Attention Head Interactive Dual Attention Transformer for Hyperspectral Image Classification

Cuiping Shi, *Member, IEEE*, Shuheng Yue, and Liguo Wang, *Member, IEEE*

***Abstract*— In recent years, transformer has attracted the attention of many researchers in the field of remote sensing due to its ability to model global information. However, it is difficult to extract local features such as textures and edges of images, thereby limiting the performance of transformer-based hyperspectral image classification (HSIC). Currently, most existing transformer models for HSIC improve their performance by combining the powerful feature extraction ability of convolution, which also introduces a large number of trainable parameters and increases model complexity. To address this issue, this article proposes a dual attention transformer for attention head interaction (DAHIT) for HSIC. First, a spatial local bias module (SLBM) was designed in the spatial branch, which introduces local priors to extract local features effectively without introducing numerous trainable parameters. Then, an attention head interaction module (AHIM) was proposed, which can make the interaction of information obtained by different attention heads. Finally, a diagonal mask multiscale dual attention module (DAM) was constructed in the spectral branch to enhance the attention to the correlation among different spectral bands through diagonal masks and then to extract features at different scales through feature vectors. Through a series of experiments, the proposed DAHIT is evaluated on four commonly used HSI datasets. The experimental results show that compared with other advanced methods, the proposed DAHIT method exhibits excellent classification performance, demonstrating the effectiveness of the proposed method in HSIC.**

***Index Terms*— Attention head, hyperspectral image classification (HSIC), multihead attention, transformer.**

I. Introduction

HYPERSPECTRAL images (HSIs) receive electromagnetic signals within a narrow and dense wavelength range through advanced sensors, obtaining spectral information from hundreds of bands. HSIs have the characteristics of spatial–spectral integration, containing rich spectral information and spatial information of land cover. Hyperspectral data have been widely used in fields, such as environmental monitoring [1], biomedicine [2], precision agriculture [3], [4], and geological prospecting [5]. The purpose of HSI classification (HSIC) is to assign unique category labels to each sample of HSIs based on a given set of ground cover categories.

After decades of development, both manual feature extraction methods and deep learning-based feature extraction methods have made great progress in HSIC tasks. Traditional manual feature extraction methods can select the most representative band feature set through direct identification, such as constrained band selection (CBS) [6] and clustering-based band selection [7]. In addition to directly selecting the band feature set, traditional manual feature extraction methods can also learn representative feature sets through linear or nonlinear transformations, such as principal component analysis (PCA) [8] and local linear embedding (LLE) [9]. Unlike traditional manual shallow feature extraction methods, deep learning-based feature extraction methods extract discriminative features in a hierarchical manner. For example, convolutional neural networks (CNNs) [10], [11], [12], graph CNNs (GCNs) [13], [14], [14], [16], graph attention networks (GATs) [17], [18], generate adversarial networks (GANs) [19], [20], and transformer [21], [22].

In recent years, in order to enable the classification network of HSIs to be trained in one domain and achieve satisfactory classification results in another different domain, the research on cross-scene HSIC has attracted wide attention. Based on the idea of domain generalization, Zhang et al. [51] proposed a single-source domain expansion network (SDENet). SDENet is trained in the source domain through generative adversarial learning and tested in the target domain. The semantic encoding and morphological encoder are used as generators, while the discriminator using supervised contrastive learning (CL) is used to learn domain-invariant representations. In order to better utilize HSIs and light detection and ranging (LiDAR) data, Zhang et al. [52] proposed a structural optimization transmission network (SOT-Net). The complementary information collaboration manner and the redundancy exclusion operator was be redesigned in SOT-Net. The correlation of multisource semantics is enhanced by SOT Net. To address the issues of domain bias and prototype instability in training and testing datasets in few-shot learning (FSL), Liu et al. [55] proposed a refined prototypical CL network for FSL (RPCL-FSL). RPCL-FSL integrates supervised CL and FSL into end-to-end networks for small sample HSIC. To alleviate

Manuscript received 24 April 2024; revised 9 June 2024; accepted 10 July 2024. Date of publication 15 July 2024; date of current version 26 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42271409 and in part by Heilongjiang Science Foundation Project of China under Grant LH2021D022. *(Corresponding author: Shuheng Yue.)*

Cuiping Shi is with the College of Information Engineering, Huzhou University, Huzhou 313000, China, and also with the Department of Communication Engineering, Qiqihar University, Qiqihar 161000, China (e-mail: shicuiping@qqhru.edu.cn).

Shuheng Yue is with the Department of Communication Engineering, Qiqihar University, Qiqihar 161000, China (e-mail: 2021910320@qqhru.edu.cn).

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

Digital Object Identifier 10.1109/TGRS.2024.3427769

domain transfer in FSL, RPCL-FSL designed a fusion training strategy to reduce feature differences between training and testing datasets. To address the issue of insufficient labeled samples in HSIC tasks, Liu et al. [57] proposed a method based on self-supervised learning of spectral masking (SSLSM) for HSIC. The training of SSLSM includes two steps: self-supervised pretraining and fine-tuning. SSLSM uses spectral reconstruction as a pretext task, and by reconstructing masked spectra, the model's feature extraction ability in classification tasks during fine-tuning is greatly improved.

In order to better reduce the dimensionality of HSI spectra, Gao et al. [49] proposed a deep learning-based band selection method called DLBSTD, in which the representative bands with key target information can be extracted. In order to better address the issue of class imbalance in HSIs, Gao et al. [50] proposed a spectral aggregation and separation network (SASN) with a target band random mask (TBRM). Multiple representative background selection strategies (MRBS) are used in the training sample set of SASN to select diverse and representative background training sets. GCNs are widely used in HSI classification tasks due to their powerful feature learning ability. Qin et al. [53] extended the original GCN to the second order, taking into account spectral and spatial neighborhood information. To solve the problem of adjacency matrix consuming a large amount of memory resources in GCN. Liu et al. [54] proposed a fast dynamic graph convolution and CNN parallel network (FDGC), which can adaptively capture topology information and extend GCN to large graphs. Due to the complex spatial variability of HSIs, existing graph construction based on HSIs is always inaccurate, and graph structure-based HSI methods often suffer from oversmoothing issues. To address the aforementioned issues, Yang et al. [58] proposed a deep network with adaptive graph structure integration (DNAGSI). DNAGSI can dynamically learn the graph structure of HSIs and improve the robustness of the model. Moreover, a joint loss with center loss is proposed in DNAGSI to learn the similarity relationship between HSI pixels and aggregate the intraclass graph features.

In the field of computer vision, CNNs have become the most widely used backbone networks. For CNN-based methods, researchers have considered and designed many effective networks from the perspective of network framework [23], [24], [25], [26], [27], [28] and effective module utilization [29], [30], [31]. Due to the ability of 3-D convolution to extract spatial–spectral joint features, He et al. [32] designed a multiscale method based on 3-D convolution. Unlike traditional convolutions, Zhou et al. [33] proposed quaternion CNNs (QCNNs). This network can extract quaternion features that not only contain contextual information but also utilize quaternion algebra within quaternion element units to express structural information. Although the above methods can extract local features of the image, the limited receptive field limits the model's feature extraction ability. Shi et al. [34] designed an expansion convolution network (ECNet) based on dilation convolution, which expands the receptive field by stacking dilation convolutional layers and can extract more contextual features. Paoletti et al. [25] proposed a deep residual pyramid network (PyResNet). The depth of the network is gradually increased by the form of residuals, which not only allows to obtain a large receptive field but also avoids the hindrance of the model to convergence. Since the shape of the conventional convolutional receptive field is fixed, this may lead to the spatial structure information being ignored. To address the above problems, Shang et al. [35] proposed an HSIC method based on multiscale cross-branch response and second-order channel attention (MCRSCA). Zhu et al. [36] proposed a DHCNet method based on deformable convolution. Deformable convolution can make the receptive field of the convolution variable and avoid the neglect of spatial structure information. In order to combine the advantages of 2-D convolution and 3-D convolution, Roy et al. [37] proposed a hybrid spectral CNN (hybrid SN). To solve the problem of GCN that only extracting features from superpixel nodes and pixel-level features was ignored, Yu et al. [46] proposed a two-branch deep GCN (TBDGCN), in which the pixel level features can be extracted by CNNs. In TBGCN, a GCN module with DropEdge technology and residual connections was proposed to solve the problem of oversmoothing in GCN. Huang et al. [47] proposed a two-branch attention adaptive DA (TAADA) network that effectively utilizes the spectral–spatial joint features of domain adaptation (DA) in HSIs. Although CNNs have significant advantages in local feature extraction, due to the limitations of receptive fields, they are unable to obtain a global receptive field and capture global contextual information.

HSIs have high spatial resolution and hundreds of spectral bands. It is difficult to model the long-range dependence of HSIs space and spectrum using CNN-based methods. In recent years, transformer has been widely used in the field of remote sensing due to its multihead self-attention (MHSA) mechanism, achieving results comparable to or even superior to CNNs. Vision transformer (VIT) proposed by Dosovitskiy et al. [38] is the most classical visual transformer. Considering that spectra have sequence properties, Hong et al. [39] proposed a spectral transformer (SF) from the perspective of spectral sequences. It designed a grouping embedding module that can learn information from adjacent HSI spectrum and extract local features of spectrum. The group-aware hierarchical transformer (GAHT) proposed by Mei et al. [40] solves the problem of excessive discretization of features extracted by MHSA in HSIC. In order to better represent advanced semantic information and extract spectral–spatial features, spectral–spatial feature tokenization transformer (SSFTT) [41] has been proposed. SSFTT converts shallow spectral–spatial information into labeled semantic information through the Gaussian semantic weighting module, which can make the deep semantic features represented more in line with the distribution characteristics of the sample. Roy et al. [41] proposed the spectral–spatial morphological attention transformer (MorphFormer), which improves the interaction of structural and morphological information between HSI Token and CLS Token through spectral–spatial morphological convolution operations. Unlike the transformer-based methods mentioned above, Zhang et al. [42] proposed a convolutional transformer mixer (CTMixer), which combines the advantages of CNNs

and transformers to construct a group of parallel residual convolutions to extract shallow features and uses transformer to extract deep features, achieving an effective combination of CNNs and transformers. In order to fully utilize the rich spectral, spatial, and semantic information in HSIs, Xie et al. [48] proposed a new semantic and spatial–spectral feature fusion transformer (S3FFT). Spatial attention and channel attention are used by S3FFT to extract shallow spatial–spectral features, while transformer-based modules are used by S3FFT to extract advanced fusion features.

Transformer includes MHSA mechanism and feedforward neural network (FFN), where MHSA is the core component of transformer. The high-level semantic information of the image can be extracted by the transformer, but the HSIC requires not only the high-level semantic information of the image but also the extraction of local features of the image [43]. However, it is difficult for transformer to extract localized features of an image. In order to better extract the local features of the image and improve the classification performance of HSIs, in this article, a dual attention transformer for attention head interaction (DAHIT) is proposed. Specifically, in order to enhance the transformer's ability to represent the local features of HSIs, DAHIT was designed with two branches for feature extraction for spatial and spectral information, respectively, and modules for enhancing the local feature extraction ability were also designed in each branch. First, in the spectral branch, a diagonal mask multiscale dual attention module (DAM) is proposed. DAM masks the attention information on the diagonal before SoftMax activation of the attention matrix, allowing the attention module to pay more attention to the relationships between adjacent pixels. To avoid information loss caused by covering diagonal lines, DAM designed an uncovered attention block. In diagonal masked attention blocks, multiscale feature extraction is performed on the value vector, enabling diagonal masked attention blocks to obtain richer feature information. Then, in the spatial branch, in order to extract spatial local features, the spatial local bias module (SLBM) is proposed. SLBM utilizes maximum pooling to extract spatial local prior information of the image and then cascades and fuses it with features that have not been extracted locally to add local prior information to the classification features. Next, this article proposes an attention head interaction module (AHIM) for information interaction between different attention heads, improving the model's feature representation ability. Finally, cascade the features extracted from the two branches and perform HSIC.

The main contributions of this article are given as follows.

1) In this article, a DAM method is designed for extracting local spectral features. The diagonal mask in DAM is utilized to enhance attention to the correlation between different spectral bands and to extract the local contextual relationships of images. To avoid the information loss, an unmasked attention block is constructed to supplement attention information.
2) In order to extract spatial local features, an SLBM is constructed in the spatial branch. By utilizing the advantage of maximum pooling to extract local features such as image textures and edges, local prior information is introduced into the spatial branch to enhance its ability to extract local features.
3) This article proposes an AHIM module that interacts with attention information captured between different attention heads of MHSA. The single distribution of attention heads is changed into a mixed distribution, and the representation ability of MHSA is enhanced.

II. Methodology

The overall framework of the proposed DAHIT is shown in Fig. 1. In Fig. 1, C represents the number of output channels extracted by shallow layers in the spectral branch, C1 represents the number of output channels extracted by shallow layers in the spatial branch, C1/2 is the number of output channels in the intermediate layer, and C + 1 is the number of feature channels cascaded by cls-tokens in the spatial transformer. Assume that the network input is $X \in \mathbb{R}^{H\times W\times B}$, where H and W denote the spatial size of the HSIs and B is the number of spectral bands of the HSIs. First, in the data preprocessing stage, PCA is performed on the spectral dimension to obtain $X \in \mathbb{R}^{H\times W\times b}$. b ($b < B$) represents the number of spectral bands after PCA dimensionality reduction. Then, a two-branch network is designed during the feature extraction stage with spectral and spatial branches. In order to extract the local information of the image, an SLBM and a multiscale diagonal mask DAM are constructed in the spatial and spectral branches, respectively. The AHIM is designed to enhance the representational ability of the DAHIT model. Finally, the features extracted from the two branches are cascaded and then classified through the classification layer to obtain the predicted label set $Y \in \mathbb{R}^{H\times W} = \{y_i | y_1, y_2, \ldots, y_N\}$, where N is the value of maximum label.

A. Attention Head Interaction Module (AHIM)

Currently, most transformer models used in the field of remote sensing improve their performance in downstream tasks by combining the powerful feature extraction ability of convolution, such as S2FTNet [56]. A new attention module AHIM of DAHIT proposed in this article is designed from the perspective of attention head interaction. The single low-rank distribution of a single attention head is transformed into a mixed distribution of different attention head interactions, enhancing the model's feature representation ability.

The multihead mechanism in MHSA can extract attention information from different subspaces. However, the extracted attention information exists in isolation from each other, which largely limits the feature representation ability of MHSA. Therefore, an AHIM is proposed in this article. Although traditional MHSAs can learn attention information from different subspaces, they are isolated from each other. Unlike MHSA, the information extracted from different subspaces of AHIM is processed by the AHIM for interaction between different subspace information. The feature extraction capability of the network is enhanced by information interaction operations in different subspaces. AHIM interacts information between different attention heads on query, key, and value vectors. Query vectors are used to measure the degree of correlation

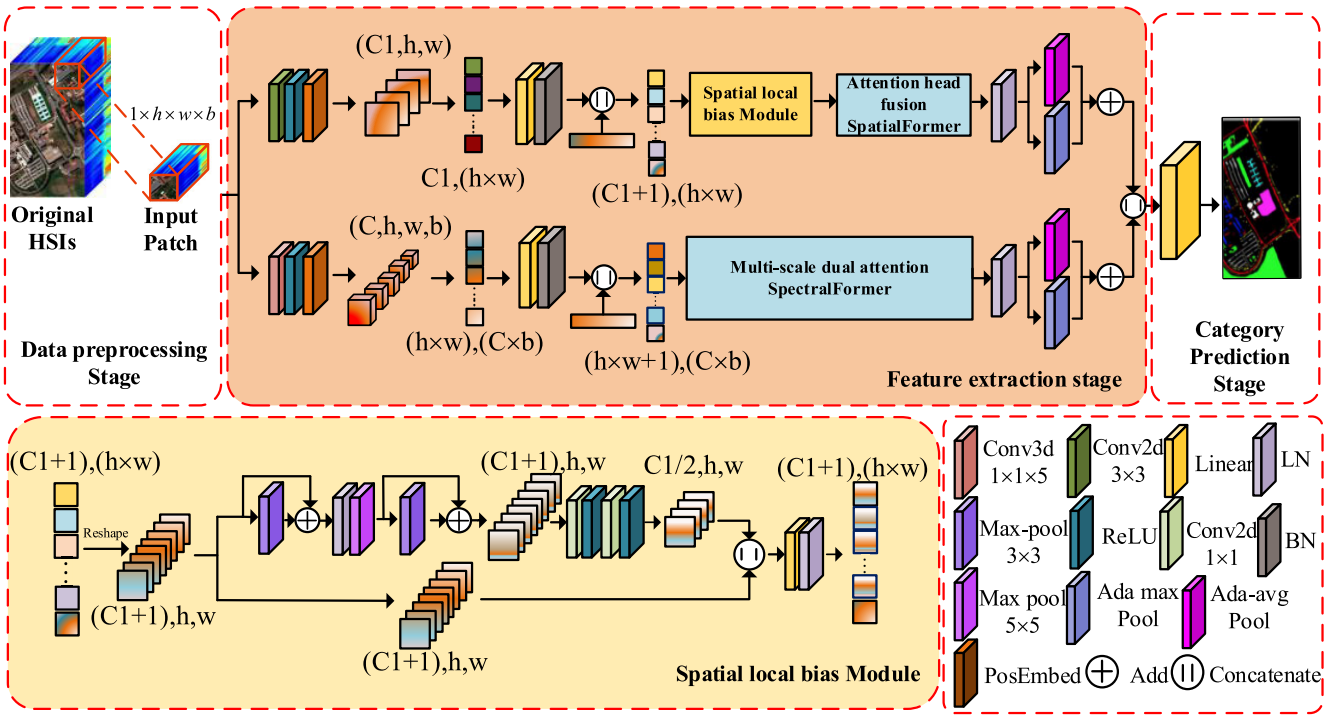

Fig. 1. Overall framework of the proposed DAHIT.

between key vectors at other positions. According to the correlation between the query vector and the key vector, the value vector is used for information aggregation. The different subspace information in the query and key obtained from traditional MHSA mapping exists in isolation. In AHIM, the features learned from different subspaces of query and key are cross mapped to a larger feature representation space. After being mapped to a larger feature representation space, information exchange occurs in each subspace. The features in different subspaces of query and key are not isolated. In order to ensure sufficient interaction of subspace information, query and key are once again mapped back from the large feature representation space to the original subspace. From a single low-rank distribution of a single attention head to a mixed distribution of interactions between different attention heads, the representation ability of the original distribution is greatly enhanced. Fig. 2 shows the structure diagram of AHIM.

For AHIM, first, input $X_T \in \mathbb{R}^{n\times c}$ obtained by linear mapping to $Q, K, V \in \mathbb{R}^{h\times n\times c}$. Then, in order to enhance the representation ability of attention, the information between different attention heads of Q, K, and V is interacted. Through two 2-D convolutions with different numbers of channels and convolution kernel size of 1×1, the information between different attention heads is mixed and mapped from the small attention space to the large one, and the information of different attention heads completes the preliminary interaction. In order to fully interact with the information of the attention head and ensure that the number of attention heads before and after the interaction is the same, a second interaction is performed on the attention information after the initial interaction, mapping from a large attention space to a small attention space. The calculation process of AHIM is

$$Q_I, K_I, V_I = \mathrm{HI}(Q, K, V) \tag{1}$$

$$\mathrm{HI} = \delta(\delta(x * W_{1\times1} + b_{1\times1}) * W_{1\times1} + b_{1\times1}) \tag{2}$$

$$\mathrm{SA} = \mathrm{soft\,max}\left(\frac{Q_I K_I}{\sqrt{d_k}}\right) V_I \tag{3}$$

$$\mathrm{AHIM} = \mathrm{Concat}(\mathrm{SA}_1, \mathrm{SA}_2, \ldots \mathrm{SA}_h) W_o \tag{4}$$

where δ, $W_{1\times1}$, and $b_{1\times1}$ denote the weights and bias of the nonlinear activation function ReLU, and convolution kernel size 1×1, respectively; Concat$(\cdot)$ denotes the cascade function; and W_o is the linear mapping weights.

B. Diagonal Mask Multiscale Dual Attention Module (DAM)

MHSA is a core component of transformer, which can model long-distance dependencies. However, the classification task of HSIs requires not only modeling long-distance dependencies but also modeling local information. Therefore, in this article, a diagonal mask multiscale DAM is designed. Unlike other multiscale attention methods, DAM only extracts multiscale features from values, not from query and key vectors. This is because extracting multiscale features from query and key vectors can lead to an increase in attention logits, resulting in model instability and gradient explosion. The value vector is a feature vector weighted by attention, and multiscale feature extraction of the value vector not only avoids the problem of gradient explosion but also achieves the extraction of rich multiscale features, thus improving the classification performance of the model. DAM is a dual-branch attention module, including unmasked attention block and diagonal masked multiscale attention block. First, diagonal masked

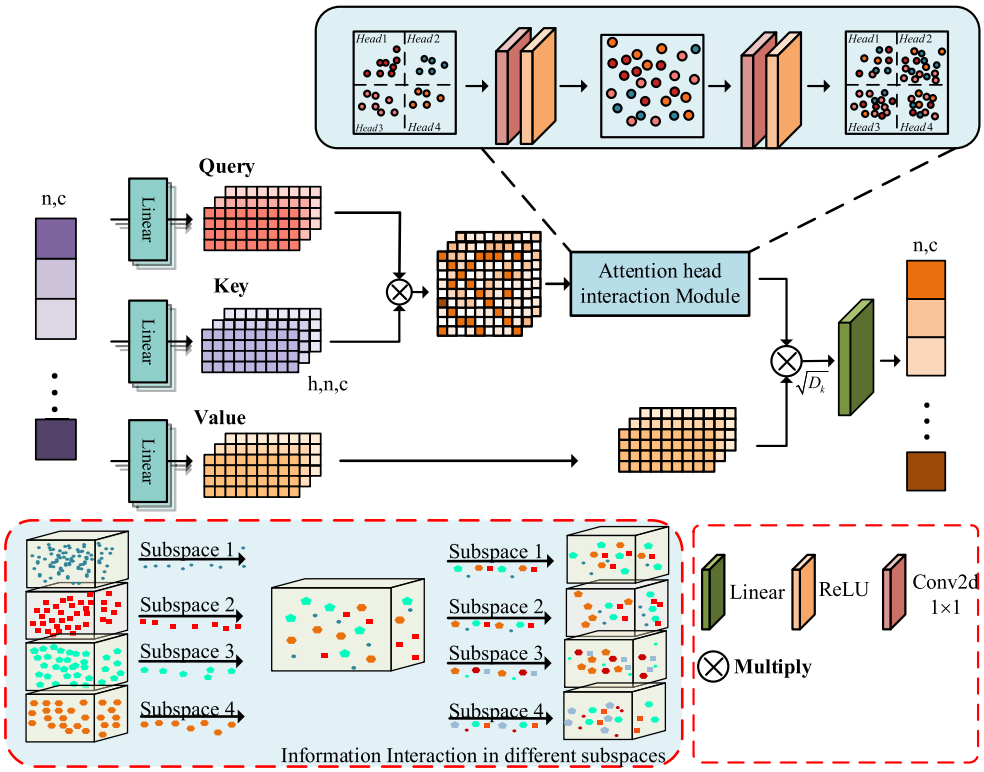

Fig. 2. Structure diagram of AHIM.

multiscale attention blocks enhance attention to the correlation between different bands by masking the information on the diagonal in the attention matrix, which can extract local spectral information. Then, considering the advantages of multiscale features, multiscale feature extraction was performed on the value vector using convolutional blocks of different scales in the diagonal mask attention block. Query, key, and value are mapped by MHSA, where query and key calculate the correlation between different positional features through vector multiplication, and the value is weighted with attention based on the calculated correlation. Therefore, in DAM, only multiscale feature extraction is considered for the value, avoiding unnecessary interference caused by multiscale convolution on the correlation calculation of query and key, which leads to excessive attention logit and suppresses the feature extraction ability of the model. Next, considering that the attention information on the diagonal in the diagonal masked multiscale attention block is masked, in order to avoid key information loss, an unmasked attention block is designed. Unmasked attention blocks can not only avoid the loss of key information but also model the long-range dependencies of the spectrum. Finally, the features extracted from the dual attention branch are fused through cascading and linear layers. Fig. 3 shows the structural diagram of DAM.

MHSA obtains the correlation matrix between different spectral bands by multiplying the query vector and the key vector and then nonlinearly activates the obtained correlation matrix using the SoftMax function. After activation, the values of the correlation matrix are all between 0 and 1, and the sum of the correlation values in each column is 1. Among them, the correlation values on the diagonal of the correlation matrix represent the correlation between the spectral bands themselves. In order to enhance attention to the correlation between different spectral bands, we designed a mask to cover the values of the correlation diagonal. After SoftMax activation, bands with strong correlation will get larger correlation values. Fig. 4 shows the activation process of the correlation matrix.

To avoid the problem of insufficient feature extraction at a single scale, a 2-D convolution is used to extract multiscale features from the value vector in diagonal masked attention blocks. This not only extracts features at different scales but also extracts local features of the spectrum. The calculation process is

$$V_m = \mathrm{MLP}(\mathrm{Concat}(V \Theta W_{3\times3} + b_{3\times3}, V \Theta W_{5\times5} + b_{5\times5})) \quad (5)$$

$$\mathrm{MLP}(x) = \delta(f_{\mathrm{BN}}(x \Theta W_{1\times1} + b_{1\times1})) \quad (6)$$

$$\mathrm{DMSA}(Q, K, V) = \mathrm{soft\,max}\left(\frac{DM(QK^T)}{\sqrt{d_k}}\right)V_m \quad (7)$$

$$\mathrm{DMMHSA} = \mathrm{Concat}(\mathrm{DMSA}_1, \mathrm{DMSA}_2, \ldots, \mathrm{DMSA}_h)W_o \quad (8)$$

where $W_{3\times3}$ and $W_{5\times5}$ represent the weights for convolutional kernel with sizes 3×3 and 5×5, respectively; $b_{3\times3}$ and $b_{5\times5}$ represent the biases for convolutional kernel with sizes 3×3 and 5×5, respectively; δ and f_{BN} represent the nonlinear activation functions ReLU and batch normalization, respectively; Θ is the convolution operator; and $DM(\cdot)$ is the diagonal mask function.

Since the diagonal masked attention block covers the correlation information on the diagonal, to avoid the problem of critical information being lost due to masking, an unmasked attention block is designed to extract the global features of

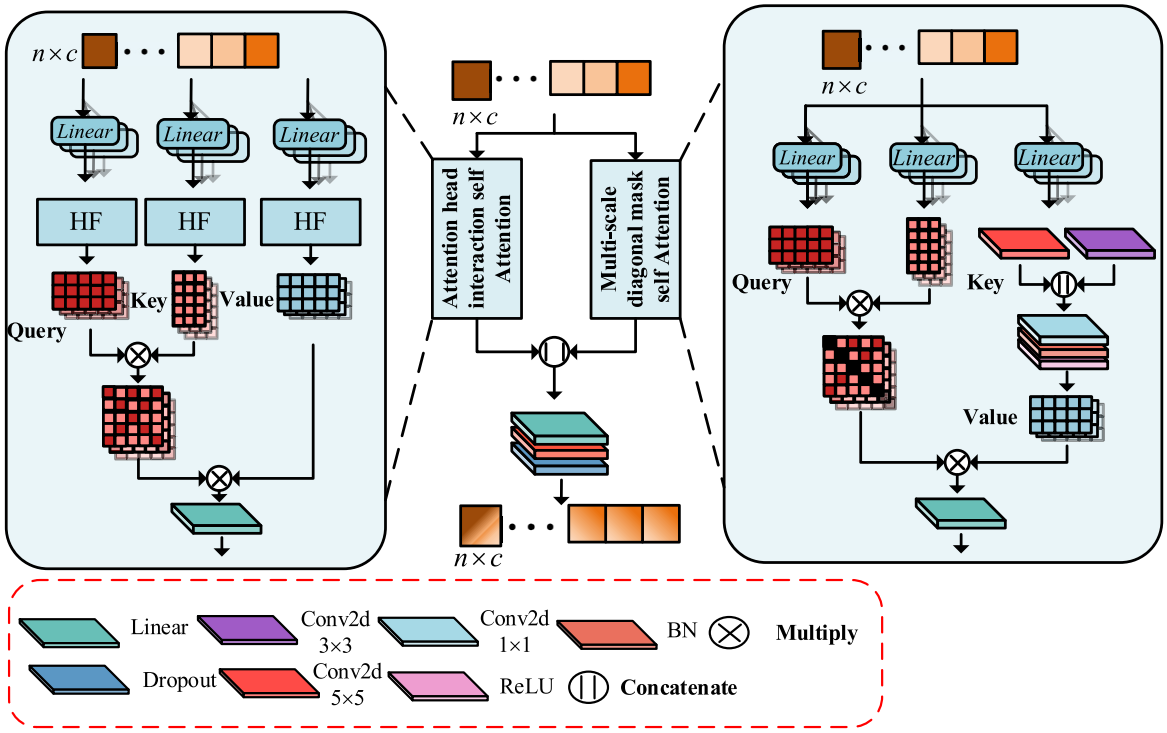

Fig. 3. Structural diagram of DAM.

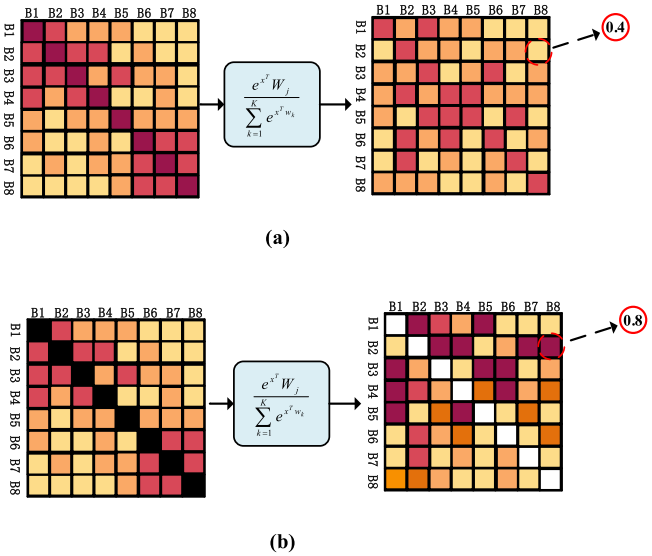

Fig. 4. Process of activating the correlation matrix. (a) Process of activating the correlation matrix of nondiagonal masks. (b) Process of activating the correlation matrix of diagonal masks.

the spectrum as a separate branch of attention. The unmasked attention block has no diagonal mask, so the correlation information on the diagonal is not lost. The AHIM module is used in unmasked attention blocks, where different attention heads interact with each other, resulting in more accurate attention information extracted from the attention blocks and improving the model's expressive ability. Therefore, unmasked attention blocks can not only extract global features of the spectrum but also solve the problem of information loss on the diagonal of diagonally masked attention blocks. The calculation process is given as follows:

$$Q_I, K_I, V_I = \mathrm{HI}(Q, K, V) \tag{9}$$

$$\mathrm{SA} = \mathrm{soft\,max}\left(\frac{Q_I K_I^T}{\sqrt{d_k}}\right) V \tag{10}$$

$$\mathrm{MHSA} = \mathrm{Concat}(\mathrm{SA}_1, \mathrm{SA}_2, \ldots, \mathrm{SA}_h) W_o. \tag{11}$$

The features obtained from the two attention branches are fused through cascading and linear layers. The computational process is

$$F_{DA} = \mathrm{Concat}\left(F_{A1}, F_{A2}\right) \tag{12}$$

$$F_A = \mathrm{Dropout}\left(\mathrm{f}_{gelu}(\mathrm{F}_{DA} * \mathrm{W}_o + \mathrm{b}_o)\right) \tag{13}$$

where Dropout$(\cdot)$ and $f_{\mathrm{gelu}}(\cdot)$ represent the Dropout function and the Gaussian error linear cell, respectively, and b_o represents the bias of the linear mapping.

C. Spatial Local Bias Module (SLBM)

Local spatial feature extraction is crucial for HSIC. However, although traditional transformer methods can model long-distance dependencies, they cannot effectively extract local spatial features. Currently, most transformer-based HSIC methods extract local features, such as texture and edges through convolution, e.g., HIT [22] and CTMixer [42]. However, introducing convolution into transformer resulted in a significant increase in the number of trainable parameters and an increase in model complexity. This article proposes an SLBM in the spatial branch. SLBM can enable the network model to perform local induction bias and allow the model to pay more attention to local feature extraction of images. Fig. 5 shows the structural diagram of SLBM.

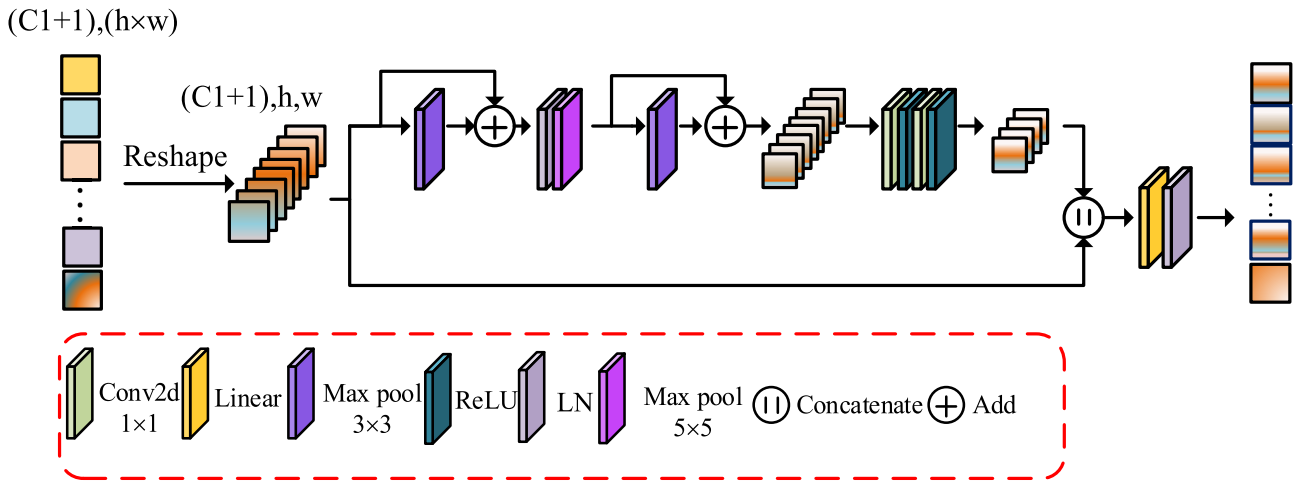

Fig. 5. Structural diagram of SLBM.

SLBM is a local feature extraction module with a dual-branch structure. Suppose that the input of SLBM is $F_T \in \mathbb{R}^{65\times(h\times w)}$, and first, to facilitate the extraction of local features of the image, the vector $F_T \in \mathbb{R}^{65\times(h\times w)}$ is converted into a matrix $F'_T \in \mathbb{R}^{65\times h\times w}$. Then, maximum pooling with sizes of 5×5 and 3×3 is utilized to extract spatial local features. In order to fully extract the local features of the image, the local features are first extracted by maximum pooling with a smaller window 3×3 and finally with a larger window 5×5. In the process of extracting local features, residual connections were added to avoid information loss caused by pooling. The local feature extraction process of SLBM is

$$F'_T = \text{Reshape}(F_T) \tag{14}$$

$$F''_T = f_{5\times5}\left(\text{LN}\left(\text{Max}_{3\times3}\left(F'_T\right)\right)\right) \tag{15}$$

$$\text{Max}_{3\times3}(x) = f_{3\times3}(x) + x \tag{16}$$

where Reshape$(\cdot)$ denotes the shape transformation function; $f_{3\times3}$ and $f_{5\times5}$ denote the maximum pooling with size 3×3 and 5×5, respectively; and LN$(\cdot)$ denotes the layer normalization.

Then, the extracted local features are fused into different channels through point convolution, and the channels are shrunk to output feature $F_L \in \mathbb{R}^{c_1\times h\times w}$. Finally, the extracted local features $F_L \in \mathbb{R}^{c_1\times h\times w}$ and $F'_T \in \mathbb{R}^{65\times h\times w}$ are fused by cascading and linear layers to obtain $F_{LT} \in \mathbb{R}^{65\times(h\times w)}$. The calculation procedure is

$$F_L = \delta\left(\delta\left(F''_T * W_{1\times1} + b_{1\times1}\right) * W_{1\times1} + b_{1\times1}\right) \tag{17}$$

$$F_m = \text{Concat}\left(F_L, F'_T\right) \tag{18}$$

$$F_{LT} = \text{LN}(F_m * W_o + b_o). \tag{19}$$

D. Implementation Process

Taking the Salinas dataset as an example, this section describes the implementation details of the proposed DAHIT algorithm. The spatial size of the Salinas dataset is 512×217 and the number of spectral bands is 200. First, the input data $X \in \mathbb{R}^{S\times S\times30}$ are obtained after data preprocessing. Then, the shallow spatial features are extracted in the spatial branch by a 2-D convolution block to obtain $X_{sa} \in \mathbb{R}^{64\times S\times S}$, and the shallow spectral features are extracted in the spectral branch by a 3-D convolution block to obtain $X_{se} \in \mathbb{R}^{C\times S\times S\times30}$. The shallow features X_{sa} and X_{se} obtained from the two branches are input to the transformer of the corresponding branch to extract the deep features, respectively. For the spectral branch, $X'_{se} \in \mathbb{R}^{169\times64}$ is obtained by flattening and linearly mapping $X_{se} \in \mathbb{R}^{C\times S\times S\times30}$, and then, local and global features of the spectrum are extracted using a transformer with DAM. For the spatial branch, $X'_{sa} \in \mathbb{R}^{64\times64}$ is obtained by flattening and linearly mapping $X_{sa} \in \mathbb{R}^{64\times S\times S}$. Spatial local information is extracted by SLBM, and then, transformer with AHIM is utilized to extract spatial global features. The features extracted from the two branches are dimensionally reduced through average pooling and maximum pooling, respectively. Finally, these features are cascaded to obtain $X_c \in \mathbb{R}^{128,}$. Algorithm 1 is the implementation details of the proposed DAHIT method.

III. Experimental Results and Analysis

In order to evaluate the effectiveness of the proposed DAHIT method, a series of experiments were conducted on four commonly used datasets, i.e., Indian Pines (IP) dataset, Pavia dataset, and Salinas dataset, and compared our method with some state-of-the-art methods.

A. Dataset Description

To verify the effectiveness of the proposed method, four widely used HSI datasets were selected in this article, including the IP (IP) dataset, the Salinas Valley (SV) dataset captured by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor, the University of Pavia (UP) dataset captured by the Reflective Optical Spectroscopy Imaging System (ROSIS-3) sensor, and Houston 2013 collected by the HSI analysis team and NCALM on the University of Houston campus and nearby urban areas (as shown in Tables I–IV). The IP, Pavia, Salinas, and Houston 2013 datasets were randomly

Algorithm 1 Process of the DAHIT Method

Inputs: Hyperspectral data $X \in \mathbb{R}^{h\times w\times B}$, labels $Y \in \mathbb{R}^{h\times w}$, number of bands after PCA $b = 30$, the spatial size of the cube $s \times s$, and training sample proportion q%.
Output: The predicted labels for the test dataset.
1: Set the batch size to 64, learning rate lr to 0.005 for Adam optimizer, training rounds T = 200.
2: The output after PCA is $X_{pca} \in \mathbb{R}^{h\times w\times b}$.
3: Divide 3D cube $X_{pca} \in \mathbb{R}^{h\times w\times b}$, and divide the training samples according to a ratio of q% ratio, and the rest are test samples.
4: **for** i = 1 to T do
5: Process the spatial branch with two-dimensional convolutional blocks to obtain $X_{sa} \in \mathbb{R}^{64\times S\times S}$, and the spectral branch with three-dimensional convolutional blocks to obtain $X_{se} \in \mathbb{R}^{C\times S\times S\times 30}$.
6: In the spatial branch, perform flattening and mapping on $X_{sa} \in \mathbb{R}^{64\times S\times S}$ to obtain $X_{se} \in \mathbb{R}^{C\times S\times S\times 30}$.
7: Implementation of SLBM.
8: Implements the Transformer with AHIM.
9: In the spectral branch, perform flattening and mapping on $X_{se} \in \mathbb{R}^{C\times S\times S\times 30}$ to obtain $X'_{se} \in \mathbb{R}^{169\times 64}$.
10: Implements the Transformer with DAM.
11: Features extracted from spectral and spatial branches are dimensionally reduced through pooling, and then cascaded to obtain $X_c \in \mathbb{R}^{128,1}$.
12: Output the prediction category for each sample.
end for

TABLE I
CATEGORY NAMES AND SAMPLE DIVISION OF THE INDIAN PINES DATASET

Category	Label color	Land cover	Training	Test
1		Alfalfa	4	42
2		Corn-notill	142	1286
3		Corn-mintill	82	748
4		Corn	23	214
5		Grass-pasture	48	435
6		Grass-trees	72	658
7		Grass-pasture-mowed	3	25
8		Hay-windrowed	47	431
9		Oats	3	17
10		Soybean-notill	97	875
11		Soybean-mintill	245	2210
12		Soybean-clean	59	534
13		Wheat	20	185
14		Woods	126	1139
15		Bldg-Grass-Tree-Drivers	38	348
16		Stone-Steel-Towers	9	84
Total	/	/	1018	9231

divided into training samples at training ratios of 10%, 1%, 1%, and 5%, respectively. After dividing the training samples in the dataset, the remaining samples are used as test samples.

B. Experimental Setup

1) Evaluation Indicators: All experiments in this article were performed on a platform with an Intel(R) Core (TM) i9-9900K CPU, an NVIDIA GeForce RTX 3090Ti GPU, and 128 GB of random access memory, using the framework of Pytorch. In addition, three common evaluation indicators are chosen to evaluate the classification performance of the mode, i.e., overall accuracy (OA), average accuracy (AA),

TABLE II
CATEGORY NAMES AND SAMPLE DIVISION OF THE PAVIA DATASET

Category	Label color	Land cover	Training	Test
1		Asphalt	66	6565
2		Meadows	186	18463
3		Gravel	20	2079
4		Trees	30	3034
5		Painted metal sheets	13	1332
6		Bare Soil	50	4979
7		Bitumen	13	1317
8		Self-Blocking Bricks	36	3646
9		Shadows	9	938
Total	/	/	423	42353

TABLE III
CATEGORY NAMES AND SAMPLE DIVISION OF THE SALINAS DATASET

Category	Label color	Land cover	Training	Test
1		Brocoil-green-weeds_1	20	1989
2		Brocoil-green-weeds_2	37	3689
3		Fallow	19	1957
4		Fallow-rough-plow	13	1381
5		Fallow-smooth	26	2652
6		Stubble	39	3920
7		Celery	35	3544
8		Grapes-untrained	112	11159
9		Soil-vinyard-develop	62	6141
10		Corn-senesced-green-weeds	32	3246
11		Lettuce-romaine-4wk	10	1056
12		Lettuce-romaine-5wk	19	1908
13		Lettuce-romaine-6wk	9	907
14		Lettuce-romaine-7wk	10	1060
15		Vinyard-untrained	72	7198
16		Vinyard-vertical-trellis	18	1789
Total	/	/	533	53596

TABLE IV
CATEGORY NAMES AND SAMPLE DIVISION OF THE HOUSTON 2013 DATASET

Category	Label color	Land cover	Training	Test
1		Healthy Grass	63	1188
2		Stressed Grass	63	1191
3		Synthetic Grass	35	662
4		Trees	62	1182
5		Soil	62	1180
6		Water	16	309
7		Residential	63	1205
8		Commerical	62	1182
9		Road	63	1189
10		Hightway	61	1166
11		Railway	62	1173
12		Parking Lot 1	62	1171
13		Parking Lot 2	23	446
14		Tennis Court	21	407
15		Running Track	33	627
Total	/	/	751	14278

TABLE V
IMPACT OF DIFFERENT MODULES ON OA

Case	Component				DataSets			
	Basic	AHIM	DAM	SLBM	Indian Pines	Pavia	Salinas	Houston 2013
1	√	–	–	–	96.69	94.08	95.93	96.52
2	√	√	–	–	97.51	95.57	98.32	97.16
3	√	–	√	–	97.96	95.03	97.66	96.93
4	√	–	–	√	97.73	96.58	98.40	97.53
5	√	√	√	–	98.31	97.11	98.86	97.46
6	√	√	√	√	**98.83**	**98.58**	**99.09**	**97.89**

and Kappa coefficient. Here, OA represents the ratio of the number of correctly classified samples to the total number of samples, AA represents the average classification accuracy of each category, and the Kappa coefficient is an evaluation indicator used to measure the robustness of the model.

2) Comparison Methods: In order to validate the effectiveness of the proposed method, a variety of state-of-the-art networks based on CNNs and transformers are chosen for comparison, including LS2CM-Res [44], HybridSN [36], PyResNet [25], MCRSCA [35], VIT [38], SSTN [45], SSFTT [40], SpectralFormer [39], MorphFormer [41], and CTMixer [42]. Among them, LS2CM-Res is a lightweight classification method based on CNN. HybrideSN is a hybrid 2DCNN and 3DCNN CNN network. PyResNet is a residual CNN classification network consisting of pyramidal bottleneck residual blocks and convolutional layers. Unlike the above methods, MCRSCA is a CNN classification method based on second-order channel attention. VIT is a classic transformer-based image classification method. SpectralFormer rethinks the problem of HSIC in terms of spectral sequence properties and designs a transformer-based SpectralFormer classification network. SSFTT. SSFTT, MorphFormer, and CTMixer are designed differently from pure transformer networks such as VIT and SpectralFormer. They are HSIC networks that combine the advantages of transformers and CNNs. SSTN is a spectral space transformer that determines the order of hierarchical and block-level selection in a network through the factorized architecture search (FAS) framework.

C. Model Analysis

1) Ablation Experiments: In the proposed DAHIT method, the network mainly consists of four parts, namely, Basic, AHIM, DAM, and SLBM. In order to better verify the impact of each part on the classification performance, some ablation experiments were conducted on four commonly used datasets and the experimental results are shown in Table V. Among them, "√" means that the module is contained in the network, and "-" means that the network does not contain the module. There are a total of six cases. As can be seen from Table V, the first case has the lowest OA in the four datasets. The second case is the addition of AHIM in the first basic network,

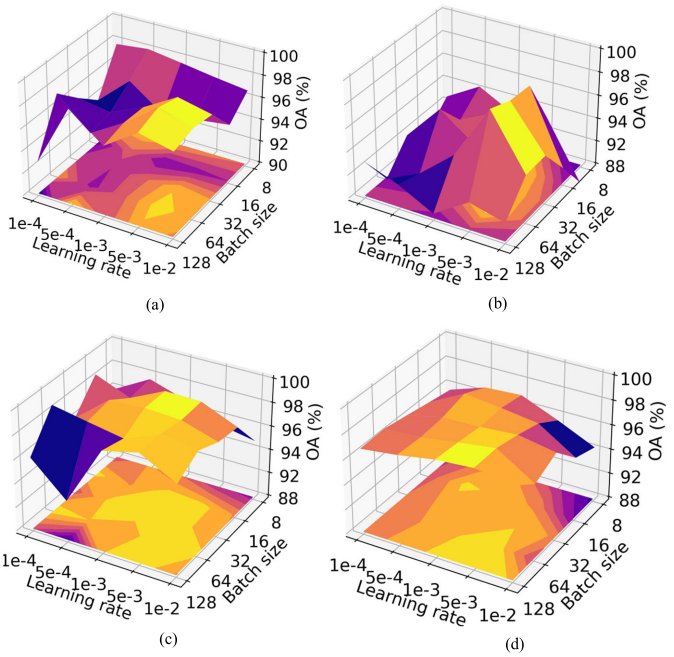

Fig. 6. Effect of different learning rates and batch sizes on OA. (a) Experimental results on Indian Pines dataset. (b) Experimental results on Pavia dataset. (c) Experimental results on Salinas dataset. (d) Experimental results on Houston 2013 dataset.

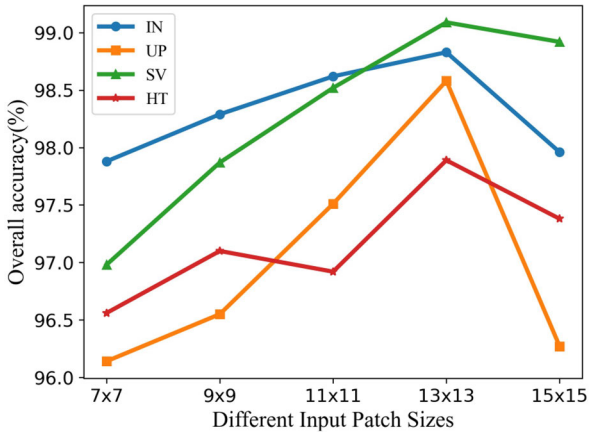

Fig. 7. Impact of different input sizes on classification performance.

where significant improvements can be observed across the four datasets, with the most significant improvement being achieved in Salinas. The third and fourth cases, respectively, add DAM and SLBM based on the first case. From Table V, it can be seen that compared with the first case, the third and fourth cases have improved on all four datasets. The fifth case, in which AHIM and DAM are added to the network, shows the greatest improvement on the Pavia dataset, with a 1.54% improvement in OA values compared to Case II with only AHIM. In the last case, when the network contains all components, it can be seen from Table V that the OA values have achieved the optimal results on all four datasets. The ablation experiments fully prove the effectiveness of the above components.

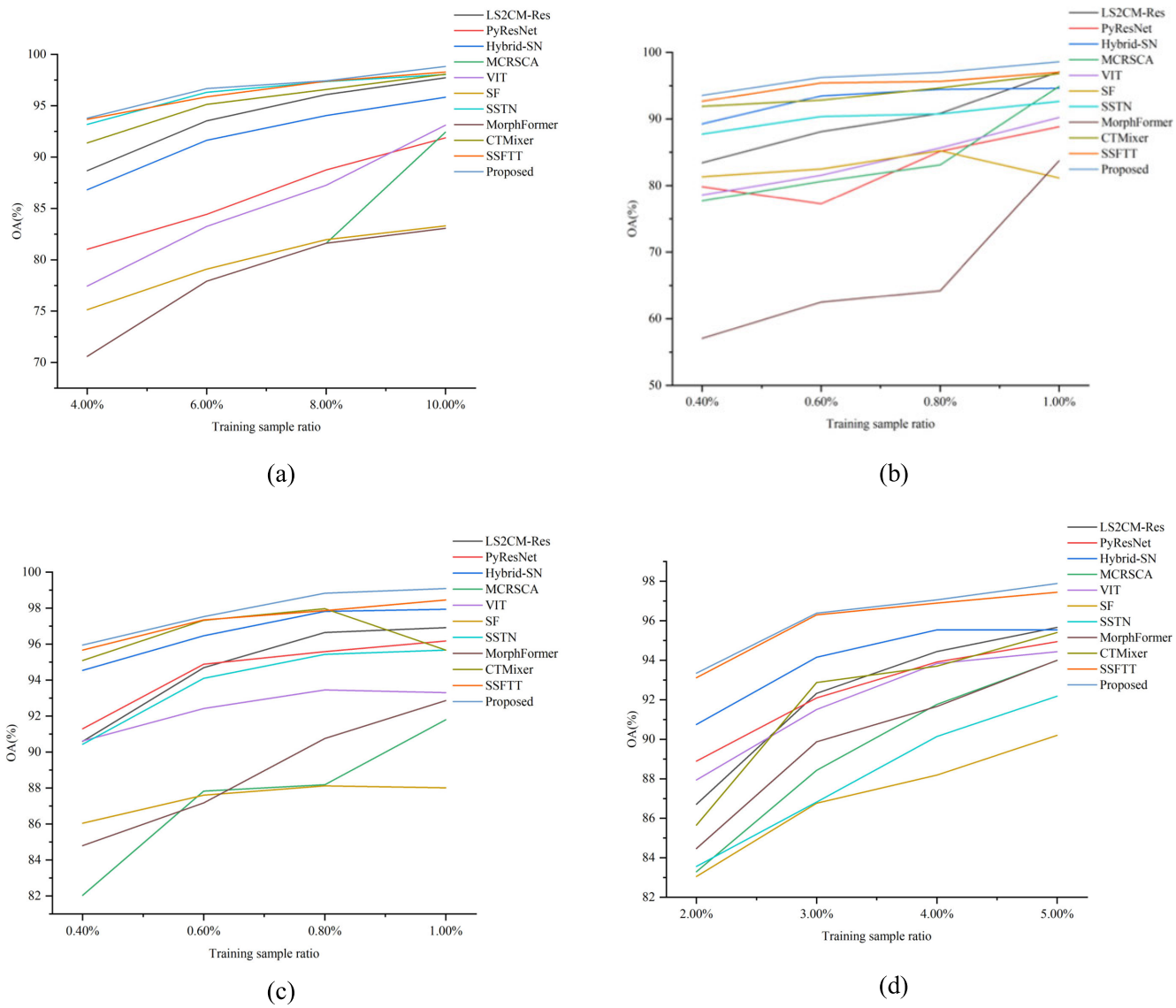

Fig. 8. OA values on different datasets under different training sample rates. (a)–(d) OA values on the Indian Pines, Pavia, Salinas, and Houston 2013 datasets under different training sample rates, respectively.

2) Parameter Analysis: During the model training process, the combination of different batch sizes and learning rates can have an impact on the classification performance of the model. In order to select a suitable learning rate and batch size for DAHIT, some experiments with different learning rates and batch sizes were conducted on three commonly used datasets. Among them, the learning rate is selected from $\{1e-4, 5e-4, 1e-3, 5e-3, 1e-2\}$ and the batch size is selected from the set $\{128, 64, 32, 16, 8\}$. The experimental results are shown in Fig. 6.

Fig. 6 shows the experimental results on IP, Pavia, Salinas, and Houston 2013 datasets. Different colors represent different ranges of OA values, with yellow indicating the highest OA value and blue indicating the lowest OA value. From Fig. 6, it can be observed that the model is sensitive to different learning rates and batch sizes on the same dataset.

In the IP dataset, we can see that the optimal batch size is 64. When the batch size is 64, the OA value increases as the learning rate increases and reaches a maximum at a learning rate of $5e-3$. In the Pavia dataset, the optimal learning rate and the batch size are $5e-3$ and 64, respectively. In the Salinas dataset, we can see that when the learning rate is $5e-3$, all batch sizes can achieve good results. In the Houston 2013 dataset, it can be seen that when the learning rate of the model is set to $5e-3$, the classification performance of the model is better than that of other learning rates, and when the learning rate is $5e-3$, it can be seen from Fig. 7(d) that as the batch size decreases, the OA value also shows a gradually decreasing trend. In the Houston 2013 dataset, when the number of batches was set to 64 and the learning rate was set to $5e-3$, the OA value of the model reaches its optimal value. Therefore, through experiments on selecting different learning rates and batch sizes for the model, it can be found that the optimal learning rate and batch size of the proposed DAHIT method are $5e-3$ and 64, respectively.

TABLE VI
CLASSIFICATION RESULTS ON THE INDIAN PINES DATASET (OPTIMAL CLASSIFICATION RESULTS ARE BOLDED)

Methods	CNNs				Transformers						
	LS2CM-Res	PyResNet	Hybrid-SN	MCRSCA	VIT	SF	SSTN	MorphFormer	CTMixer	SSFTT	Proposed
OA(100%)	97.73±0.45	91.87±3.00	95.83±1.12	92.43±1.54	93.10±1.28	83.31±1.37	98.04±0.39	83.06±1.53	98.07±1.14	98.27±0.27	**98.83±0.30**
AA(100%)	96.46±1.37	92.22±2.57	95.19±1.33	89.11±2.21	94.49±1.27	79.26±2.46	92.02±0.42	76.20±4.13	97.64±0.95	97.75±0.77	**97.33±0.88**
Kappa(100%)	97.41±0.51	90.72±3.40	95.24±1.28	91.35±1.77	92.12±1.47	78.82±1.58	97.76±0.44	80.56±1.83	97.81±1.30	98.03±0.31	**98.67±0.34**
1	94.35±0.67	96.02±3.60	99.16±1.84	76.76±7.13	95.45±3.89	62.24±34.09	80.00±40.00	57.50±3.33	98.73±2.24	99.66±1.00	**100.00±0.00**
2	98.01±1.06	87.69±4.82	91.86±4.16	91.7±3.39	91.03±2.85	77.66±3.50	98.01±0.78	74.20±3.10	98.52±1.32	98.20±1.21	98.07±0.65
3	97.41±1.28	91.87±2.45	94.00±2.79	89.83±4.57	91.84±2.22	76.88±2.91	97.09±0.65	75.04±5.90	**99.54±5.30**	97.32±1.07	98.67±0.90
4	98.46±1.82	94.76±6.02	95.70±2.67	79.26±6.58	94.87±3.91	67.68±6.18	99.64±0.91	54.79±8.59	99.37±0.97	97.56±2.09	**100.00±0.00**
5	97.02±0.84	89.75±18.79	99.11±0.70	94.77±8.21	94.62±1.57	87.65±3.28	97.32±1.63	68.72±3.85	97.66±1.83	98.90±1.05	**99.54±0.75**
6	98.29±1.17	96.05±1.92	98.49±0.99	99.16±2.76	98.43±1.33	90.11±4.28	98.37±0.56	94.50±1.25	98.99±0.73	99.48±0.46	**99.49±0.46**
7	85.70±1.63	89.85±1.77	**92.67±10.37**	80.45±11.87	92.58±4.81	79.95±9.66	88.44±6.04	40.91±3.40	93.83±12.52	92.57±13.03	82.76±10.58
8	99.75±1.00	98.28±2.20	98.66±1.74	99.79±2.05	97.87±1.33	89.41±1.64	96.40±2.52	97.38±0.21	**99.95±0.15**	99.52±0.56	99.31±0.79
9	97.06±7.89	86.85±14.65	85.06±17.58	69.38±17.94	**100.00±0.00**	66.95±30.68	30.00±40.00	61.54±2.12	96.64±3.39	97.51±5.79	94.44±7.70
10	95.60±1.71	93.68±3.05	93.56±2.65	88.1±2.90	91.73±1.46	78.30±3.46	97.67±1.02	84.65±3.28	94.99±6.87	97.01±1.21	**99.54±1.04**
11	98.58±0.67	92.52±3.80	97.15±0.97	93.45±3.06	91.23±2.68	80.53±2.32	98.31±0.54	88.03±6.21	98.78±0.92	98.61±0.69	**99.10±0.31**
12	95.50±3.22	87.62±5.53	95.13±2.84	79.83±3.14	89.31±4.03	71.34±7.67	96.90±1.17	72.45±3.91	96.15±4.25	97.34±1.61	**97.36±1.60**
13	99.45±1.27	94.94±10.80	99.45±1.09	99.15±2.47	99.09±1.32	91.24±6.45	99.88±0.73	97.59±2.03	99.34±0.99	99.47±1.58	**100.00±0.00**
14	99.29±0.38	97.2±2.48	98.11±1.43	97.82±1.92	96.02±0.64	91.07±1.78	98.79±0.69	95.63±4.03	99.22±0.59	99.20±0.77	**100.00±0.00**
15	95.16±2.18	93.63±3.55	95.67±3.54	89.45±5.73	92.98±2.72	76.82±5.59	97.41±1.69	65.35±7.46	95.60±5.37	97.22±2.36	**99.71±1.46**
16	93.74±4.32	84.79±12.87	89.26±6.51	**96.89±2.57**	94.82±4.57	91.17±4.13	97.47±1.60	90.91±7.27	95.00±4.16	94.44±3.27	89.13±2.73

TABLE VII
CLASSIFICATION RESULTS ON THE PAVIA DATASET (OPTIMAL CLASSIFICATION RESULTS ARE BOLDED)

Methods	CNNs				Transformers						
	LS2CM-Res	PyResNet	Hybrid-SN	MCRSCA	VIT	SF	SSTN	MorphFormer	CTMixer	SSFTT	Proposed
OA(100%)	97.08±0.34	88.82±1.41	94.62±2.33	94.91±162	90.21±1.05	81.16±2.22	92.62±1.60	83.73±0.98	96.82±3.50	97.03±0.71	**98.58±0.43**
AA(100%)	96.10±0.62	89.05±2.15	92.57±2.68	92.43±1.83	88.52±1.41	77.50±5.35	89.22±7.91	75.96±2.76	95.78±2.62	96.10±1.08	**97.56±0.38**
Kappa(100%)	96.12±0.45	84.98±1.97	92.63±3.16	93.23±2.11	86.89±1.43	74.33±3.22	90.11±2.18	77.99±1.86	95.79±4.76	96.05±0.94	**98.11±0.58**
1	96.17±0.99	83.75±6.56	93.64±3.50	94.67±2.37	87.50±2.61	82.84±4.74	84.29±4.46	85.39±2.30	97.10±1.12	96.30±1.97	**99.27±1.01**
2	98.70±0.89	93.39±3.86	97.91±2.17	98.90±1.56	93.47±1.51	83.20±3.05	97.67±0.95	97.49±3.92	99.20±3.45	98.98±0.36	**99.33±0.60**
3	90.99±4.49	76.76±11.98	82.36±6.92	79.93±4.69	73.05±6.24	47.41±14.73	75.39±29.03	39.30±4.38	84.33±15.50	93.44±3.35	**97.99±3.11**
4	**97.35±1.68**	95.49±3.23	96.95±1.61	95.78±0.98	94.98±2.68	85.35±4.94	95.54±0.90	86.69±10.12	96.90±1.73	96.69±1.35	96.02±1.09
5	98.18±3.80	95.58±6.15	97.11±2.56	**99.67±1.64**	98.13±4.05	95.74±2.09	97.16±1.09	93.93±4.49	98.45±2.22	98.12±2.64	98.73±2.00
6	98.99±0.49	93.66±6.34	97.47±1.51	90.24±6.65	89.31±4.03	77.38±8.52	99.37±0.48	61.50±3.27	98.56±0.51	98.17±0.99	**100.00±0.00**
7	94.48±4.86	90.61±11.32	90.84±8.60	81.86±8.22	84.62±5.33	57.58±25.80	78.34±44.64	56.34±11.58	95.81±3.99	97.25±2.43	**99.47±1.78**
8	92.04±2.56	77.01±12.44	81.26±3.59	91.01±5.06	82.04±3.63	69.12±8.06	75.47±7.44	67.27±7.34	**95.36±5.95**	88.73±3.27	94.38±2.84
9	97.97±3.79	95.19±5.66	95.58±4.08	99.80±3.78	93.61±7.19	98.98±0.58	**99.80±0.68**	95.69±17.95	96.39±1.64	97.25±1.24	96.86±1.48

3) Different Input Space Sizes: The HSIC method proposed in this article is based on the HSI 3-D cube method. The spatial size of the input cube has an impact on the classification accuracy of the network. In order to select the most appropriate spatial size of the input cube for the network, some experiments were conducted with different spatial sizes on four datasets, i.e., IP, Pavia, Salinas, and Houston 2013. The selected space sizes for the experiment are 7×7, 9×9, 11×11, 13×13, and 15×15, and the experimental results are shown in Fig. 7. From the Fig. 7, we can see that in the IP dataset, as the input space size increases, the overall classification accuracy shows a tendency of first increasing and then decreasing. It can be seen that the network has the best classification accuracy when the input spatial size is 13×13. In the Pavia dataset and Houston 2013 dataset, the classification accuracy is highest when the input spatial size is 13×13. It can be observed that the classification accuracy of the network at this point is more sensitive to changes in the input spatial, and the curve is steeper. This is because Pavia has many small targets, and when the input space size changes, it is easy to confuse samples of other categories in the input samples. From Fig. 7, we can see that

TABLE VIII
CLASSIFICATION RESULTS ON THE SALINAS DATASET (OPTIMAL CLASSIFICATION RESULTS ARE BOLDED)

Methods	CNNs				Transformers						
	LS2CM-Res	PyResNet	Hybrid-SN	MCRSCA	VIT	SF	SSTN	MorphFormer	CTMixer	SSFTT	Proposed
OA(100%)	96.92±0.50	96.18±0.69	97.94±0.45	91.80±1.05	93.3±0.49	88.01±1.18	95.67±0.82	92.86±0.68	95.67±3.59	98.45±2.78	**99.09±0.20**
AA(100%)	97.92±0.44	97.35±0.94	98.2±0.53	93.41±1.08	95.86±0.66	92.18±7.29	97.52±6.57	94.46±0.74	97.96±0.87	98.94±1.28	**99.15±0.23**
Kappa(100%)	96.57±0.55	95.75±0.77	97.71±0.50	90.86±1.17	92.54±0.54	86.63±1.32	95.17±1.21	92.05±0.76	95.18±4.03	98.28±0.31	**98.98±0.22**
1	**100.00±0.00**	99.14±2.03	99.24±2.02	92.98±0.32	97.47±4.32	96.59±1.55	**100.00±0.00**	97.82±1.48	**100.00±0.00**	99.96±0.15	**100.00±0.00**
2	99.76±0.66	99.81±0.22	99.56±0.62	95.86±2.02	98.54±1.57	93.2±1.73	99.66±6.57	97.76±1.80	98.43±4.43	99.97±0.07	**100.00±0.00**
3	97.22±1.04	99.86±0.21	99.94±0.14	90.24±5.76	96.16±2.97	90.55±2.33	97.02±1.21	93.32±2.23	99.73±0.37	99.85±0.29	**100.00±0.00**
4	96.05±5.98	95.12±6.74	97.4±1.66	98.91±1.74	97.85±1.36	96.74±0.84	97.54±2.41	93.96±2.48	97.08±3.03	98.44±1.21	**99.93±2.18**
5	**99.45±0.67**	98.82±1.13	97.6±1.92	95.98±1.20	97.14±0.98	96.31±2.26	98.45±0.38	94.91±1.82	99.08±0.93	99.20±0.81	98.40±0.42
6	99.95±1.30	99.7±0.34	99.68±0.73	99.91±0.28	98.84±0.87	99.29±0.51	99.84±0.04	97.90±1.00	99.75±0.48	99.97±0.06	**100.00±0.00**
7	99.4±0.62	99.71±0.40	99.56±0.48	99.78±1.98	99.03±1.74	96.53±1.56	99.65±0.47	98.64±0.72	**100.00±0.00**	99.88±0.26	**100.00±0.00**
8	96.88±2.70	95.88±3.86	96.93±0.88	87.38±3.56	86.66±1.29	76.09±3.59	91.49±4.03	88.30±2.84	95.48±10.78	97.22±1.12	**99.40±1.01**
9	99.52±0.20	99.81±0.26	99.81±0.18	99.27±0.58	99.07±0.45	98.21±0.88	99.29±0.26	97.95±0.96	99.76±0.20	99.81±0.24	**99.84±0.36**
10	96.9±2.98	98.67±1.05	98.38±1.13	91.80±3.51	95.23±3.44	86.11±3.16	97.67±1.28	92.36±3.84	97.40±1.97	**99.40±0.42**	98.82±1.42
11	96.54±4.41	**98.16±1.22**	96.14±2.86	76.70±5.43	93.24±2.66	85.90±1.88	96.25±4.15	96.81±1.80	97.76±2.38	97.33±2.22	95.75±0.79
12	99.26±0.28	97.83±2.32	99.56±0.64	98.98±11.43	98.65±2.53	95.56±0.80	98.43±1.58	92.27±1.99	**99.90±0.19**	99.45±0.53	99.27±0.70
13	98.83±1.34	94.58±6.54	96.91±3.40	97.39±3.25	98.18±3.06	94.29±2.13	**99.16±0.91**	93.67±0.72	97.57±4.03	98.84±1.24	97.7±0.72
14	98.88±1.12	96.1±6.68	96.4±2.78	97.45±3.60	98.52±2.47	92.57±4.07	**99.16±1.51**	94.30±2.84	99.02±1.83	98.87±1.02	98.47±2.18
15	88.35±3.45	85.61±4.62	94.43±2.33	77.06±5.51	81.45±2.47	66.38±4.19	86.79±3.02	84.40±2.67	86.66±11.10	95.00±1.84	**96.84±1.33**
16	99.80±1.02	98.72±1.91	99.61±0.59	94.82±1.97	97.78±1.33	97.25±0.92	99.95±0.02	96.92±1.46	99.75±0.73	99.87±023	**100.00±0.00**

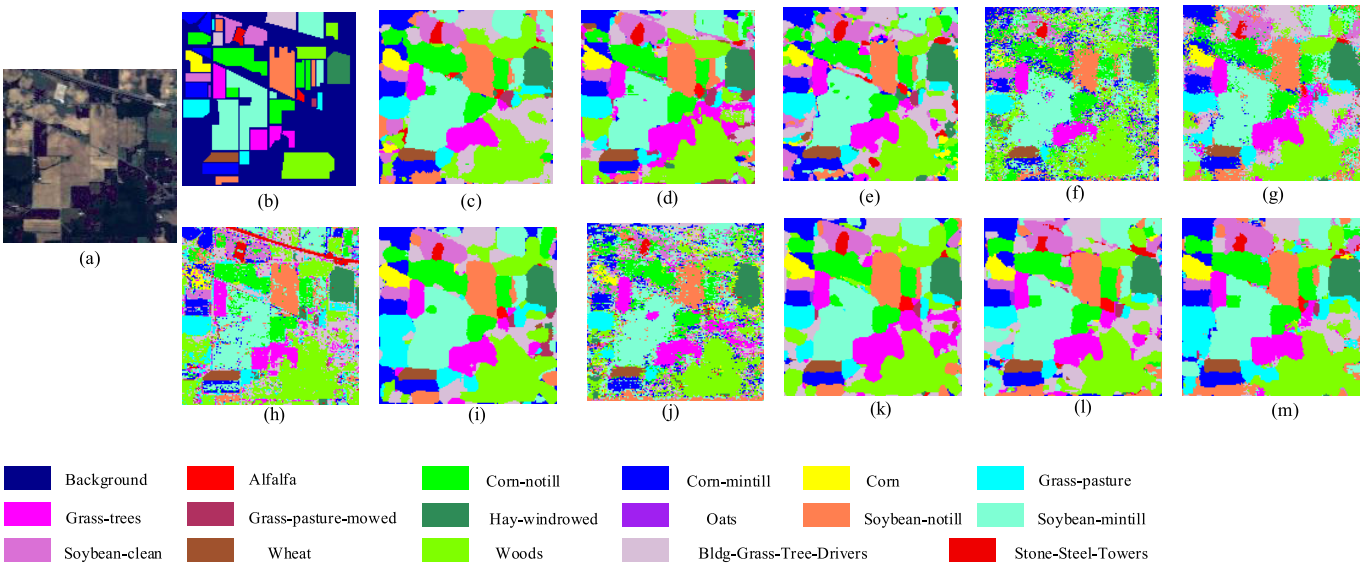

Fig. 9. Classification maps obtained by each classification method on the Indian Pines dataset. (a) Pseudo-color map; (b) real feature map; and (c)–(m) classification maps for LS2CM-Res (97.73%), PyResNet (91.87%), HybridSN (95.83%), MCRSCA (92.43), VIT (93.10%), SpectralFormer (81.52%), SSTN (98.04%), MorphFormer (83.06%), CTMixer (98.07%), SSFTT (98.27%), and DAHIT (98.83%), respectively.

in the Salinas dataset, as the input space size increases, the classification accuracy increases. At a spatial size of 13 × 13, the classification accuracy reaches its highest. When the spatial sizes are 13 × 13 and 15 × 15, the classification accuracy does not differ significantly. This is because the network proposed in this article is an HSIC method based on patch blocks. As the spatial size of the patch input into the network increases, the spatial information contained within

TABLE IX
CLASSIFICATION RESULTS ON THE HOUSTON 2013 DATASET (OPTIMAL CLASSIFICATION RESULTS ARE BOLDED)

Methods	CNNs				Transformers						
	LS2CM-Res	PyResNet	Hybrid-SN	MCRSCA	VIT	SF	SSTN	MorphFormer	CTMixer	SSFTT	Proposed
OA(100%)	95.66±1.64	94.94±1.79	95.54±1.63	93.99±0.54	94.43±0.51	90.30±0.72	92.18±0.90	94.00±0.54	95.40±2.61	97.44±0.52	**97.89±0.43**
AA(100%)	96.04±1.23	95.53±1.62	95.94±1.32	94.46±0.53	95.06±0.47	90.69±0.78	93.63±0.84	94.46±0.53	96.27±1.98	97.54±0.53	**97.82±0.43**
Kappa(100%)	95.31±1.77	94.53±1.93	93.50±0.59	85.51±2.04	93.97±0.55	89.51±0.78	91.54±0.97	93.51±0.59	95.03±2.82	97.23±0.56	**97.72±0.47**
1	92.54±3.19	96.53±2.65	96.34±2.53	88.00±5.16	96.29±2.21	95.09±2.71	89.00±4.06	96.34±2.54	92.08±5.30	98.00±1.09	**99.58±1.61**
2	98.20±2.57	97.11±3.87	97.78±1.03	97.06±1.19	97.85±0.76	97.27±1.64	94.18±3.57	97.78±1.03	95.07±8.35	**98.69±0.90**	98.65±0.60
3	**99.83±0.42**	98.59±1.89	98.46±1.08	98.80±2.01	99.79±0.38	99.30±0.54	98.46±1.88	98.46±1.09	99.12±2.34	99.54±0.54	99.69±1.39
4	94.81±3.56	95.79±1.87	94.40±1.26	96.01±1.44	96.83±0.85	97.97±1.91	98.67±1.61	94.40±1.26	96.89±5.36	97.03±1.60	**99.32±1.84**
5	99.64±0.48	98.45±2.65	98.55±2.53	94.04±2.73	99.24±0.52	96.53±1.77	99.25±0.92	98.55±2.53	99.30±0.93	99.46±0.60	**99.79±0.43**
6	97.73±3.76	96.91±2.80	98.07±1.72	97.51±2.37	98.84±0.94	97.77±2.98	97.89±3.89	98.07±1.72	**99.32±1.30**	98.00±3.28	95.46±3.28
7	94.91±1.98	95.64±1.74	82.38±3.20	79.24±3.33	93.45±2.98	86.91±3.88	91.44±4.93	82.38±3.29	94.55±5.84	**97.32±1.44**	96.09±1.76
8	98.40±1.52	92.14±7.90	97.53±1.16	87.02±4.04	91.17±2.32	84.39±2.93	95.24±2.50	97.53±1.16	98.09±3.07	**99.34±4.77**	97.12±0.69
9	93.66±3.83	94.24±3.98	86.39±2.99	79.45±4.34	90.84±2.44	82.53±4.26	93.51±2.73	86.39±2.99	94.62±2.97	**96.60±2.91**	91.09±2.61
10	91.47±5.68	88.56±2.45	93.91±1.86	79.88±4.72	90.62±1.76	86.56±3.06	76.75±5.04	93.91±1.86	89.84±7.44	94.80±1.12	**99.91±1.69**
11	96.41±3.21	96.23±0.89	94.58±1.72	74.47±3.02	93.60±1.75	84.83±2.26	93.69±5.59	94.58±1.72	97.03±5.76	98.17±1.57	**98.55±1.72**
12	94.54±3.72	93.98±2.51	95.52±1.68	80.24±5.62	89.30±3.38	83.18±2.81	87.36±4.31	95.52±1.68	95.39±2.87	94.68±4.21	**97.95±1.94**
13	92.92±3.44	94.98±3.37	88.28±3.22	90.28±2.98	93.07±3.55	74.34±5.22	92.91±3.61	88.28±3.22	**95.83±3.47**	93.73±4.76	93.95±4.83
14	97.36±2.88	96.67±3.88	97.48±2.26	91.32±4.05	99.26±1.74	95.82±1.17	98.48±2.35	97.48±2.26	99.39±1.74	99.05±1.96	**100.00±0.00**
15	98.18±1.53	94.24±11.99	97.17±1.94	97.41±1.24	95.75±4.45	97.79±1.32	97.55±1.83	97.17±1.94	97.53±1.44	98.65±1.50	**100.00±0.00**

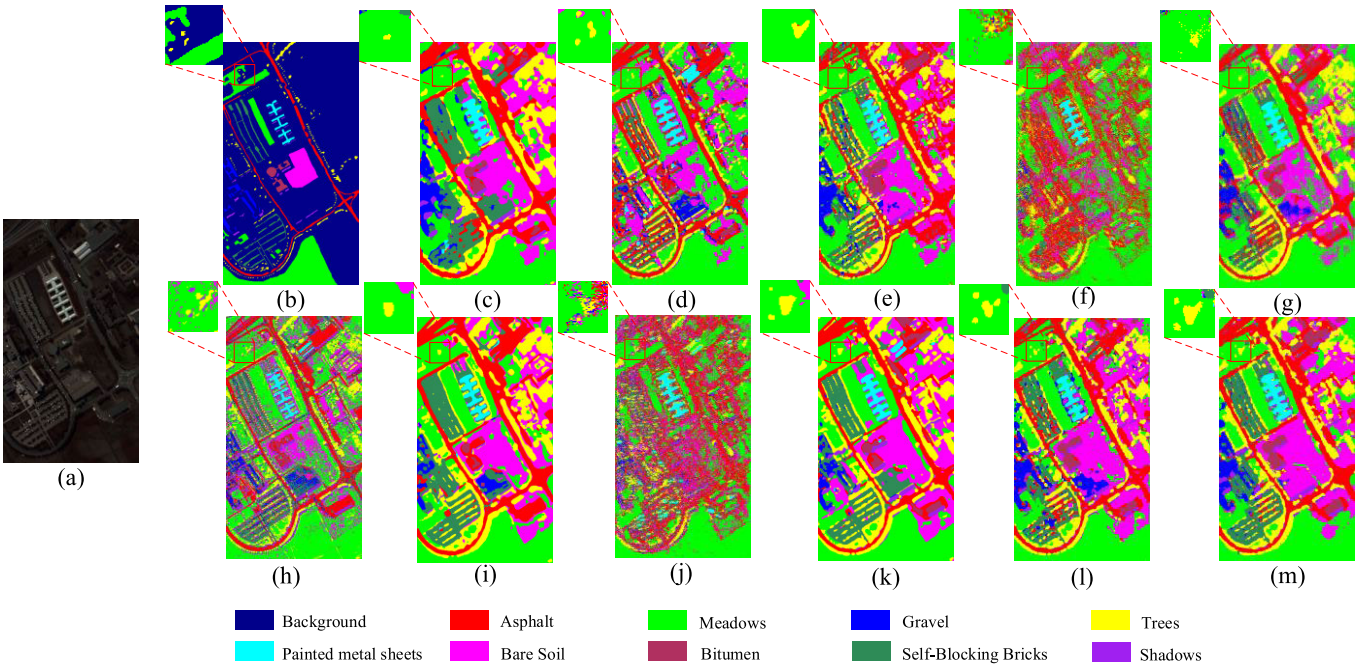

Fig. 10. Classification maps obtained by each classification method on the Pavia dataset. (a) Pseudo-color map; (b) real feature map; and (c)–(m) classification maps for LS2CM-Res (97.08%), PyResNet (88.82%), HybridSN (94.46%), MCRSCA (94.91%), VIT (90.21%), SpectralFormer (81.16%), SSTN (92.62%), MorphFormer (83.73%), CTMixer (96.82%), SSFTT (97.03%), and DAHIT (98.58%), respectively.

the patch becomes richer. Therefore, when the spatial size is increased from 7 × 7 to 13 × 13, the patch contains more spatial information that is beneficial for classification. There are many small targets in the IP and Pavia datasets. Therefore, when the space size is increased to 15 × 15, there will be interference information in the input patch, and the network

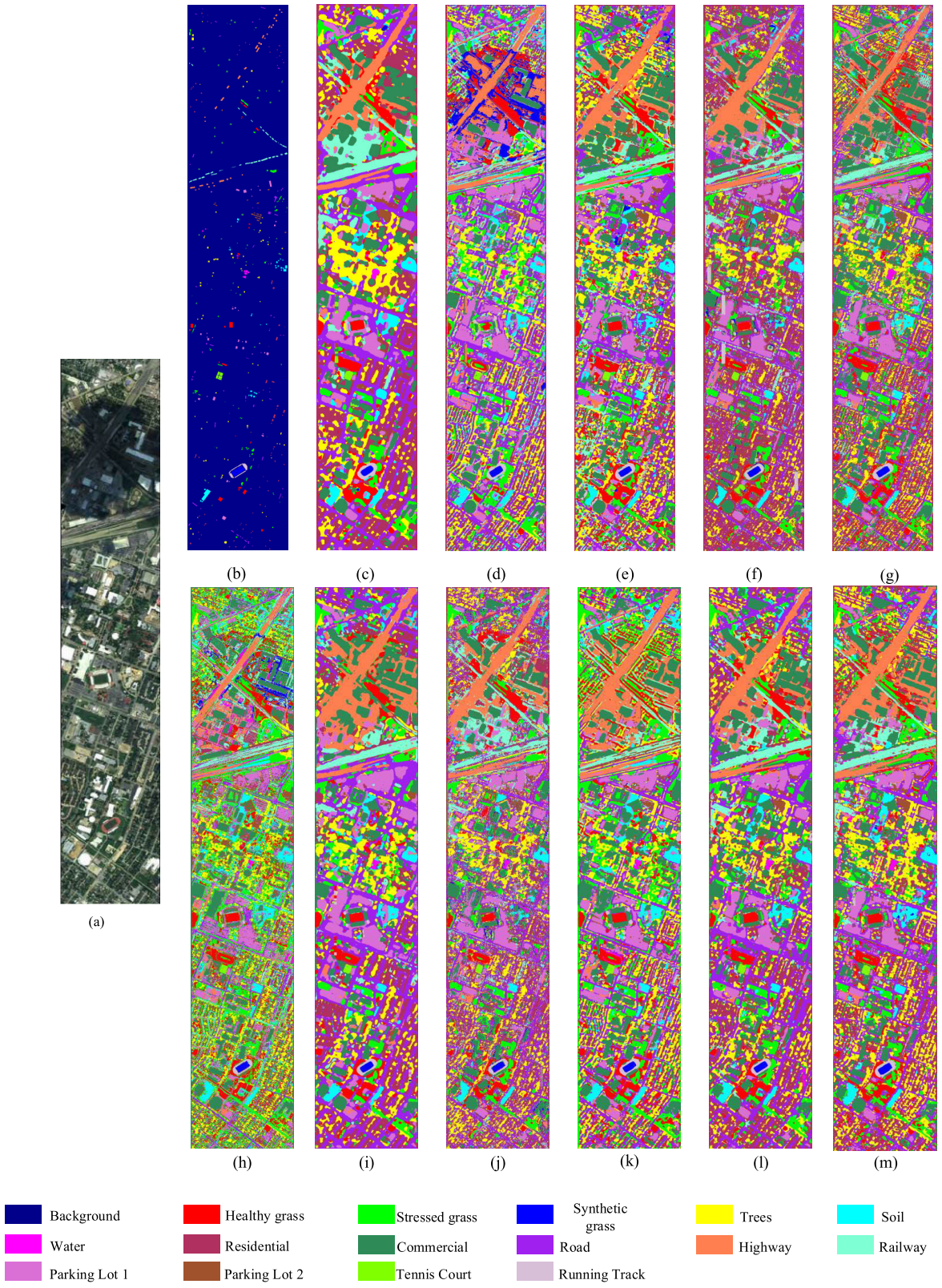

Fig. 11. Classification maps obtained by each classification method on the Houston 2013 dataset. (a) Pseudo-color map; (b) real feature map; and (c)–(m) classification maps for LS2CM-Res (95.66%), PyResNet (94.94%), HybridSN (95.54%), MCRSCA (93.99%), VIT (94.43%), SpectralFormer (90.30%), SSTN (92.18%), MorphFormer (94.00%), CTMixer (95.40%), SSFTT (97.44%), and DAHIT (97.89%), respectively.

classification accuracy will be reduced. From Fig. 7, it can be seen that in the Salinas dataset, due to the small number of small-sized ground cover in Salinas, the fluctuation of the OA curve of Salinas is small when the spatial size is increased from 13 × 13 to 15 × 15. In summary, the most suitable input space size for the proposed method is 13 × 13. In this

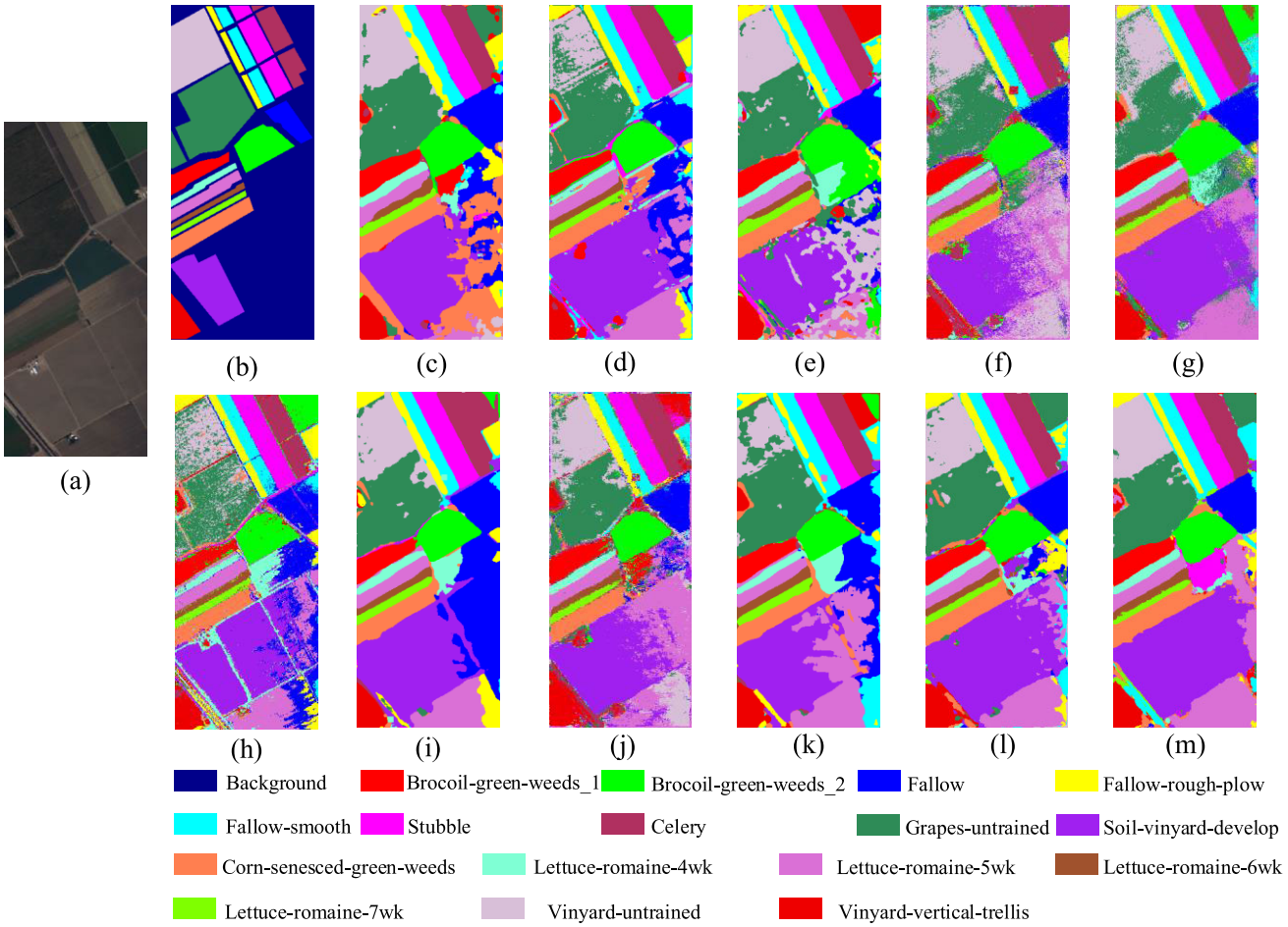

Fig. 12. Classification maps obtained by each classification method on the Salinas dataset. (a) Pseudo-color map; (b) real feature map; and (c)–(m) classification maps for LS2CM-Res (96.92%), PyResNet (96.18%), HybridSN (97.94%), MCRSCA (91.80%), VIT (93.30%), SpectralFormer (88.01%), SSTN (95.67%), MorphFormer (86.66%), CTMixer (95.67%), SSFTT (98.45%), and DAHIT (99.09%), respectively.

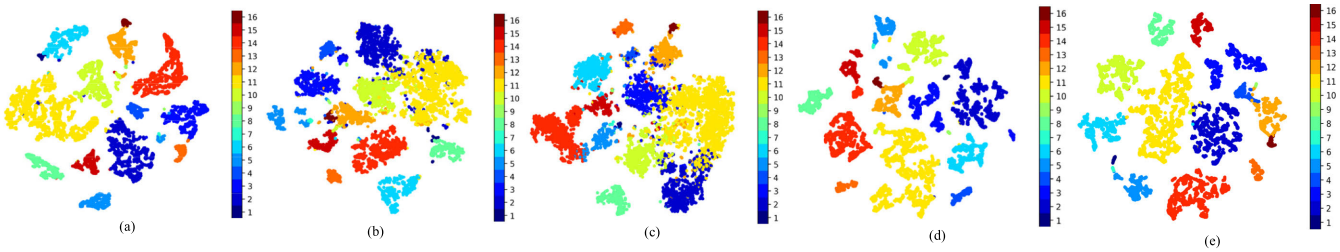

Fig. 13. t-SNE visualization on the Indian Pines dataset. (a) LS2CM-Res (97.73%). (b) VIT (93.10%). (c) MorphFormer (83.06%). (d) CTMixer (98.07%). (e) DAHIT (98.83%).

case, the input samples of the network contain more spatial information, which is beneficial for the network to extract more discriminative classification features.

4) Different Sample Proportions: In order to better prove the performance of the proposed model, this article compares all methods under different training samples. The sample proportion set selected in the IP dataset is {4.0%, 6.0%, 8.0%, 10.0%}, the sample proportion set selected in the Pavia and Salinas datasets is {0.4%, 0.6%, 0.8%, 1.0%}, and the sample proportion set selected in the Houston 2013 dataset is {2.0%, 3.0%, 4.0%, 5.0%}. The OA values of the four datasets at different training sample ratios are shown in Fig. 8. From Fig. 8, it can be seen that the proposed DAHIT not only achieves satisfactory classification results with a larger training sample ratio but also has higher OA values than those of other comparison methods with small training sample ratio. Specifically, it can be clearly observed in the Pavia, Salinas, and Houston 2013 datasets that the training accuracy of the proposed DAHIT method is superior to that of other methods in any training sample ratio and has significant advantages. Based on the above analysis, the proposed DAHIT method has achieved satisfactory classification performance on four different datasets. The OA values are still relatively high with a small proportion of training samples. This indicates that the proposed method can not only provide good classification performance but also has excellent robustness.

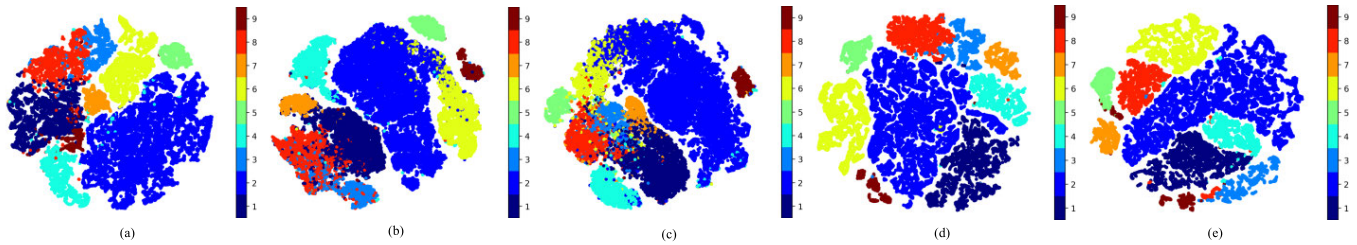

Fig. 14. t-SNE visualization on the Pavia dataset. (a) LS2CM-Res (97.08%). (b) VIT (90.21%). (c) MorphFormer (83.73%). (d) CTMixer (96.82%). (e) DAHIT (98.58%).

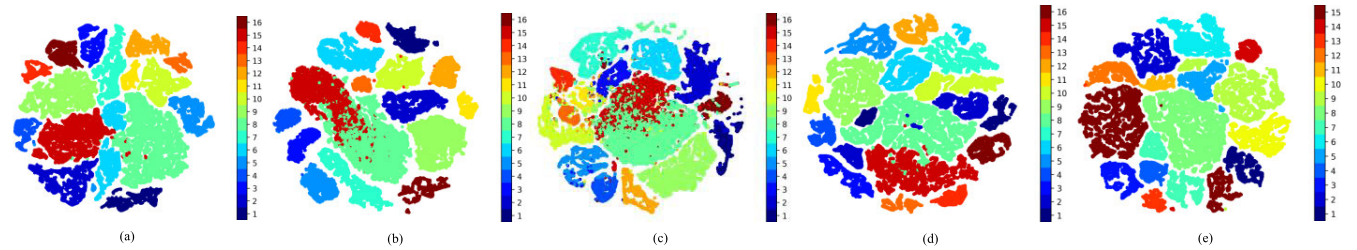

Fig. 15. t-SNE visualization on the Salinas dataset. (a) LS2CM-Res (96.92%). (b) VIT (93.30%). (c) MorphFormer (86.66%). (d) CTMixer (95.67%). (e) DAHIT (99.09%).

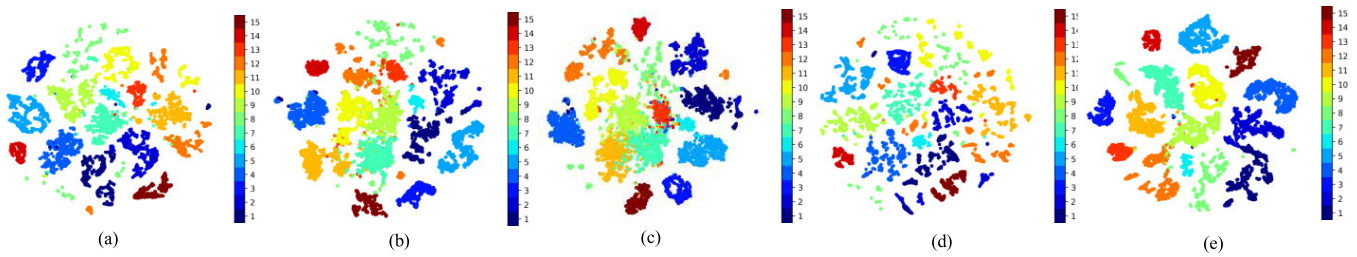

Fig. 16. t-SNE visualization on the Houston 2013 dataset. (a) LS2CM-Res (95.66%). (b) VIT (94.43%). (c) MorphFormer (94.00%). (d) CTMixer (95.40%). (e) DAHIT (97.89%).

D. Analysis of Experimental Results

1) Quantitative Analysis: Tables VI–IX show the OA, AA, Kappa, and the classification accuracy of each category for all methods on the four datasets of IP, Pavia, and Salinas, with the best results highlighted in bolded. From Tables V to VII, it can be seen that the CNN-based method benefited from its powerful local feature extraction ability and achieved good classification results in all four datasets. However, due to the poor performance of CNN in global feature extraction, CNN-based methods are prone to performance bottlenecks. In addition, transformer-based methods can extract global features of images, but the classification results obtained solely using transformer methods, such as VIT, SF, and MorphFormer, are not satisfactory. The classification networks constructed by combining CNN and transformer, such as CTMixer and SSTN, have achieved good classification results.

From Tables VI to IX, it can be observed that the proposed classification method outperforms all comparison methods in the four datasets. This is because the proposed DAHIT method in this article solves the problem of unsatisfactory performance of transformer in extracting local features. By introducing spatial local bias and spectral branch attention diagonal mask, the network's local feature extraction ability for HSI is enhanced. We have designed an AHIM to further enhance the feature representation ability of MHSA. On the IP, Pavia, and Salinas datasets, the OA values of the proposed DAHIT method were 0.76%, 1.76%, and 3.42% higher than those of the transformer method with better classification performance, respectively. In transformer-based classification methods, SF achieved poor classification results. This is because SF only considered the spectral properties of HSIs, and the rich spatial information of HSIs was ignored by SF. SF was unable to pay attention to the spatial features of HSIs, resulting in poor classification results on the IP, Salinas, and Pavia datasets. Similarly, compared to CNN-based methods, the OA values of the proposed DAHIT method are 1.1%, 1.5%, and 1.15% higher. In the IP dataset, the proposed network achieves 100% classification accuracy in categories 1, 4, 13, and 14. In the Pavia dataset, there are six categories that have the best results among all comparison methods. In the Salinas dataset, there are six categories with a classification accuracy of 100%, namely, Categories 1, 2, 3, 6, 7, and 16. From Table IX, it can be observed that the proposed DAHIT achieved the best OA values for a total of eight categories in the Houston 2013 dataset and achieved 100% classification results in the tennis course and running track categories.

TABLE X
TRAINING TIME (S) AND TESTING TIME (S) OF EACH METHOD ON THE FOUR DATASETS

Methods		Indian Pines		Pavia		Salinas		Houston 2013	
		training	test	training	test	training	test	training	test
CNNs	LS2CM-Res	**0.30**	0.81	**0.12**	2.96	0.18	7.27	0.17	0.90
	PyResNet	2.31	7.00	0.97	36.30	1.22	45.51	1.73	11.50
	Hybrid-SN	0.34	0.95	0.14	4.91	0.18	6.09	0.25	1.55
	MCRSCA	0.52	1.63	0.21	7.53	0.30	10.39	0.38	2.39
Transformer	VIT	**0.30**	1.00	0.16	3.71	0.16	6.29	0.09	0.60
	SF	0.42	1.37	0.18	5.81	0.23	9.00	0.28	1.90
	SSTN	0.39	1.18	0.16	5.34	0.23	7.71	0.28	1.72
	MorphFormer	0.33	1.77	0.20	3.22	0.26	6.89	0.35	1.62
	CTMixer	0.41	1.37	0.15	5.10	0.23	8.29	0.27	1.74
	SSFTT	0.32	**0.47**	**0.12**	2.37	0.19	6.11	**0.12**	0.75
	Proposed	0.33	0.80	**0.12**	**2.19**	**0.17**	6.07	0.23	**0.65**

TABLE XI
PARAMS AND FLOPS OF EACH METHOD ON THE FOUR DATASETS

Methods		Indian Pines		Pavia		Salinas		Houston 2013	
		Params	FLOPs	Params	FLOPs	Params	FLOPs	Params	FLOPs
CNNs	LS2CM-Res	**9.86K**	0.16G	**9.40K**	0.16G	**9.86K**	0.16G	**9.79K**	0.16G
	PyResNet	21.86M	2.12G	21.85M	2.12G	21.86M	2.12G	21.85M	2.12G
	Hybrid-SN	796.80K	2.03G	795.90K	2.03G	796.80K	2.03G	796.67K	2.03G
	MCRSCA	990.50K	3.73G	989.60K	3.73G	990.50K	3.73G	136.99K	3.73G
Transformer	VIT	137.06K	1.08G	136.60K	1.08G	137.06K	1.08G	216.94K	1.08G
	SF	333.43K	2.15G	155.18K	0.86G	343.19K	2.12G	17.36K	1.35G
	SSTN	20.49K	**0.11G**	13.62K	**0.07G**	20.69K	**0.11G**	63.34K	**0.09G**
	MorphFormer	63.30K	0.33G	63.15K	0.33G	63.60K	0.33G	623.58K	0.33G
	CTMixer	625.74K	4.74G	621.18K	4.74G	625.74K	4.74G	623.58K	4.74G
	SSFTT	148.49K	0.45G	148.49K	0.45G	148.49K	0.45G	148.49K	0.45G
	Proposed	103.86K	0.50G	103.41K	0.50G	103.86K	0.50G	103.80K	0.50G

2) Visual Assessment: In order to analyze the classification performance of the model more intuitively, we visualize the classification results of each method. Figs. 9–12 show the classification results of each classification method on IP, Pavia, Houston 2013, and Salinas datasets, respectively. From Figs. 9 to 12, it can be seen that the visual effect of the proposed DAHIT method is the closest to the ground maps of all four datasets. As shown in Fig. 9, the CNN-based classification method on the IP dataset is the worst for those edge categories with pretzel noise. The classification method based solely on transformer cannot extract local features well, so its visualization results are also unsatisfactory. The methods combining CNN and transformer, such as SSTN and CTMixer, have relatively smooth classification results and almost no salt-and-pepper noise. For the visualization results of the Pavia dataset, compared to the CNN method lacking global feature extraction and the transformer method with weaker local feature extraction, the method combining CNN and transformer has the least salt-and-pepper noise in the classification visualization of the bare soil category, which is closest to the classification effect of the real bare soil category. Compared with other classification methods on the Salinas dataset, the proposed DAHIT method has the closest visualization results to the real distribution for the Grapes-untrained category.

Figs. 13–16 show the t-SNE visualization results of the five methods on IP, Pavia, Salinas, and Houston2013 datasets, respectively. It can be observed that the proposed DAHIT method achieves satisfactory clustering results. For the IP dataset, compared with the other four transformer's methods, the proposed DAHIT method has smaller distances within classes and larger distances among classes. Compared with the pure transformer approach, the method of combining CNN with transformer benefits from the rich global and local features and can provide the best t-SNE visualization results. As shown in Fig. 14, for the Pavia dataset, there is less category-to-category confusion in our proposed method. Compared with other methods, the proposed DAHIT can better cluster Categories 6 and 7, not only without category confusion

but also with smaller intraclass distances and larger interclass distances. For the Salinas dataset, the t-SNE visualization result of our proposed method is also the best. As shown in Fig. 16, compared to the other four methods, the proposed method has the best clustering performance on the Houston 2013 dataset. Different categories were clearly separated, and the geometric distance between similar categories was small.

3) Model Complexity Analysis: DAHIT achieved satisfactory classification results. In order to evaluate the model complexity of DAHIT, a time complexity analysis was conducted on the IP, Pavia, Salinas, and Houston 2013 datasets from the perspectives of training and testing time. Table X shows the time consumption of DAHIT on four datasets. As shown in Table X, the training time of the proposed method on the Pavia dataset and Salinas dataset is the shortest compared to other methods. On the Pavia and Houston 2013 datasets, DAHIT outperformed other methods in terms of testing time. On the IP dataset, the training time of DAHIT is slightly longer than that of the LS2CM Res and VIT methods, and the testing time of it is also slightly longer than that of the SSFTT method. In summary, the method proposed in this article not only has satisfactory classification performance and good robustness but also has relatively low model complexity.

In order to further analyze the complexity of the model, the floating-point operations (FLOPs) and parameter quantity of all methods were provided, as shown in Table XI. From Table XI, it can be observed that compared to transformer-based methods, most CNN-based methods have higher FLOPs. This is because CNN-based models require a large number of convolutional layers to extract rich local features of images, which leads to larger FLOPs. Transformer-based methods can extract global semantic information of images through MHSA and feedforward networks (FFNs). Compared with convolution-based methods, the FLOPs and parameter quantity of the proposed method are lower than those of all convolution-based methods except for LS2CM-Res. This is because LS2CM-Res is specially designed from a lightweight network perspective and has lower model complexity. Compared with transformer-based methods, the proposed DAHIT method outperforms more than half of the transformer methods in terms of FLOPs and Params. Specifically, compared to CTMixer that uses convolution to extract local features from images, the method proposed in this article exhibits significant advantages in both FLOPs and parameter quantity.

IV. Conclusion

For the classification task of HSI, not only global features but also local features are particularly important. Transformer can extract global features of images, but its performance in local feature extraction is not very good. This article proposes a DAHIT method for HSIC. First, an SLBM module was designed in the spatial branch for extracting spatial local features of HSIs. To enhance the representation ability of MHSA, an AHIM module was constructed for information interaction between different attention heads. Then, in order to extract the local features of the spectrum and enhance the attention of the attention module to the correlation between different spectral bands, a DAM module was carefully designed in the spectral branch. Finally, the features extracted from the two branches are cascaded and fused. Extensive experiments have shown that our proposed DAHIT achieves excellent classification performance for HSIs compared to other state-of-the-art classification methods. In the future, we will further optimize transformer's ability to extract local features of HSIs, introducing the local bias of convolution into the transformer network, and strive to achieve more excellent HSIC performance.

References

[1] X. Yang and Y. Yu, "Estimating soil salinity under various moisture conditions: An experimental study," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2525–2533, May 2017, doi: 10.1109/tgrs.2016.2646420.

[2] S. S. M. Noor et al., "The properties of the cornea based on hyperspectral imaging: Optical biomedical engineering perspective," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Bratislava, Slovakia, 2016, pp. 1–4, doi: 10.1109/IWSSIP.2016.7502710.

[3] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral–spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652–675, Mar. 2013, doi: 10.1109/JPROC.2012.2197589.

[4] J. Ling, Z. Zeng, Q. Shi, J. Li, and B. Zhang, "Estimating winter wheat LAI using hyperspectral UAV data and an iterative hybrid method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8782–8794, 2023, doi: 10.1109/JSTARS.2023.3317499.

[5] L. Sun, F. Wu, T. Zhan, W. Liu, J. Wang, and B. Jeon, "Weighted nonlocal low-rank tensor decomposition method for sparse unmixing of hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1174–1188, 2020, doi: 10.1109/JSTARS.2020.2980576.

[6] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1575–1585, Jun. 2006, doi: 10.1109/tgrs.2006.864389.

[7] A. MartÌnez-UsÓMartinez-Uso, F. Pla, J. M. Sotoca, and P. GarcÌa-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 4158–4171, Dec. 2007, doi: 10.1109/tgrs.2007.904951.

[8] S. Prasad and L. M. Bruce, "Limitations of principal components analysis for hyperspectral target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 625–629, Oct. 2008, doi: 10.1109/LGRS.2008.2001282.

[9] M. Wang, J. Yu, L. Niu, and W. Sun, "Unsupervised feature extraction for hyperspectral images using combined low rank representation and locally linear embedding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, 2017, pp. 1428–1431, doi: 10.1109/ICASSP.2017.7952392.

[10] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016, doi: 10.1109/TGRS.2016.2584107.

[11] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017, doi: 10.1109/TGRS.2016.2636241.

[12] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016, doi: 10.1109/TGRS.2016.2543748.

[13] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021, doi: 10.1109/TGRS.2020.3015157.

[14] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 241–245, Feb. 2019, doi: 10.1109/LGRS.2018.2869563.

[15] Q. Liu, L. Xiao, J. Yang, and Z. Wei, "CNN-enhanced graph convolutional network with pixel- and superpixel-level feature fusion for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8657–8671, Oct. 2021, doi: 10.1109/TGRS.2020.3037361.

[16] X. He, Y. Chen, and P. Ghamisi, "Dual graph convolutional network for hyperspectral image classification with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5502418, doi: 10.1109/TGRS.2021.3061088.

[17] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 1559–1572, 2022, doi: 10.1109/TIP.2022.3144017.

[18] J. Bai, B. Ding, Z. Xiao, L. Jiao, H. Chen, and A. C. Regan, "Hyperspectral image classification based on deep attention graph convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5504316, doi: 10.1109/TGRS.2021.3066485.

[19] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 212–216, Feb. 2018, doi: 10.1109/LGRS.2017.2780890.

[20] J. Wang, S. Guo, R. Huang, L. Li, X. Zhang, and L. Jiao, "Dual-channel capsule generation adversarial network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5501016, doi: 10.1109/TGRS.2020.3044312.

[21] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020, doi: 10.1109/TGRS.2019.2934760.

[22] X. Yang, W. Cao, Y. Lu, and Y. Zhou, "Hyperspectral image transformer classification networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528715, doi: 10.1109/TGRS.2022.3171551.

[23] A. Santara et al., "BASS net: Band-adaptive spectral–spatial feature learning neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5293–5301, Sep. 2017, doi: 10.1109/TGRS.2017.2705073.

[24] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018, doi: 10.1109/TGRS.2018.2794326.

[25] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019, doi: 10.1109/TGRS.2018.2860125.

[26] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017, doi: 10.1109/TIP.2017.2725580.

[27] 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, doi: 10.1109/TGRS.2022.3188529.

[28] L. Fang, D. Zhu, J. Yue, B. Zhang, and M. He, "Geometric-spectral reconstruction learning for multi-source open-set classification with hyperspectral and LiDAR data," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 10, pp. 1892–1895, Oct. 2022, doi: 10.1109/JAS.2022.105893.

[29] X. Ma, S. Ji, J. Wang, J. Geng, and H. Wang, "Hyperspectral image classification based on two-phase relation learning network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10398–10409, Dec. 2019, doi: 10.1109/TGRS.2019.2934218.

[30] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021, doi: 10.1109/TGRS.2020.3007921.

[31] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, Jan. 2020, doi: 10.1109/TGRS.2019.2933609.

[32] M. He, B. Li, and H. Chen, "Multi-scale 3D deep convolutional neural network for hyperspectral image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3904–3908, doi: 10.1109/ICIP.2017.8297014.

[33] H. Zhou, X. Zhang, C. Zhang, and Q. Ma, "Quaternion convolutional neural networks for hyperspectral image classification," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106234, doi: 10.1016/j.engappai.2023.106234.

[34] C. Shi, D. Liao, T. Zhang, and L. Wang, "Hyperspectral image classification based on expansion convolution network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5528316, doi: 10.1109/TGRS.2022.3174015.

[35] R. Shang, H. Chang, W. Zhang, J. Feng, Y. Li, and L. Jiao, "Hyperspectral image classification based on multiscale cross-branch response and second-order channel attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532016, doi: 10.1109/TGRS.2022.3184117.

[36] J. Zhu, L. Fang, and P. Ghamisi, "Deformable convolutional neural networks for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 8, pp. 1254–1258, Aug. 2018, doi: 10.1109/LGRS.2018.2830403.

[37] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020, doi: 10.1109/LGRS.2019.2918719.

[38] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[39] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615, doi: 10.1109/TGRS.2021.3130716.

[40] 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, doi: 10.1109/TGRS.2022.3207933.

[41] S. K. Roy, A. Deria, C. Shah, J. M. Haut, Q. Du, and A. Plaza, "Spectral–spatial morphological attention transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5503615, doi: 10.1109/TGRS.2023.3242346.

[42] J. Zhang, Z. Meng, F. Zhao, H. Liu, and Z. Chang, "Convolution transformer mixer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3208935.

[43] D. Liao, C. Shi, and L. Wang, "A complementary integrated transformer network for hyperspectral image classification," *CAAI Trans. Intell. Technol.*, vol. 8, no. 4, pp. 1288–1307, Jan. 2023, doi: 10.1049/cit2.12150.

[44] Z. Meng, L. Jiao, M. Liang, and F. Zhao, "A lightweight spectral–spatial convolution module for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3069202.

[45] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514715, doi: 10.1109/TGRS.2021.3115699.

[46] L. Yu, J. Peng, N. Chen, W. Sun, and Q. Du, "Two-branch deeper graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5506514, doi: 10.1109/TGRS.2023.3257369.

[47] Y. Huang et al., "Two-branch attention adversarial domain adaptation network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540813, doi: 10.1109/TGRS.2022.3215677.

[48] E. Xie et al., "Semantic and spatial–spectral feature fusion transformer network for the classification of hyperspectral image," *CAAI Trans. Intell. Technol.*, vol. 8, no. 4, pp. 1308–1322, 2023, doi: 10.1049/cit2.12201.

[49] H. Gao, Y. Zhang, Z. Chen, S. Xu, D. Hong, and B. Zhang, "A multidepth and multibranch network for hyperspectral target detection based on band selection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5506818, doi: 10.1109/TGRS.2023.3258061.

[50] H. Gao, Y. Zhang, Z. Chen, F. Xu, D. Hong, and B. Zhang, "Hyperspectral target detection via spectral aggregation and separation network with target band random mask," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515516, doi: 10.1109/TGRS.2023.3288739.

[51] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1498–1512, 2023, doi: 10.1109/TIP.2023.3243853.

[52] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and LiDAR data classification based on structural optimization transmission," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3153–3164, May 2023, doi: 10.1109/TCYB.2022.3169773.

[53] A. Qin, C. Liu, Z. Shang, and J. Tian, "Spectral–spatial graph convolutional networks for semel-supervised hyperspectral image classification," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit. (ICWAPR)*, Chengdu, China, 2018, pp. 89–94, doi: 10.1109/ICWAPR.2018.8521407.

[54] Q. Liu, Y. Dong, Y. Zhang, and H. Luo, "A fast dynamic graph convolutional network and CNN parallel network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5530215, doi: 10.1109/TGRS.2022.3179419.
[55] Q. Liu et al., "Refined prototypical contrastive learning for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5506214, doi: 10.1109/TGRS.2023.3257341.
[56] D. Liao, C. Shi, and L. Wang, "A spectral–spatial fusion transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5515216, doi: 10.1109/TGRS.2023.3286950.
[57] W. Liu et al., "Self-supervised feature learning based on spectral masking for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4407715, doi: 10.1109/TGRS.2023.3310489.
[58] B. Yang, F. Cao, and H. Ye, "A novel method for hyperspectral image classification: Deep network with adaptive graph structure integration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5523512, doi: 10.1109/TGRS.2022.3150349.

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

Her doctoral dissertation won the Nomination Award of Excellent Doctoral Dissertation of Harbin University of Technology (HIT) in 2016. From 2017 to 2020, she held a post-doctoral research position at the College of Information and Communications Engineering, Harbin Engineering University, Harbin. She is currently a Professor with the Department of communication engineering, Qiqihar University, Qiqihar, China. Since 2024, she has been working with the College of Information Engineering, Huzhou University, Huzhou, China. She has published two academic books about remote sensing image processing and more than 90 papers in journals and conference proceedings. Her main research interests include remote sensing image processing, pattern recognition, and machine learning.

Shuheng Yue received the bachelor's degree from Shandong Jiaotong University, Jinan, China, in 2021. He is currently pursuing the master's degree with Qiqihar University, Qiqihar, China.

His research interests include hyperspectral image processing and machine learning.

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

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

文献检索报告

哈尔滨工业大学图书馆 **HITLIB**

报告编号：**HIT202404901**

本次查收查引工作是根据委托人提供的作者姓名、组织机构及文献列表进行的，委托人信息如下：

姓名：石翠萍

机构：齐齐哈尔大学

检索范围：

- 科学引文索引（Science Citation Index Expanded）： 1900年-2024年
- 期刊引证报告（Journal Citation Reports）： 1997年-2023年
- 中科院期刊分区表： 2005年-2023年

检索结果：

检索类型	数据库	年份范围	记录数
SCI-E 收录	SCI-EXPANDED	2024	1
JCR 影响因子	JCR	1997 - 2023	1
中科院期刊分区	中科院分区	2005 - 2023	1

检索人员：张 玲

哈尔滨工业大学科技查新站

教育部科技查新站（L39）

2024年9月9日

附件一：**SCI-E** 收录

#	作者	标题	来源出版物	JCR影响因子	中科院分区	文献类型	入藏号
1	**Shi, CP**; Yue, SH; Wang, LG	Attention Head Interactive Dual Attention Transformer for Hyperspectral Image Classification	***IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*** 2024, 62: 5523720.	• 7.5 (2023);	• 小类(升级版) (2023) 遥感 **[2区]**; • 小类(升级版) (2023) 成像科学与照相技术 **[2区]**; • 小类(升级版) (2023) 工程：电子与电气 **[2区]**; • 小类(升级版) (2023) 地球化学与地球物理 **[1区]**; • 大类(升级版) (2023) 地球科学 **[1区]**;	J Article	WOS:001276870500031
						合计	1

第 1 条，共 1 条：
标题： Attention Head Interactive Dual Attention Transformer for Hyperspectral Image Classification
作者： Shi, CP (Shi, Cuiping); Yue, SH (Yue, Shuheng); Wang, LG (Wang, Liguo)
来源出版物： IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING **卷：** 62 **文献号：** 5523720 **出版年：** 2024
入藏号： WOS:001276870500031
文献类型： Article **出版物类型：** J
作者地址： [Shi, Cuiping] Huzhou Univ, Coll Informat Engn, Huzhou 313000, Peoples R China.; [Shi, Cuiping; Yue, Shuheng] Qiqihar Univ, Dept Commun Engn, Qiqihar 161000, Peoples R China.; [Wang, Liguo] Dalian Nationalities Univ, Coll Informat & Commun Engn, Dalian 116000, Peoples R China.
所属机构： Huzhou University; Qiqihar University; Dalian Minzu University
通讯作者地址： Yue, SH (corresponding author), Qiqihar Univ, Dept Commun Engn, Qiqihar 161000, Peoples R China.
电子邮件地址： shicuiping@qqhru.edu.cn; 2021910320@qqhru.edu.cn; wangliguo@hrbeu.edu.cn

出版商：IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC 出版商城市：PISCATAWAY 出版商地址：445 HOES LANE, PISCATAWAY, NJ 08855-4141 USA

Web of Science 类别：Geochemistry & Geophysics; Engineering, Electrical & Electronic; Remote Sensing; Imaging Science & Photographic Technology

研究方向：Geochemistry & Geophysics; Engineering; Remote Sensing; Imaging Science & Photographic Technology

IDS 号：ZQ9E6

ISSN: 0196-2892 **eISSN:** 1558-0644

基金资助机构和授权号：National Natural Science Foundation of China [42271409]; Heilongjiang Science Foundation Project of China [LH2021D022]

基金资助致谢：This work was supported in part by the National Natural Science Foundation of China under Grant 42271409 and in part by Heilongjiang Science Foundation Project of China under Grant LH2021D022.

JCR 影响因子：

期刊	**JCR** 影响因子	指标年份
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING	7.5	2023

中科院期刊分区：

期刊	类型	学科类别	分区	指标年份	TOP
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING	大类(升级版)	地球科学	1	2023	否
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING	小类(升级版)	工程：电子与电气 (ENGINEERING, ELECTRICAL & ELECTRONIC)	2	2023	-
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING	小类(升级版)	地球化学与地球物理 (GEOCHEMISTRY & GEOPHYSICS)	1	2023	-
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING	小类(升级版)	成像科学与照相技术 (IMAGING SCIENCE & PHOTOGRAPHIC TECHNOLOGY)	2	2023	-
IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING	小类(升级版)	遥感 (REMOTE SENSING)	2	2023	-

End