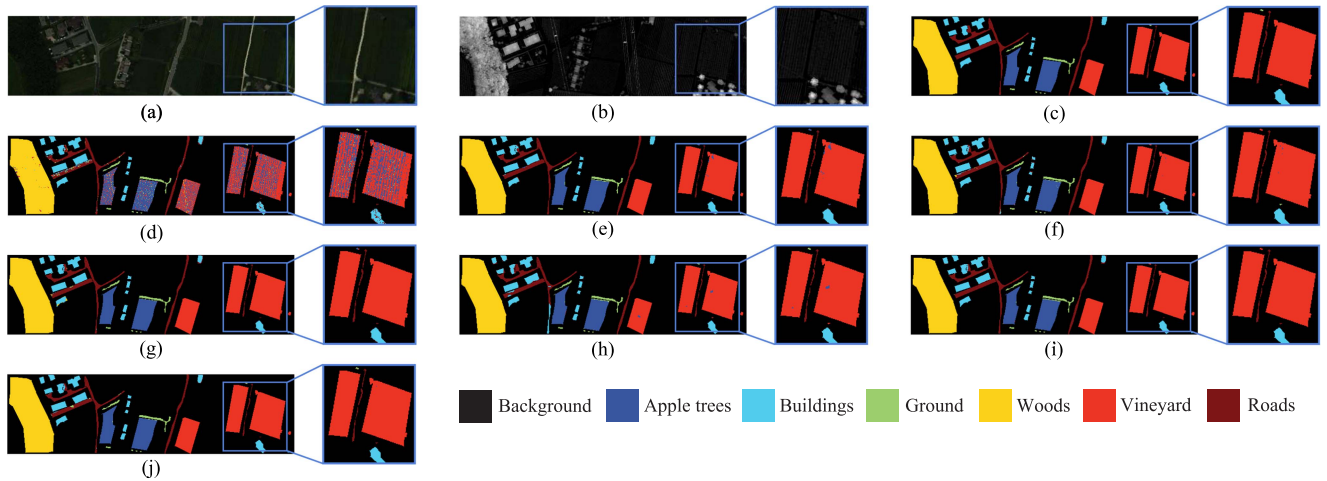


Fig. 4. Classification maps by different methods on Trento dataset. (a) Pseudocolor image for HSI. (b) LiDAR-based DSM. (c) Ground-truth map. (d) EndNet (83.13%). (e) FusAtNet (99.26%). (f) MDL-cross (99.13%). (g) S2E (99.37%). (h) HCT-Net (98.31%). (i) GLT-Net (99.27%). (j) MHST (99.45%).

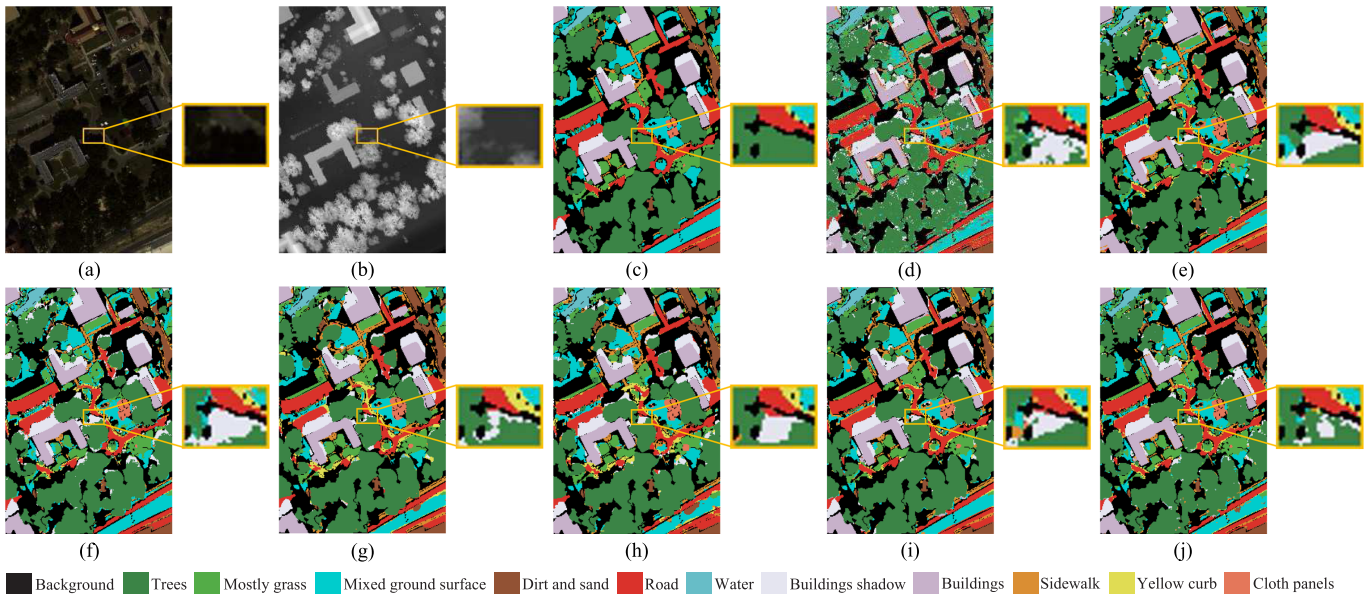


Fig. 5. Classification maps by different methods on MUUFL dataset. (a) Pseudocolor image for HSI. (b) LiDAR-based DSM. (c) Ground-truth map. (d) EndNet (81.24%). (e) FusAtNet (85.45%). (f) MDL-cross (84.89%). (g) S2E (85.60%). (h) HCT-Net (84.20%). (i) GLT-Net (87.86%). (j) MHST (88.71%).

methods, the visualized classification results of MHST are closer to the ground-truth map, resulting in better classification performance. Furthermore, MHST exhibits superior classification performance compared to other methods, resulting in smoother classification results. On the other hand, several classification methods, such as EndNet, FusAtNet, and MDL-cross tend to have more isolated data points. Specifically, almost all other methods incorrectly classify some “vineyard” as “apple trees” on Trento dataset, as shown in Fig. 4. In the locally enlarged image, some methods classify this category as the “vineyard” which is geographically adjacent, for the “ground” with fewer samples. In contrast, the classification results of MHST are almost completely correct, consistent with the results shown in Table II. For MUUFL dataset shown in Fig. 5, the proposed method

produces clearer classification boundaries. Such as, MHST’s results are closest to the ground-truth map for the boundary between “road” and “trees,” while the boundaries produced by other methods are not only more blurred and difficult to distinguish, but also produce more classification errors, e.g., classifying “trees” as “buildings shadow.” In conclusion, MHST has demonstrated its effectiveness in both quantitative and visual analysis, showcasing strong classification performance.

3) *Computational Complexity Analysis*: Table IV shows the computational complexity of different methods, including the trainable parameters in the backpropagation phase and the testing time on the MUUFL dataset. It can be seen that the number of trainable parameters in CNNs-transformer-based networks is higher than in most CNNs-based networks. The EndNet method

TABLE IV
COMPARISON OF COMPUTATIONAL COMPLEXITY AND TESTING TIME(S) OF DIFFERENT METHODS ON THE MUUFL DATA

Methods	EndNet	FusAtNet	MDL-Cross	S2E	HCT-Net	GLT-Net	MHST
Trainable Params	381.45K	138.48M	445.58K	745.86K	1.76M	1.83M	3.45M
Testing time(s)	4.50	31.57	7.98	8.45	7.80	22.79	25.28

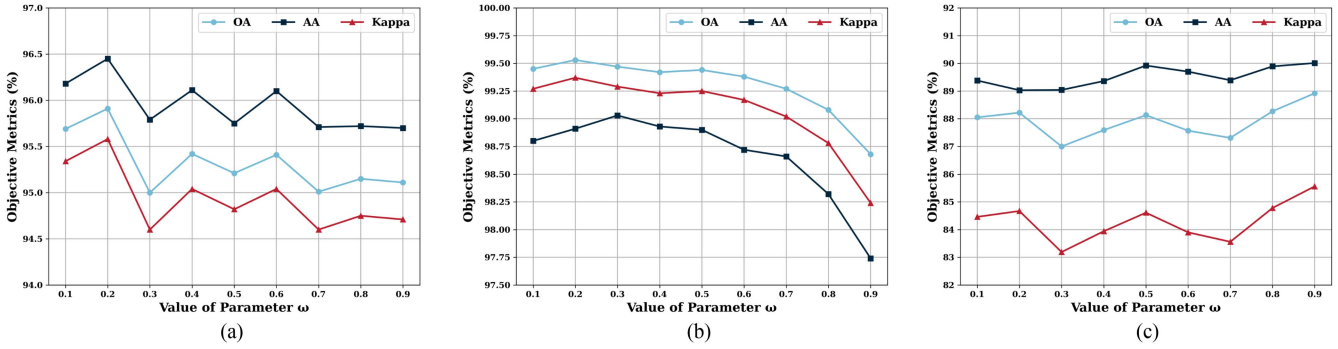


Fig. 6. Effects of the weight parameter ω on classification performance. (a) Houston2013. (b) Trento. (c) MUUFL.

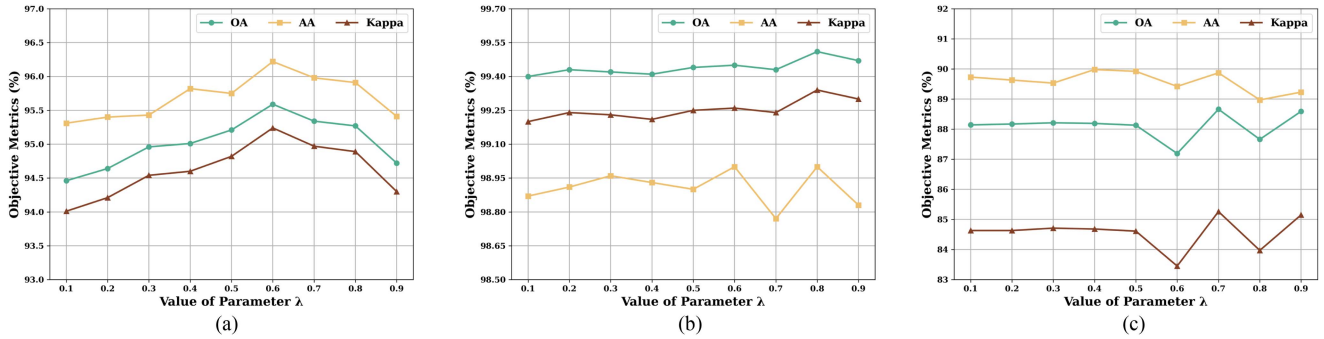


Fig. 7. Effects of the weight parameter λ on classification performance. (a) Houston2013. (b) Trento. (c) MUUFL.

has the fewest trainable parameters, while the FusAtNet uses the most parameters. The proposed model's parameter counts and testing time are slightly higher than those of the GLT-Net with a similar network structure because we embedded multiple layers of head selection pooling transformer blocks after the conventional transformer encoder, increasing the computational cost while improving the accuracy of land-cover classification. Within an acceptable range of testing time and computational cost, MHST exhibits the best classification performance.

D. Parameters Analysis

1) *Weight Coefficient*: Different feature fusion weights ω and decision classification weights λ have influences on classification performance. Then, we set the default values of ω and λ to 0.5, while keeping other hyperparameters fixed, and set the values of ω and λ from 0.1 to 0.9 in fixed increments of 0.1. Figs. 6 and 7 show the different values of ω and λ and their corresponding OA, AA, and Kappa on three datasets. It

can be observed that while ω is less than 0.5, which means fewer HSI feature feeds, MHST achieves optimal classification performance on two datasets (Houston2013 and Trento). Nevertheless, while λ exceeds 0.5, indicating that the classification decision relies more on the features extracted by the head selection pooling transformer, the proposed model delivers the best classification results across all three datasets. In addition, while λ is too large or too small, it will lead to varying degrees of decline in classification performance. These observations collectively underscore the importance of integrating fused features from HSI and LiDAR for further spectral feature extraction, while also affirming the efficacy of designed head selection pooling transformer.

2) *Training Samples (TS) Number*: To better validate the robustness and generalizability of MHST, we systematically vary the quantity of TS and analyze the corresponding trends in overall classification accuracy. Specifically, we randomly select different samples of each land-cover category for training, and the remaining samples are used for testing. The number

TABLE V
IMPACT OF THE NUMBER OF TS PER CLASS ON MHST

Houston2013				Trento				MUUFL			
TS	OA (%)	AA (%)	Kappa (%)	TS	OA (%)	AA (%)	Kappa (%)	TS	OA (%)	AA (%)	Kappa (%)
20	96.19	96.80	95.88	20	99.45	99.09	99.26	60	88.71	90.08	85.32
40	96.15	96.62	95.84	40	99.68	99.49	99.57	70	89.40	89.46	86.13
60	97.57	98.01	97.37	60	99.70	99.51	99.60	80	88.61	90.61	85.22
80	98.22	98.34	98.07	80	99.72	99.50	99.63	90	88.35	91.46	84.87
100	99.14	99.21	99.07	100	99.74	99.57	99.65	100	88.68	90.49	85.23
120	99.09	99.28	99.02	120	99.78	99.66	99.71	110	88.74	91.39	85.33
140	99.53	99.59	99.49	140	99.75	99.58	99.66	120	89.43	92.21	86.22
160	99.26	99.42	99.19	160	99.85	99.75	99.80	130	90.10	91.96	87.05
180	99.35	99.50	99.30	180	99.82	99.71	99.76	140	87.83	91.34	84.17
200	99.62	99.71	99.58	200	99.81	99.70	99.75	150	90.31	92.44	87.27

The bold values represents the maximum value of each column.

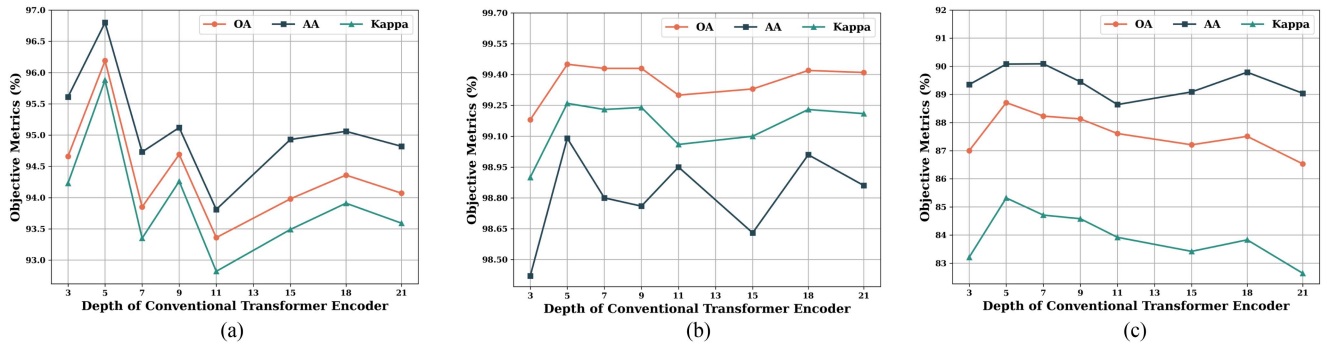


Fig. 8. Effects of the depth of conventional transformer block on classification performance. (a) Houston2013. (b) Trento. (c) MUUFL.

of selected TS are increased from 20 to 200 in steps of 20 on Houston2013 and Trento datasets. For MUUFL dataset, the number of TS is increased from 60 to 150 in steps of 10. All the training parameters are set the same as the original settings, and the evaluation metrics remained as OA, AA, and Kappa.

Table V shows the classification performance of MHST under different TS sizes. On all three datasets, the three evaluation metrics exhibit a trend of fluctuating improvement with increasing TS. Taking the Houston2013 dataset as an example, the performance of proposed network notably improves as the TS increase from 40 to 100, benefiting from additional samples providing more feature and interfeature relationship information. Nevertheless, the performance gain from 100 to 200 samples is only about 20% of the gain observed from 40 to 100 samples. This might be because the model already captured most crucial features with 100 TS for each category, so increasing the TS size may not bring the same level of information gains. The marginal reduction on classification performance could be attributed to potential model overfitting at a particular training data, leading to fluctuations in the classification performance, rather than a consistent increase. In general, with more TS fed, the proposed MHST is able to extract strong features to improve the accuracy of classification and generalize across varying data volumes, resulting in consistent and strong classification performance.

3) *Depth of Conventional Transformer Encoder*: The depth of the conventional transformer encoder L_1 will affect the

model's feature representation ability. As shown in Fig. 8, by varying the size of L_1 on three datasets to explore its impact on classification performance, it can be observed that while the encoder depth is 5, the proposed model can achieve the best classification performance. In addition, increasing the depth does not necessarily lead to better classification performance. With the increase in depth, the model's classification performance on the three datasets shows a decreasing trend. On the other hand, a conventional transformer encoder that is too shallow may not be sufficient to capture complex data patterns and features, which can also lead to a decrease in model performance. This indicates the importance of determining the optimal encoder depth through experiments to balance representation ability and prevent overfitting.

E. Ablation Studies

Due to some trainable parameters located within the head selection pooling transformer module, we conducted ablation experiments to specifically assess the effectiveness of the head selection decision network and pooling operation. The experimental results, based on three objective metrics (OA, AA, and Kappa) from three datasets, are depicted in Fig. 9, the discrepancy in classification evaluation metrics among different datasets signifies the varying advantages of two modules in feature extraction. For Houston2013 dataset with higher scene

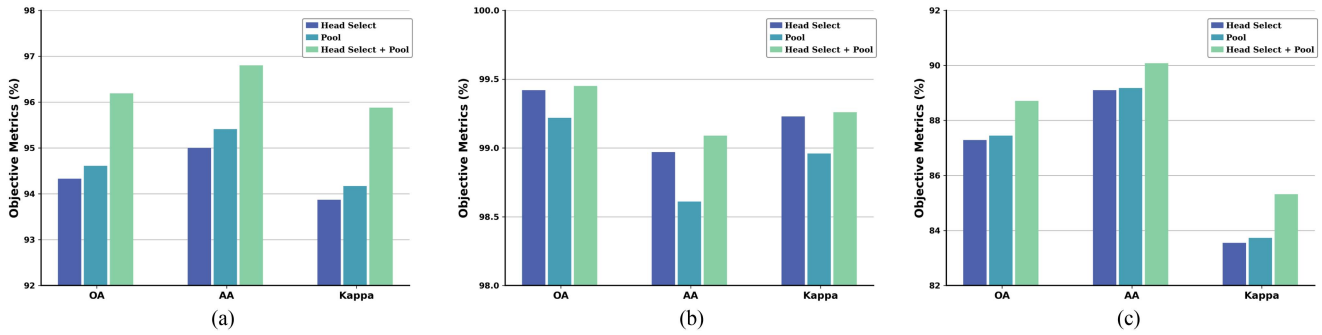


Fig. 9. Impact of head selection decision network and pooling operation on classification performance. (a) Houston2013. (b) Trento. (c) MUUFL.

complexity, benefiting from the integration of pooling operation in transformer, the second variant can focus on global features while also considering the extraction of local features, thereby paying more attention to information near the target classified pixels. Consequently, compared to first variant, which omits the pooling operation, the second variant achieves slightly higher classification accuracy. Conversely, as for the Trento dataset, it has lower scene complexity, which leads to limited local feature information. Since the head selection decision network enhances the global feature extraction capability, the first variant achieves a higher score than the second variant. Simultaneously, the contribution of locally extracted features by the pooling operation module to classification accuracy is reduced. Thus, the scores of the first variant is only slightly lower than both two modules.

For MUUFL dataset, only using the head selection decision network or pooling operation leads to lower classification accuracy, by around 1.3%, as compared to the combined use of both modules. Overall, the experimental findings and objective analysis suggest that both head selection decision network and pooling operation have a positive impact on land-cover classification.

V. CONCLUSION

This article focuses on the detailed and nonredundant features in spatial and spectral dimension for efficiently HSI and LiDAR classification. The MSFE module effectively accounts for the overall patterns and intricate local attributes of various-scale land-covers. Adaptive global feature extraction module could adaptively select the heads in transformer to avoid feature redundancy caused by the participation of all heads. Furthermore, we validated the effectiveness of MHST under different feature fusion ratios and verified the performances of proposed MHST from multiple dimensions, such as different TS and ablation experiments. Moreover, there are some details in MHST that deserve improvement. For example, structural features are crucial for HSI and LiDAR data classification, so one of the key focuses in future work is how to fully utilize the selected structural features of HSI. Specifically, we will further explore how to improve the head selection decision network to generate a more effective head selection strategy, reasonably allocate weights to the retained heads, and make the model further select

the retained nonredundant features after filtering out redundant features. Simultaneously, since word tokens carry more specific information, the model could pay more detailed attention to local features. Therefore, one of the main research contents in the future is to explore how to more effectively integrate word tokens in order to optimize the model's understanding of HSI and LiDAR data.

REFERENCES

- [1] W. Qi, C. Huang, Y. Wang, X. Zhang, W. Sun, and L. Zhang, "Global-local three-dimensional convolutional transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5510820.
- [2] Z. Chen, D. Hong, and H. Gao, "Grid network: Feature extraction in anisotropic perspective for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5507105.
- [3] C. Zhao et al., "Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning," *IEEE Trans. Image Process.*, vol. 32, pp. 3606–3621, 2023.
- [4] H. Gao, H. Feng, Y. Zhang, S. Xu, and B. Zhang, "AMSSE-Net: Adaptive multiscale spatial-spectral enhancement network for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5531317.
- [5] W. Dong, T. Yang, J. Qu, T. Zhang, S. Xiao, and Y. Li, "Joint contextual representation model-informed interpretable network with dictionary aligning for hyperspectral and LiDAR classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6804–6818, Nov. 2023.
- [6] E. Ouyang, B. Li, W. Hu, G. Zhang, L. Zhao, and J. Wu, "When multigranularity meets spatial-spectral attention: A hybrid transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401118.
- [7] W. Li, Q. Du, and B. Zhang, "Combined sparse and collaborative representation for hyperspectral target detection," *Pattern Recognit.*, vol. 48, no. 12, pp. 3904–3916, 2015.
- [8] Z. Chen et al., "Global to local: A hierarchical detection algorithm for hyperspectral image target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5544915.
- [9] Z. Chen et al., "Temporal difference-guided network for hyperspectral image change detection," *Int. J. Remote Sens.*, vol. 44, no. 19, pp. 6033–6059, 2023.
- [10] Q. Meng et al., "Joint inversion of gravity and magnetic data with tetrahedral unstructured grid and its application to mineral exploration," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5915514.
- [11] Z. Chen et al., "Local aggregation and global attention network for hyperspectral image classification with spectral-induced aligned superpixel segmentation," *Expert Syst. Appl.*, vol. 61, 2023, Art. no. 5915514.
- [12] J. Xia, W. Liao, and P. Du, "Hyperspectral and LiDAR classification with semisupervised graph fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 666–670, Apr. 2020.
- [13] F. Jahan, J. Zhou, M. Awrangjeb, and Y. Gao, "Inverse coefficient of variation feature and multilevel fusion technique for hyperspectral and LiDAR data classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 367–381, 2020.

- [14] 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.
- [15] S. Jia, X. Zhou, S. Jiang, and R. He, "Collaborative contrastive learning for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5507714.
- [16] L. Song, Z. Feng, S. Yang, X. Zhang, and L. Jiao, "Discrepant Bidirectional interaction fusion network for hyperspectral and LiDAR data classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5510605.
- [17] Z. Xu, W. Jiang, and J. Geng, "Dual-branch dynamic modulation network for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514813.
- [18] M. Zhang, W. Li, Q. Du, L. Gao, and B. Zhang, "Feature extraction for classification of hyperspectral and LiDAR data using patch-to-patch CNN," *IEEE Trans. Cybern.*, vol. 50, no. 1, pp. 100–111, Jan. 2020.
- [19] 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.
- [20] X. Wang, Y. Feng, R. Song, Z. Mu, and C. Song, "Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data," *Inf. Fusion*, vol. 12, pp. 1–18, 2022.
- [21] J. Li, Y. Liu, R. Song, Y. Li, K. Han, and Q. Du, "Sal² 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.
- [22] 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.
- [23] 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.
- [24] X. Zhao et al., "Joint classification of hyperspectral and LiDAR data using hierarchical random walk and deep CNN architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7355–7370, Oct. 2020.
- [25] 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.
- [26] H. Song et al., "Joint classification of hyperspectral and LiDAR data using binary-tree transformer network," *Remote Sens.*, vol. 15, no. 11, 2023, Art. no. 2706.
- [27] Y. Hu, H. He, and L. Weng, "Hyperspectral and LiDAR data land-use classification using parallel transformers," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 703–706.
- [28] Y. Zhang, Y. Peng, B. Tu, and Y. Liu, "Local information interaction transformer for hyperspectral and LiDAR data classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1130–1143, 2023.
- [29] 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, 2024, Art. no. 111190.
- [30] J. Feng, J. Zhang, and Y. Zhang, "Multi-view feature learning and multi-level information fusion for joint classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5528613.
- [31] Z. Xue, X. Yu, X. Tan, B. Liu, A. Yu, and X. Wei, "Multiscale deep learning network with self-calibrated convolution for hyperspectral and LiDAR data collaborative classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514116.
- [32] S. Roy, A. Deria, D. Hong, M. Ahmad, A. Plaza, and J. Chanussot, "Hyperspectral and LiDAR data classification using joint CNNs and morphological feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530416.
- [33] S. Fang, K. Li, and Z. Li, "S² ENet: Spatial-spectral cross-modal enhancement network for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6504205.
- [34] S. Jia et al., "Multiple feature-based superpixel-level decision fusion for hyperspectral and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1437–1452, Feb. 2021.
- [35] H. Zhang, J. Yao, L. Ni, L. Gao, and M. Huang, "Multimodal attention-aware convolutional neural networks for classification of hyperspectral and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3635–3644, 2023.
- [36] W. Dong, T. Zhang, J. Qu, S. Xiao, T. Zhang, and Y. Li, "Multibranch feature fusion network with self-and cross-guided attention for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530612.
- [37] L. Zhou, J. Geng, and W. Jiang, "Joint classification of hyperspectral and LiDAR data based on position-channel cooperative attention network," *Remote Sens.*, vol. 14, no. 14, 2022, Art. no. 3247.
- [38] D. Xiu, Z. Pan, Y. Wu, and Y. Hu, "MAGE: Multisource attention network with discriminative graph and informative entities for classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539714.
- [39] Y. Feng, L. Song, L. Wang, and X. Wang, "DSHFNet: Dynamic scale hierarchical fusion network based on multi-attention for hyperspectral image and LiDAR data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5522514.
- [40] J. Li, Y. Ma, R. Song, B. Xi, D. Hong, and Q. Du, "A triplet semisupervised deep network for fusion classification of hyperspectral and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540513.
- [41] D. Hong, L. Gao, R. Hang, B. Zhang, and J. Chanussot, "Deep encoder-decoder networks for classification of hyperspectral and LiDAR data," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art. no. 5500205.
- [42] 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.
- [43] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "FusAtNet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and LiDAR classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 416–425.
- [44] 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.
- [45] Z. Xue, X. Tan, X. Yu, B. Liu, A. Yu, and P. Zhang, "Deep hierarchical vision transformer for hyperspectral and LiDAR data classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3095–3110, 2022.
- [46] Y. Zhang et al., "Multimodal transformer network for hyperspectral and LiDAR classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514317.
- [47] M. Zhang, F. Gao, T. Zhang, Y. Gan, J. Dong, and H. Yu, "Attention fusion of transformer-based and scale-based method for hyperspectral and LiDAR joint classification," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 650.
- [48] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214.
- [49] J. Bai et al., "Hyperspectral image classification based on multibranch attention transformer networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535317.
- [50] S. Mei, C. Song, M. Ma, and F. Xu, "Hyperspectral image classification using group-aware hierarchical transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539014.
- [51] Y. Peng, Y. Zhang, B. Tu, Q. Li, and W. Li, "Spatial Cspectral transformer with cross-attention for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537415.
- [52] J. Zou, W. He, and H. Zhang, "LESSFormer: Local-enhanced spectral-spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535416.
- [53] R. Song, Y. Feng, W. Cheng, Z. Mu, and X. Wang, "BS2T: Bottleneck spatial-spectral transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532117.
- [54] Z. Xue, Q. Xu, and M. Zhang, "Local transformer with spatial partition restore for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4307–4325, 2022.
- [55] K. Ni and C. Yuan, "GPCNet: Global-context pyramidal and class-balanced network for high-resolution SAR remote sensing image classification," *J. Appl. Remote Sens.*, vol. 16, no. 3, 2022, Art. no. 036510.
- [56] I. Duta, L. Liu, F. Zhu, and L. Shao, "Pyramidal convolution: Rethinking convolutional neural networks for visual recognition," 2020, *arXiv:2006.11538*.
- [57] H. Fan et al., "Multiscale vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6824–6835.
- [58] L. Meng et al., "AdaViT: Adaptive vision transformers for efficient image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12299–12308.



Kang Ni (Member, IEEE) received the M.S. degree in electronics and communications engineering from the Changchun University of Technology, Jilin, China, in 2016, and the Ph.D. degree in signal and information processing from the Nanjing University of Aeronautics and Astronautics, Jiangsu, China, in 2020.

He is a Lecturer with the School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include machine learning, SAR image processing, and computer vision.

Dr. Ni is a Member with the Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing, China.



Zhizhong Zheng received the B.S. degree in electronic engineering from Jilin University, Changchun, China, in 2001, and the Ph.D. degree in control engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2021.

He is currently a Team Leader of hyperspectral remote sensing and intelligent perception of the Nanjing University of Posts and Telecommunications, Nanjing, China. He has developed a series of equipments, such as space remote sensing camera, airborne hyperspectral imaging instrument, UAV hyperspectral

remote sensing system, little field spectrometer, and core scan spectral imaging system, as well as spectral data processing and information extraction software.



Duo Wang is currently working toward the bachelor's degree in artificial intelligence with the Nanjing University of Posts and Telecommunications, Nanjing, China.

His research interests include machine learning, multisource remote sensing image processing, and computer vision.



Peng Wang (Senior Member, IEEE) received the B.E. degree in microelectronics and the Ph.D. degree in information and communication engineering from the College of Information and Communications Engineering, Harbin Engineering University, Harbin, China, in 2012 and 2018, respectively.

He is currently an Associate Professor with the Key Laboratory of Radar Imaging and Microwave Photonics, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He is also a Hong Kong Scholar with the Department of

Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong. In 2016, he was a Visiting Ph.D. Student with the Grenoble Images Parole Signals Automatics Laboratory, Grenoble Institute of Technology, Saint Martin d Hres, France. He has authored two books and more than 50 articles. His research interests include remote-sensing imagery processing and machine learning.

Dr. Wang is a Reviewer of more than 20 international journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Remote Sensing of Environment*, IEEE TRANSACTIONS ON IMAGE PROCESSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS.