Deformable Convolution-Enhanced Hierarchical Transformer With Spectral-Spatial Cluster Attention for Hyperspectral Image Classification

Yu Fang, Le Sun, *Senior Member, IEEE*, Yuhui Zheng, *Member, IEEE*, and Zebin Wu, *Senior Member, IEEE*

***Abstract*— Vision Transformer (ViT), known for capturing non-local features, is an effective tool for hyperspectral image classification (HSIC). However, ViT's multi-head self-attention (MHSA) mechanism often struggles to balance local details and long-range relationships for complex high-dimensional data, leading to a loss in spectral-spatial information representation. To address this issue, we propose a deformable convolution-enhanced hierarchical Transformer with spectral-spatial cluster attention (SClusterFormer) for HSIC. The model incorporates a unique cluster attention mechanism that utilizes spectral angle similarity and Euclidean distance metrics to enhance the representation of fine-grained homogenous local details and improve discrimination of non-local structures in 3D HSI and 2D morphological data, respectively. Additionally, a dual-branch multiscale deformable convolution framework augmented with frequency-based spectral attention is designed to capture both the discrepancy patterns in high-frequency and overall trend of the spectral profile in low-frequency. Finally, we utilize a cross-feature pixel-level fusion module for collaborative cross-learning and fusion of the results from the dual-branch framework. Comprehensive experiments conducted on multiple HSIC datasets validate the superiority of our proposed SClusterFormer model, which outperforms existing methods. The source code of SClusterFormer is available at https://github.com/Fang666666/HSIC_SClusterFormer.**

***Index Terms*— Hyperspectral image classification, hierarchical Transformer, multi-feature, cluster attention.**

I. Introduction

HYPERSPECTRAL images (HSIs) involve the acquisition of extensive data comprising hundreds of contiguous spectral bands, which are combined with spatial information to construct a data cube for the analysis of land covers [1]. The applications of HSIs span diverse fields, including environmental monitoring, geological exploration, precision agriculture, national defense [2], medical imaging [3], [4] and forensic science [5]. Recently, researchers have conducted numerous investigations in the field of HSI processing [6], [7], [8], [9], [10], [11], [12], [13], [14]. Notably, hyperspectral image classification (HSIC) has emerged as a pivotal research topic, leveraging spectral-spatial cues in HSI to establish a direct correspondence between pixels and predefined land cover categories. However, HSIC faces substantial challenges, including the inherent high dimensionality of the data, spectral band similarities, data redundancy, and limitations in spatial resolution.

The challenges posed by the characteristics of HSI data require researchers to extract higher-quality features to improve classification performance. Over the past decades, traditional HSIC methods have preliminarily addressed issues such as band redundancy and the large volume of data by extracting carefully crafted manual features. In the early days, support vector machine (SVM) was widely used for spectral-spatial feature extraction, and several variants were developed to address issues of stability and discriminative power [15], [16]. Notably, multi-kernel SVM achieved better classification results by incorporating spatial relationships into spectral correlations [17], while the combination of extended morphological attribute profiles (EMAP) with SVM also yielded competitive results [18], [19]. Additionally, K-nearest neighbor (KNN) [20] and random forest (RF) [21], [22] methods leveraged spectral information to capture the correlations between different bands and facilitate category mapping. However, these traditional methods rely on manual feature extraction, which requires a high level of specialized domain knowledge and offers limited generalization capabilities when applied to complex feature data.

Recently, significant progress has been made in deep neural network (DNN)-based HSIC processing methods, which can automatically learn data features, handle nonlinear relationships, and perform well on large and high-dimensional datasets [23]. Convolutional neural networks (CNNs) have been particularly successful in HSIC applications due to their strong ability to capture both spectral and spatial information. HU et al. introduced CNN into HSIC, and output the spectral information of the image in the form of a one-dimensional convolution, obtaining exciting results [24]. Zhao and Du

Received 25 March 2024; revised 5 November 2024; accepted 13 December 2024. Date of publication 1 January 2025; date of current version 27 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant U23B2006 and Grant 62471239. The associate editor coordinating the review of this article and approving it for publication was Prof. Manoranjan Paul. *(Corresponding author: Le Sun.)*

Yu Fang is with the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: cse_yfang@njust.edu.cn).

Le Sun and Yuhui Zheng are with the School of Computer Science and Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: sunlecncom@163.com; zheng_yuhui@nuist.edu.cn).

Zebin Wu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: zebin.wu@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2024.3522809, provided by the authors.

Digital Object Identifier 10.1109/TIP.2024.3522809

introduced a two-dimensional convolution, and the spatial convolutional feature extracted in multiple spectral dimensions could accurately process detailed features [25]. Since then, DNNs have been rapidly developed in the field of HSIC. Paoletti's team used deep residual structures for HSIC releasing the limitations of CNNs on generalization of high-dimensional redundant information, and obtained excellent results [26]. To extract deeper joint spatial-spectral features, Uzair et al. proposed a local similarity projection Gabor filtering (LSPGF) method, which combined a local similarity projection (LSP) of CNN as well as a 2D-Gabor filtering to respond more texture and edge features [27]. In the process of spatial-spectral extraction, the hybrid of 3D-CNN and 2D-CNN is often used to ensure the relevance of individual bands to the context of feature details as well as to expand the convolutional receptive field. But the inefficient connectivity of the two convolutional networks will reduce the efficiency of the parameters used in the computation. Ma's team proposed a lightweight hybrid CNN method that utilized a deeply separable convolution to explore the efficiency of connectivity between convolutions of different dimensions with guaranteed accuracy [28]. However, the over-reliance of DNNs on convolutional operations often causes these methods to focus excessively on fine textures and image edges, while overlooking the balance needed to capture global feature information. In our work, we introduce a multi-scale deformable convolution feature with frequency-based spectral attention (MDC-FSA) to early select localized sampling focuses. This method captures regions of potential interest at various scales through flexible convolutional sampling positions while preserving the finest curve features from the high-low frequency spectral representation. By addressing the limitations of convolutional receptive fields in large-scale images, this approach mitigates the loss of spectral-spatial information extracted from high-dimensional HSI data. To enhance spatial feature representation at low resolutions, we draw on the morphological EMAP feature for spatial representation improvement. Additionally, we incorporate another deformable convolutional branch specifically for EMAP feature extraction, which significantly boosts classification performance, especially in scenarios with limited samples and imbalanced sample distribution.

The problem of convolutional feature receptive field limitations, which plague the natural images, is more evident in large-scale remote sensing images. Recently, the domain of computer vision has introduced the global model called Vision Transformer (ViT) tending to replace CNN. ViT splits the image into different patches, and converts them into the vectors named tokens. Subsequently, all tokens are subject to simultaneous attention computation through the multi-head self-attention mechanism (MHSA), which is also often referred as the token mixer. After the computation, tokens can capture the global spectral-spatial information and better model the complex long-distance dependencies. Hong et al. modeled the global dependency of neighboring spectral bands as a group, which well preserved the sequence information between spectral bands [29]. This is challenging to achieve using CNNs. The hybrid 3D-CNN and 2D-CNN approach mentioned above can also be effectively combined with Transformer. Sun et al. [30] proposed a spectral-spatial feature tokenization Transformer (SSFTT). It extracted high-level abstract features using hybrid convolutions, explored local features, and assigned Gaussian semantics to the tokens, and then used the transformer for global modeling. In addition, Fang et al. proposed a lightweight Transformer encoder, which used multiple attention tricks to achieve highly expressive tokens, and then fed them into a multi-scale pooling mixer for global computation. It can achieve high classification accuracy even under unbalanced sample distribution, and also effectively reduce the number of parameters [31].

Lately, the overall structure of ViT has begun to be immobilized, and the hierarchical algorithmic processing can more adequately handle long sequences composed of more bands. Hierarchical Transformer can flexibly adapt to data with different spectral distributions and has more powerful modeling capabilities in complex classification situations such as scarce samples. Mei et al. [32] proposed a hierarchical Transformer that utilized HSI grouped pixels for embedding inputs to exploit contextual information in a global-local manner. Xu et al. applied the sliding window mechanism to the spectral dimension and proposed a spatial-spectral 1DSwin Transformer [33]. The hierarchical Transformer solves the limitation of CNN receptive field structurally and realizes the effective combination of global and local. However, it remains unclear whether mixers utilizing MHSA can effectively balance global and local information, particularly in the context of the unique high-dimensional nature of HSI data. Often, MHSA places too much emphasis on overall modeling, neglecting the processing of local detailed features. In our work, we propose the cluster-based attention mixer called spectral cluster attention (SpeCA) and spatial cluster attention (SpaCA). They enable the computation of clusters for feature maps from both spectral and spatial perspectives, followed by the aggregated learning of intra- and inter-cluster relationships to effectively leverage local texture details and global structural patterns. Furthermore, the coupling and balance of local and non-local information are significantly enhanced through the simultaneous processing of both intra- and inter-clusters.

In summary, we propose a deformable convolution-enhanced hierarchical Transformer with spectral-spatial cluster attention (SClusterFormer) to facilitate effective large-scale classification of complex HSI. The approach makes the following contributions:

1. SClusterFormer designs a two-branch hierarchical Transformer and an effective branch fusion approach. Spectral-spatial features and morphological features are simultaneously exploited and a cross-feature pixel-level fusion strategy is used to realize an effective complementarity between the both.

2. SClusterFormer introduces a novel attention computation mechanism termed spectral cluster attention and spatial cluster attention. SpeCA clusters feature maps by computing spectral angle similarity, capturing local fine-grained features and non-local discriminative features intra- and inter-cluster. SpaCA, in turn, utilizes Euclidean distance similarity so as to understand spatial texture information and contextual

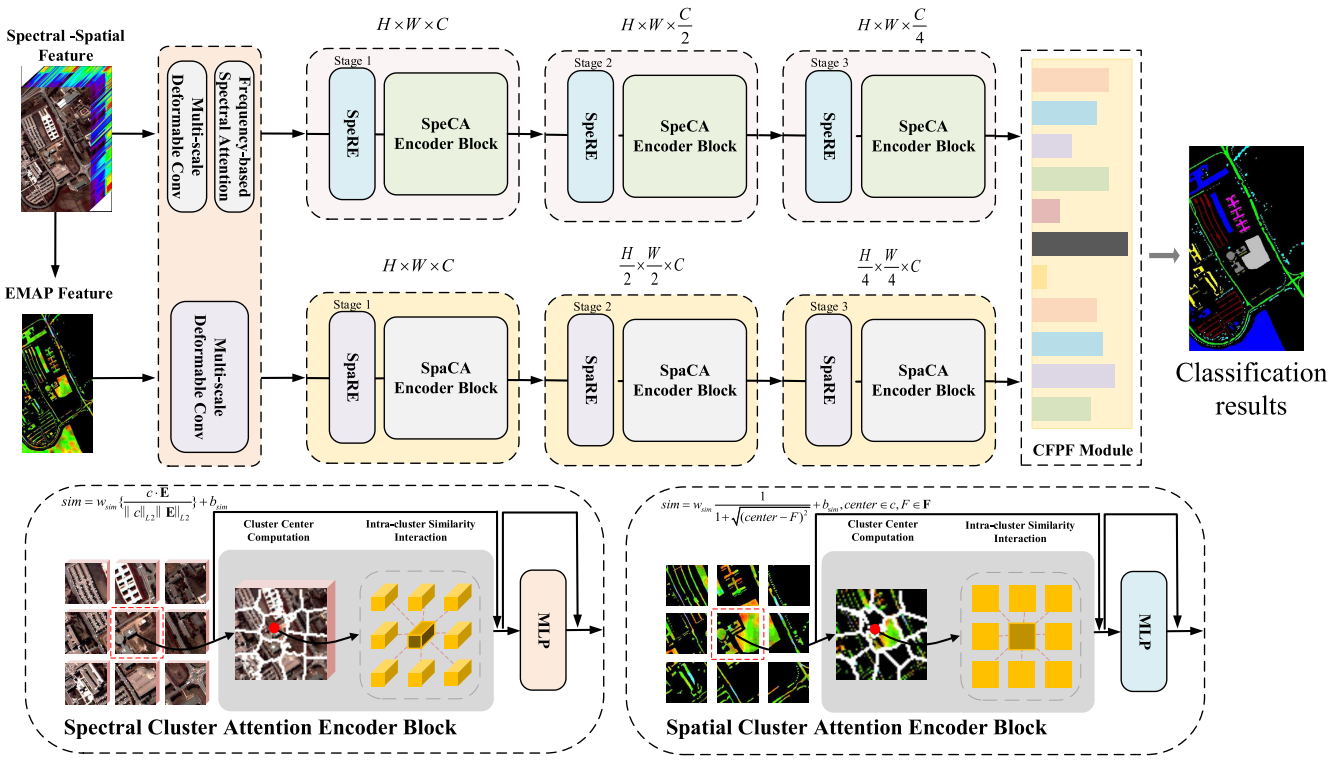


Fig. 1. Schematic of the SClusterFormer model for HSIC. The network is divided into upper and lower branches. They utilize different forms of deformable convolution to extract original HSI features and morphological EMAP features, respectively. Each of the branches consists of three similar stages. The stages in the upper branch consist of the Spectral Reducer Embedding and Spectral Cluster Attention Encoder Block, and the stages in the lower branch consist of the Spatial Reducer Embedding and Spatial Cluster Attention Encoder Block. Finally, the cross-fusion of the dual branches is realized by the CFPF module. Below the schematic is the SpeCA Encoder Block and SpaCA Encoder Block simple flow, which describes two residual structures. In the first residual, it is our proposed SpeCA and SpaCA, which both compute the cluster centers first and then the similar interactions for the clusters.

distribution. Such cluster attention effectively enhances the coupled modeling of local and non-local information.

3. SClusterFormer employs the multi-scale deformable convolution feature with frequency-based spectral attention to overcome the receptive field defects of large-scale remote sensing images, effectively retaining the key detailed features in the high-frequency and the overall trend of the spectral profile in the low-frequency. This provides a more accurate convolution representation for the subsequent transformer encoder.

4. SClusterFormer also performs well in small samples, especially for unbalanced sample distribution and large scales data. Experiments on all three datasets show the excellent results we achieved.

The remainder of this article is organized as follows. Section II reviews the knowledge related to the algorithm in depth and details of the proposed SClusterFormer. Section III gives a lot of experiments and analysis. Finally, Section IV summarizes this paper and puts forward research lines for future work.

II. Materials and Methodology

Fig. 1 illustrates the general framework diagram of the HSIC task using the proposed SClusterFormer. Its two-branches network contains three main components: a multi-scale deformable convolution feature extraction with frequency-based spectral attention (MDC-FSA) module, spectral and spatial cluster attention hierarchical Transformer, and a cross-feature pixel-level fusion (CFPF) module. Next, we explore these three main modules separately.

A. Multi-Scale Deformable Convolution Feature With Frequency-Based Spectral Attention Extraction

1) Multi-Scale Deformable Convolution Processing: The general convolution features are sampled at predetermined positions with a fixed-size convolution kernel. However, in the face of large-scale HSI data, it is difficult to accurately deal with complex surface conditions and flexible spectral curves with a fixed receptive field. Thus, the deformable convolution is used for complex landform feature extraction. Specifically, the learnable offsets are introduced additional into the convolutional receptive field, which can be closer to the actual terrain shape. Then, we further expand the deformable field by setting three scales of convolutional kernels for the upper branch and two for the lower branch, which greatly enhances the model's ability to capture contextual information. In order to avoid the curse of dimensionality caused by feature redundancy and to reduce computational complexity, we perform principal components analysis (PCA) operation for the upper branch feature to select the most representative bands before the convolution operation. Let the input HSI data $\mathbf{I} \in \mathbf{R}^{m\times n\times l}$, where $m \times n$ is the spatial dimension and l is the number of spectral bands, be downscaled by the PCA operation to $\mathbf{I}_{\text{pca}} \in \mathbf{R}^{m\times n\times b}$. The feature is then sliced from the spatial dimension into patches of size $s \times s$. Subsequently, the multi-scale deformable

convolution (MDC) extraction operation is performed, and the propagation operation of the feature cube in the i-th layer at the (x, y, z) position of the upper branch is as follows:

$$\begin{aligned}
\alpha_i^{xyz} &= \sum_{a,b,c} \mathbf{K}_L(a,b,c) \cdot \mathbf{I}_{pca}(x+a+\Delta a, y+b+\Delta b, z \\
&\quad + c + \Delta c) \\
\beta_i^{xyz} &= \sum_{a,b,c} \mathbf{K}_M(a,b,c) \cdot \mathbf{I}_{pca}(x+a+\Delta a, y+b \\
&\quad + \Delta b, z+c+\Delta c) \\
\delta_i^{xyz} &= \sum_{a,b,c} \mathbf{K}_S(a,b,c) \cdot \mathbf{I}_{pca}(x+a+\Delta a, y+b \\
&\quad + \Delta b, z+c+\Delta c) \\
v_i^{xyz} &= \text{cat}(\alpha_i^{xyz}, \beta_i^{xyz}, \delta_i^{xyz}) \otimes \text{Filter}_{1\times1\times1} + \beta_i{}^{xyz} \qquad (1)
\end{aligned}$$

where $\mathbf{K}_L$, $\mathbf{K}_M$, and $\mathbf{K}_S$ are the size of the convolution kernel at three different scales and (a, b, c) is the regular positions within the kernel. The deformable convolution introduces an offset position $(\Delta a, \Delta b, \Delta c)$ for each pixel separately compared with the original convolution making the sampling position more flexible. However, these offset positions are not integers and do not correspond directly to actual pixel locations in the feature map. To address this, we use bilinear interpolation to compute the pixel values at the offset positions, as described in **Appendix A** of the **Supplementary Material**. Additionally, we concatenate and apply unit-mapping to the convolutional features from the different scales. The features of MDC extraction in the upper branch, which capture spectral-spatial information, are completed and illustrated in Fig. 2.

In the lower branch, we primarily extract the rich spatial information from the EMAP features, which helps address the issue of spatial information being neglected due to the high-dimensional data in the upper branch. EMAP features are based on the principle of mathematical morphology, and realize the specific description of image objects and region details. For the specific mathematical principles of EMAP feature extraction, please see **Appendix B** section of **Supplementary Material**, and we directly use $\mathbf{I}_{emap} \in \mathbf{R}^{m\times n\times 1}$ for its tensor representation. Similarly, we slice the features into patches of size $s \times s$ and perform the MDC operation on them, but in this case, we only use the convolution kernel for 2D data. The propagation operation for the EMAP deformable convolutional feature in layer i at position (x, y) is as follows:

$$\begin{aligned}
\alpha_i^{xy} &= \sum_{a,b} \mathbf{K}_L(a,b) \cdot \mathbf{I}_{emap}(x+a+\Delta a, y+b+\Delta b) \\
\beta_i^{xy} &= \sum_{a,b} \mathbf{K}_S(a,b) \cdot \mathbf{I}_{emap}(x+a+\Delta a, y+b+\Delta b) \\
v_i^{xy} &= \text{cat}(\alpha_i^{xy}, \beta_i^{xy}, \delta_i^{xy}) \otimes \text{Filter}_{1\times1} + \beta_i{}^{xy} \qquad (2)
\end{aligned}$$

where $\mathbf{K}_L$ and $\mathbf{K}_S$ are the size of the convolution kernel for deformable convolution, (a, b) is the regular positions of the convolution kernel, and it also introduces an offset position $(\Delta a, \Delta b)$ for the spatial information. At this stage, the spectral-spatial features from the upper branch and the morphological features from the lower branch have been extracted. However, for the spectral sequence, the model may become overly sensitive to retained spectral noise, especially after dimensionality reduction. To mitigate this, we employ frequency-based spectral attention (FSA) in the upper branch. FSA enhances spectral features by dynamically adjusting the weights of key spectral bands, leveraging a combination of high- and low-frequency information to adapt to different scenes and conditions.

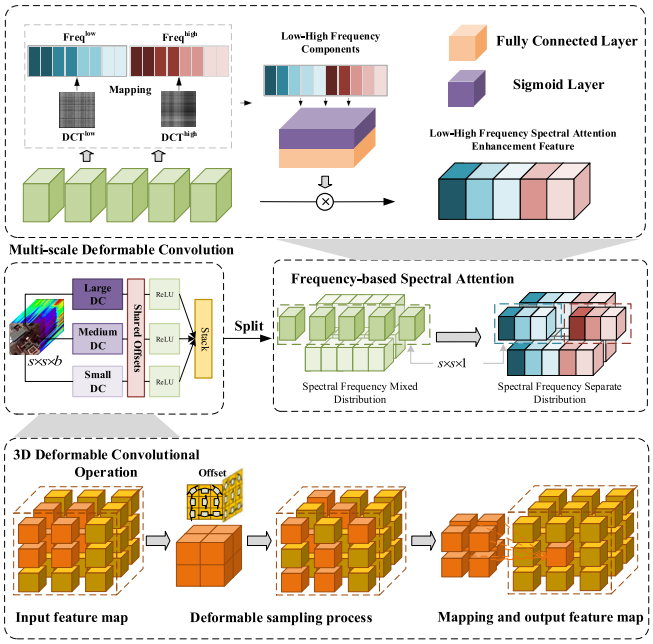


Fig. 2. Illustration of the upper branch: Multi-scale deformable convolution feature with frequency-based spectral attention (MDC-FSA).

2) Frequency-Based Spectral Attention Enhancement: The principle of channel attention is to compress different channel features using a pooling layer, enabling the calculation of attention weights. The specific mechanism can be expressed as:

$$att = sigmoid(fc(\phi(\mathbf{X}))) \qquad (3)$$

where $\mathbf{X}$ represents the input high-dimensional feature map, *sigmoid* is the activation function, *fc* denotes the fully connected layer, and *att* is the output attention weight. The ϕ function reduces the three-dimensional tensor, represented by C, H, and W (channels, height, and width), into a vector of C dimensions. Some current research suggests that the compress function ϕ can be represented using a form of frequency analysis, which is expressed as:

$$\phi_{h,w} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j} [\cos(\frac{\pi h}{H}(i+\frac{1}{2})) \cos(\frac{\pi w}{W}(j+\frac{1}{2}))] \quad (4)$$

where $i \in \{0, 1, \cdots, H-1\}$ and $j \in \{0, 1, \cdots, W-1\}$. This operation transforms the original features into the frequency domain using the discrete cosine transform (DCT). Often-used channel attention based on global average pooling (GAP) is a special case when $h, w = 0$. However, relying on GAP in some scenario can lead to significant loss of other frequency components, thus reducing the richness of the extracted feature information. To better compress the spectral

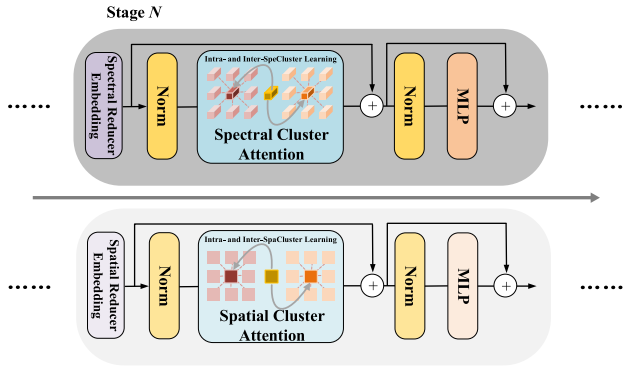


Fig. 3. Illustration of spectral and spatial cluster attention encoder block.

features of HSI data while retaining more information, we propose a frequency-based spectral attention feature enhancement method. This method retains both the low-frequency and high-frequency compressed information of the tensor, enabling the enhancement of high-frequency detail features in the spectral information. The specific calculation formula is as follows:

$$\phi_{i,j} = \text{cat}(\sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}(\sin \pi i \cdot \sin \pi j), \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}) \quad (5)$$

Consequently, the attention score at this stage is given by: $fs\_att = sigmoid(fc(\phi_{i,j})$, as represented in Fig. 2. We then enhance the multi-scale spectral-spatial deformable convolutional features in the upper branch using FSA to obtain a new feature representation $v^{xyz}_{enchanced}$.

B. Spectral and Spatial Cluster Attention Hierarchical Transformer Encoder

1) Overall Architecture: The resulting features after the above module, i.e., $\mathbf{X}_1 \in \mathbf{R}^{p\times p\times b}$ and $\mathbf{X}_2 \in \mathbf{R}^{p\times p\times c}$, already have a high capability to represent spectral and spatial information, where p is the size of the convolution feature map, b is the number of spectral information bands, and c is the number of channels with the morphological information. Next, we model the feature maps by the proposed SpeCA and SpaCA based on transformer's bi-residual structure to enhance the aggregation learning of similar local details and the discriminative computation of global patterns. As shown in Fig. 3, we propose two Transformer blocks belonging to different branches. There are three stages in one branch, where each stage in the upper branch consists of spectral reducer embedding (SpeRE) and spectral cluster attention encoder block, where each stage in the lower branch is transformed into spatial reducer embedding (SpaRE) and spatial cluster attention encoder block due to its morphological spatial information only. Spectral cluster attention encoder block and spatial cluster attention encoder block are composed of multiple spectral cluster attention (SpeCA) and spatial cluster attention (SpaCA). The main function of the SpeRE module is to slice the features along the spectral dimension and embed them into token form. The number of spectra is set to C by SpeRE, and subsequently halved for the 2nd and 3rd stage respectively, i.e., $C/2$, $C/4$. For the lower branch of SpaRE, instead of dealing with channel dimensions, it downsamples the spatial information to continuously increase the receptive field and retain more abstract, high-level feature representations. SpaRE module sets feature size to $H \times W \times C$. Thus the input sizes for each stage are $H \times W \times C$, $H/2 \times W/2 \times C$, and $H/4 \times W/4 \times C$. Finally, the results of dual branches are fed into the CFPF module for fusion operation with feature map sizes of $H \times W \times C/4$ and $H/4 \times W/4 \times C$, respectively.

2) Upper Branch: Spectral Reducer Embedding and Spectral Cluster Attention: The SpeRE module attempts to maintain the two-dimensional shape of the input data while mapping and compressing the number of spectra to fit the demands of each stage's spectral receptive field, which is different from ViT. The specific propagation formula is as follows:

$$SpeRE(\mathbf{X}_{hsi}) = \text{ReLU}(\text{BN}(\text{Conv2d}(\mathbf{X}_{hsi}))) \quad (6)$$

Among them, Conv2d is a 2D convolution operation, which aims to realize the mapping transformation of the input feature over the number of spectra. Here we set the convolution kernel size to 1×1 to ensure that the spatial information of the feature map is fixed. BN is the Batch Normal layer and ReLU is the activation function, which guarantees better convergence between different bands while introducing nonlinearities to prevent gradient vanishing. The whole process is very concise, without excessive slicing and transforming of spatial data, and retains the spectral-spatial information to a great extent.

In recent years, MHSA has been considered as the most representative module in Transformer. However, since MHSA usually attributes weights based on the relationships between different elements, and when the input is limited, global modeling can result in the loss of spatial-spectral information. Therefore, we propose a SpeCA for global modeling. The SpeCA obtains the centroids of multiple adaptive partitions based on the spectral-spatial information of the input local feature maps. By mutual learning the spectral-spatial information between the centroid and the surrounding information, local detailed features are better captured. Also, SpeCA allows information to transfer between different clusters through a multi-heads mechanism for global context. This clustering idea greatly enhances the balance of global and local information, and improves the classification performance by optimizing details from the global perspective.

We take the first stage as an example here. The data after the SpeRE module is $\mathbf{E} \in \mathbf{R}^{H\times W\times C}$. Subsequently, the SpeCA module is performed, and its structure is shown in Fig. 4. First, the input data are divided into h sub-features along the spectral dimension and are respectively reduced back to the number of spectra by 2D convolutional mapping C. A three-dimensional adaptive pooling operation is performed in each sub-feature distribution, and the adaptive window is generally set as $(1, e, e)$ to easily obtain the centroid values of each clustered region. Where $e \times e$ is the crude measurement of adaptive cluster regions, we have to ensure that $e < H, W$ to obtain an adaptive window. Their centers are the centroids that take into account the spectral-spatial information. Subsequently, we calculate the similarity between each centroid and the surrounding proximity points separately. Here we use cosine

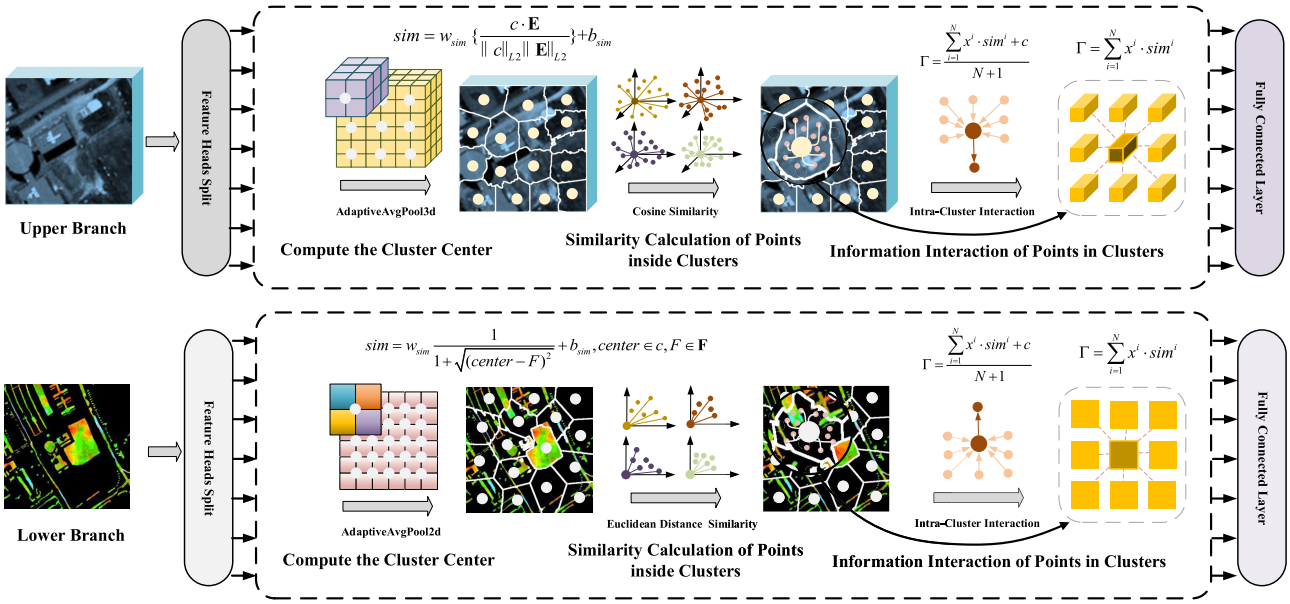


Fig. 4. Illustration of spectral and spatial cluster attention.

similarity of the spectral angle calculation and the specific formula for the calculation process is as follows:

$$sim = w_{sim} \cdot \cos(c, \mathbf{E}) + b_{sim}$$
$$\cos(c, \mathbf{E}) = \frac{c \cdot \mathbf{E}}{\|c\|_{L2} \|\mathbf{E}\|_{L2}} \tag{7}$$

where c is the centroid matrix, $\|c\|_{L2}$ and $\|\mathbf{E}\|_{L2}$ are the $L2$ normalization which map each element of the input feature onto the unit hypersphere, guaranteeing a modulus of 1. Meanwhile, we introduce learnable parameters w_{sim} and b_{sim}, which control the scaling factor and offset value of the similarity, respectively. Subsequently, we generate the mask $mask_{sim}$ of the similarity matrix sim, which retains the maximum similarity centroid information of each point in the feature. Then the cluster operation is completed, and each point in the feature map is assigned to the center point region with the highest similarity to it. Next, the aggregation operation is carried out, and the information of each point in the cluster interacts with its centroid in a signature interaction, which improves the understanding of the spectral-spatial information of each centroid. The formula for the specific interaction method is as follows:

$$\Gamma = \frac{\sum_{i=1}^{N} x^i \cdot sim^i + c}{N+1} \tag{8}$$

where x is the points within the cluster, N is their number, and c is the cluster centroid. The intra-cluster aggregation operation helps to capture local features in HSI data and effectively utilize the context information. In addition, the centroid after feature enhancement is realized requires to dispatch the clustered information to the context points so that the feature of the points within the cluster can interact with each other. The formula for the specific allocation method is as follows:

$$\Gamma = \sum_{i=1}^{N} x^i \cdot sim^i \tag{9}$$

where x is each point in the cluster after aggregation, and they interact with the centroid separately to achieve the distribution of centroid information. Each point acquires the center point feature and also retains the key information of the other points interacting.

At this point, we have completed the spectral cluster attention operation for a sub-feature, i.e., a single head. In SpeCA we perform clustering interactions of h heads at the same time, so we need to cascade the results of each head and realize the inter-cluster learning through the fully connected layer to preserve the global information:

$$out = \mathrm{FC}(\mathrm{Concat}(\Gamma_1, \Gamma_2, \ldots, \Gamma_h)) \tag{10}$$

With the SpeCA module, we re-emphasize the importance of local information in the hierarchical Transformer and enhance the communication between local and global features to achieve the balance between both.

3) Lower Branch: Spatial Reducer Embedding and Spatial Cluster Attention: As for the lower branch, its inputs are EMAP features containing only spatial information. Hence SpaRE will emphasize mapping downsampling of spatial dimensions to satisfy the demand for spatial receptive fields from multiple stages. The propagation formulation of SpaRE is the same as Formula 6. In this case, the kernel size, stride, and padding of the 2D convolution operation are changed to 3, 2, and 3/2 in the 2nd and 3rd stages to ensure that the downsampling of spatial information is realized.

We also use the first stage as an example here. The data after the SpaRE module is $\mathbf{F} \in \mathbf{R}^{H \times W \times C}$. Afterward, the SpaCA module is executed, which is also shown in Fig. 4. We again divide the input data into h sub-features along the

channel dimension. Then we adapt two-dimensional adaptive pooling instead to obtain the cluster centroids, where the adaptive window is set to (f, f) and guaranteed to be $f < H, W$. Subsequently, we compute the similarity between each centroid and the other points in the feature map using the Euclidean distance as the similarity measure. The specific formula for the calculation process is as follows:

$$sim = w_{sim}\frac{1}{1+d} + b_{sim}$$
$$d = \sqrt{(center - F)^2}, center \in c, F \in \mathbf{F} \quad (11)$$

where $center$ is the point in centroid matrix c and d is the Euclidean distance which is inversely proportional to the similarity. Likewise, we use the similarity mask matrix to retain the centroid information of the maximum similarity for each point and assign to the centroid. Next, the aggregation operation and the dispatch operation are performed, which are the same here as in Formula 8 and Formula 9. Finally, we equally need to cascade the results of each head and implement inter-cluster learning through the full connected layer. In the SpaCA module, we perform the cluster attention operation again for the spatial information using Euclidean distance, while continuously reducing the spatial scale in the subsequent stages to ensure that sufficient local-global communication are achieved for the morphological spatial features.

C. Cross-Feature Pixel-Level Fusion

After three stages of feature interaction, the output features of the upper and lower branches have already possessed high expressiveness and interpretability. The direct fusion of operations may not be able to effectively utilize both to achieve better results. Therefore, we design a new fusion strategy from the perspective of categorizing output pixels, aiming to achieve the effect of "one plus one is greater than two".

A cross-feature pixel-level fusion (CFPF) module is proposed. First, we design a cross-feature attention fusion method to interactively learn the upper and lower branches to generate a third type of feature. The importance of this step is that it exposes complementary information from different branches. Then the three types of features simultaneously perform feature mapping to transform them into pre-classification results. Finally, we introduce learnable weight coefficient variables and learn them as coefficients with the three types of features. Through this step, we can fuse the various types of features effectively and obtain a more expressive and adaptive classification result. The specific flow of CFPF is shown in Fig. 5.

We first randomly generated three different learned weight matrices $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$. Subsequently, the output of the lower branch represented by $\mathbf{T}_1$, which was originally featured as EMAP, is linearly mapped to the $\mathbf{K}$ and $\mathbf{V}$ matrices, and the output of the upper branch represented by $\mathbf{T}_2$ is linearly mapped to the $\mathbf{Q}$ matrix.

$$\mathbf{Q} = \mathbf{W}_q\mathbf{T}_2$$
$$\mathbf{K} = \mathbf{W}_k\mathbf{T}_1$$
$$\mathbf{V} = \mathbf{W}_v\mathbf{T}_1 \quad (12)$$

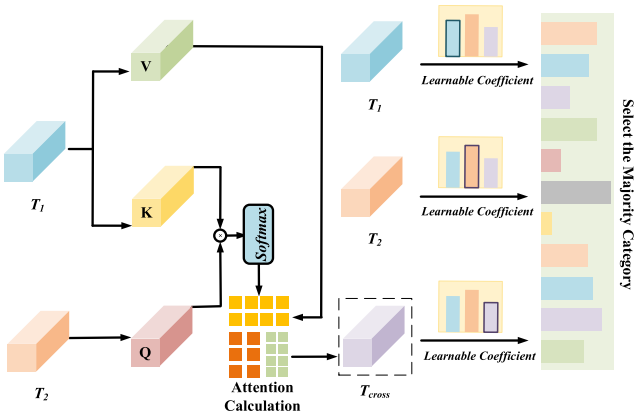


Fig. 5. Illustration of cross-feature pixel-level fusion (CFPF).

$$\mathbf{T}_{cross} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (13)$$

where d_k is the number of columns of the $\mathbf{Q}$ and $\mathbf{K}$ matrices. So far we have generated a new cross-feature. Subsequently, the three types of features $\mathbf{T}_1$, $\mathbf{T}_2$, $\mathbf{T}_{cross}$ are assigned a learnable weight coefficients, respectively, and the sum of the coefficients is guaranteed to be 1. The learnable coefficients of the $\mathbf{T}_{cross}$ features can be assigned a larger value at initialization, as it is the result of cross-learning of the other two types of features, and has stronger spectral-spatial representation. Features are subjected to a linear layer separately to obtain the pre-classification results of each pixel separately. Under the same pixel, the number of three classifications is counted by adding the weights, and the category with the highest number is the final feature category for that pixel. The fusion can be represented as follows:

$$Out(pi) = \max_{n=[1,2,\ldots,N]} \Phi(pi, n)$$
$$\Phi(pi, n) = \sum_{k=1,2,3} \theta(pi, n) \quad (14)$$

where pi is the order of the pixels, n is the order of the ground categories, k is the number of feature outputs, i.e., 3, and θ denotes the weight coefficients of each feature. In this regard, the CFPF module synthesizes the classification results of different features to provide more comprehensive and richer information, and fully exploits the complementarity between them. It makes up for the shortcomings of a single feature, better adapts to complex landform classification tasks, and enhances the robustness, and generalization ability of the overall model.

D. Algorithm Summarization for SClusterFormer

The overall process of the proposed SClusterFormer network is shown in **Algorithm 1**.

III. Experiment and Analysis

We have carried out a series of experiments on three well-known HSI datasets Pavia University, WHU-Hi-HanChuan and WHU-Hi-HongHu to demonstrate the performance advantages and drawbacks of our proposed SClusterFormer for HSIC. The description of the datasets

Algorithm 1 SClusterFormer Network

Input:

HSI data $\mathbf{I} \in \mathbf{R}^{m\times n\times l}$, ground-truth $\mathbf{Y} \in \mathbf{R}^{m\times n}$; PCA bands number b; patch size s; training sample rate $\mu\%$; SpeCA cluster centroids number c_1; SpaCA cluster centroids number c_2; attention heads number h; epochs Δ; batch size = bs; learning rate = lr;.

Output:

Predicted labels of the test dataset.

Obtain $\mathbf{I}_{pca}$ feature from HSI data and convert them into multi-scale deformable convolution features V_{pca}^{xyz} by Eq. (1).
Obtain $\mathbf{I}_{emap}$ feature from HSI data and convert them into deformable convolution features V_{emap}^{xy} by Eq. (2).
The Upper Branch feature is enhanced with FSA to obtain the$V_{enchanced}^{xyz}$
for $i = 1$ to Δ **do**
  for $j = 1$ to 3 **do**
    Execute SpeRE in the upper branch to obtain token data by Eq. (6).
    Obtaining Cluster Centroid Positions in the upper branch Using 3D Adaptive Pooling.
    Calculate the spectral angular similarity of each pixel to all centroids and retain the most similar centroids to generate clusters by Eq. (7).
    Achieve mutual interaction between pixels in the cluster and the centroid and generate feature $\mathbf{T}_1$ by Eq. (8) (9) (10).
    Execute SpaRE in the lower branch to obtain token data.
    Obtaining Cluster Centroid Positions in the lower branch Using 2D Adaptive Pooling.
    Calculate the Euclidean distance similarity of each pixel to all centroids and retain the most similar centroids to generate clusters by Eq. (11).
    Achieve mutual interaction between pixels in the cluster and the centroid and generate feature $\mathbf{T}_2$ by Eq. (8) (9) (10).
  end for
  Cross-learning for $\mathbf{T}_1$ and $\mathbf{T}_2$ generates feature $\mathbf{T}_{cross}$
  Learning the pre-classification weights and assigning them to three types of features, and fusing pixels to identify the labels.
end for
Use the test dataset with the trained model to get predicted labels.

with the sampled data is specified in the **Supplementary Material Appendix C**. We performed meticulous comparative experiments on the model's parameters to explore the most suitable parameter for the network structure. In addition, several classical and state-of-the-art methods were selected for comparison with SClusterFormer, including traditional machine learning methods, i.e., RF [34], SVM [16], classic deep learning networks, i.e., 2D-CNN [25], 3D-CNN [35], HybridSN [36], attention-based bidirectional long short-term memory network (AB-LSTM) [37] and multi-structure kernel extreme learning machine with attention fusion strategy (MSAF-KELM) [38], and transformer-based methods, i.e., SpectralFormer [29], SSFTT [30], and GAHT [32]. For all methods, we tuned the parameters to the optimum according to the original paper and performed quantitative and qualitative comparisons of the three complex HSI datasets mentioned above with the same distribution of train-test samples. Other than that, some thorough experiments such as ablation experiments, feature analysis, model fitting experiments, and sample size experiments are shown in following and **Supplementary Material Appendix D** to further investigate and analyze our method.

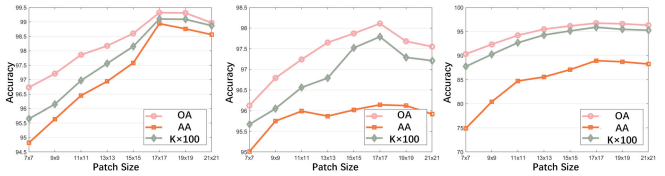
A. Experimental Setting

1) Evaluation Criteria: To precisely assess and compare the classification performance disparity between the SClusterFormer and other methods, we employ four distinct metrics for evaluating classification results: overall accuracy (OA), average accuracy (AA), kappa coefficient, and category-specific classification accuracy. Among them, the OA value is more reflective of the assessment of the overall performance. The AA values take the different sample sizes for each category into account and are more reflective of the method performance to capture sample balance. The Kappa coefficient focuses on measuring the difference between the method and randomized classification. The accuracy of each category reflects how sensitive the model performance is to the specific category. A superior value in these four metrics signifies the method's classification performance.

2) Environment Configuration: Our proposed approach was implemented in the PyTorch, while the classical traditional methods employed for comparison experiments were executed in the MATLAB R2018b environment. The computations were performed on a server equipped with an Intel Xeon Silver 4210 CPU with 128GB of RAM, complemented by an NVIDIA Geforce RTX 4090 24GB GPU. For both deep learning and transformer-based methods, we used the Adam optimizer as well as the cross-entropy loss function for the classification task. In addition, we set the learning rate to 1e-3 uniformly, and we will also specifically discuss the impact of the learning rate on our methods in Section III. To ensure sufficient GPU acceleration memory, we uniformly set the batch size to 128.

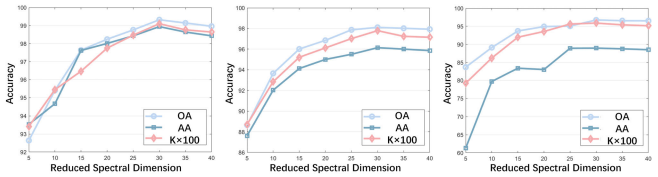
B. Parameter Analysis

In the parameter analysis section, we scrutinized and compared various types of parameters that influence the classification results. We fine-tuned these parameters to attain the optimal values. This encompasses factors such as the patch size ($s \times s$) for image partitioning, the number of reduced spectral dimensions through PCA (b), the learning rate, the number of attention heads, and the number of clusters in SpeCA and SpaCA. We set the sampling rates for the Pavia University, and WHU-Hi-HanChuan datasets to 5% and 2%, and WHU-Hi-HongHu to 100 samples per category respectively, in order to comprehensively assure the applicability



(a) (b) (c)

Fig. 6. Influence of impact on patch size for the OA, AA, and Kappa. (a) Pavia University dataset. (b) WHU-Hi-HanChuan dataset. (c) WHU-Hi-HongHu dataset.



(a) (b) (c)

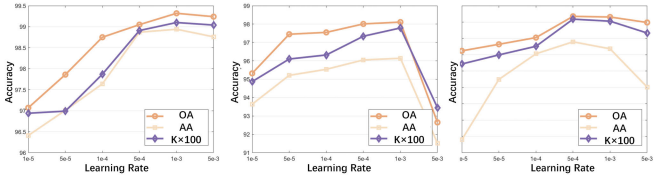
Fig. 7. Influence of impact on reduced spectral dimension for the OA, AA, and Kappa. (a) Pavia University dataset. (b) WHU-Hi-HanChuan dataset. (c) WHU-Hi-HongHu dataset.

of their parameters on data with different spatial scales and sample conditions.

1) Patch Size: Fig. 6 shows the influence of the input data's patch size on the classification accuracy. Subfigure (a) shows the results on the Pavia University dataset. It can be seen that the patch size varied with the accuracy coordinates and the overall positive correlation trend. When the patch size was 17×17, OA, AA and Kappa all increased to the highest. Subfigure (b) shows the results on the WHU-Hi-HanChuan dataset, and we find that the best results were achieved when the patch size was 17×17. Meanwhile, subfigure (c) shows the results on the WHU-Hi-HongHu dataset, which also shows that the best classification results were achieved when the patch size is 17×17. Too small a patch size can lead to a restricted receptive field for the overall network, which affects the classification performance. Also, too large a patch size will increase the computational effort and the network structure will tend to be underfitted. It can be seen that a patch size of 17×17 is the best choice to keep the subsequent higher classification performance on datasets of different scales.

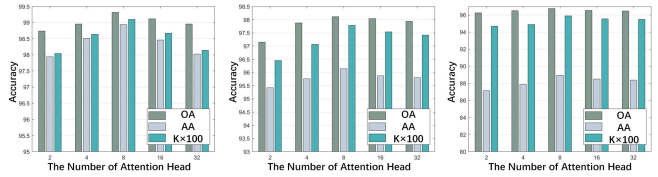
2) Reduced Spectral Dimension: We subjected the HSI data to a PCA operation before entering the network to eliminate redundant bands and reduce the amount of subsequent computation. Fig. 7 demonstrates the relationship between the number of reduced spectral dimensions and the resulting classification accuracy. All three subfigures show a positive correlation between accuracy and the number of retained dimensions up to a value of 30, and a decrease after 30. Since too few dimensions would lose significant spectral information that our proposed FSA module would not be effective, and SpeCA would not be able to capture accurate spectral information to compute the similarity in clusters. And too numerous spectral dimensions will lead to a great increase in the subsequent computation, resulting in the Hughes phenomenon. Therefore, we set the reduced spectral dimension to 30 is the most favorable for the improvement of classification accuracy.

3) Learning Rate: The learning rate, a pivotal hyperparameter in deep learning, exerts a profound influence on the convergence behavior of the objective function towards local minima. In our experimental regimen, we meticulously considered a curated set of candidate learning rates, namely $1e-5, 5e-5, 1e-4, 5e-4, 1e-3, 5e-3$. As delineated in Fig. 8, the OA, AA, and Kappa of our proposed method exhibit discernible fluctuations across three distinct datasets under varying learning rates. Evidently, Pavia University and WHU-Hi-HanChuan datasets demonstrated their highest



(a) (b) (c)

Fig. 8. Influence of impact on learning rate for the OA, AA, and Kappa. (a) Pavia University dataset. (b) WHU-Hi-HanChuan dataset. (c) WHU-Hi-HongHu dataset.



(a) (b) (c)

Fig. 9. Influence of impact on the number of attention heads for the OA, AA, and Kappa. (a) Pavia University dataset. (b) WHU-Hi-HanChuan dataset. (c) WHU-Hi-HongHu dataset.

performance on a learning rate of $1e-3$ and WHU-Hi-HongHu dataset on $5e-4$.

4) The Number of Attention Heads: When performing the cluster attention computation i.e., SpeCA and SpaCA modules, we adopt a multi-head mechanism similar to MHSA. Multiple attention computations are conducted in a block and then fused together, which effectively improves the processing of data and realizes the adequate interaction of global and local information. An insufficient number of computational heads can lead to insufficient interaction, and an excessive number of heads can lead to overattentive computation, which can lead to overfitting phenomena. Therefore, it is particularly important to select the appropriate number of heads. From Fig. 9, it can be seen that when the number of heads was 2, the AA values under the three datasets would be too low, and when the number of heads was 24, the accuracy also decreased due to over-computation. Therefore, in conjunction with the results in the figure, we set the number of heads to 8 in all subsequent experiments.

5) The Number of Clusters in SpeCA and SpaCA: There is a correlation between the number of clusters and the degree of interaction between points in the feature map when using SpeCA and SpaCA. We conducted parametric experiments on the relationship between the number of clusters and the overall accuracy. The output size of the adaptive pooling layer is

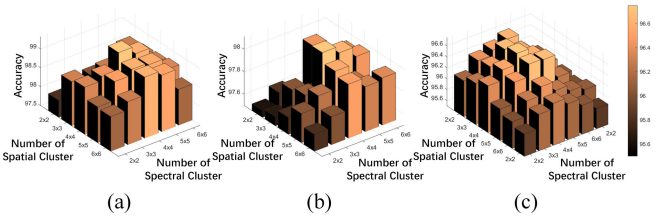


Fig. 10. Influence of impact on the number of Clusters in SpeCA and SpaCA for the OA. (a) Pavia University dataset. (b) WHU-Hi-HanChuan dataset. (c) WHU-Hi-HongHu dataset.

the centroid of the subsequent clusters, i.e., the number of clusters. As shown in Fig. 10, the number of SpeCA and SpaCA clusters were set to $2 \times 2, 3 \times 3, 4 \times 4, 5 \times 5, 6 \times 6$ under each of the three datasets, and the figure represents their relationship with z-axis accuracy. The brighter the color and the higher the bar represents the higher precision at that position. From the x-axis direction, the accuracy of 4×4 size was greater than the accuracy of the two side sizes for all three datasets. From the y-axis direction, the 4×4 size was also at its maximum. Too much clustering leads to loss of local information, while too little clustering reduces the degree of interaction of context points. Setting the appropriate number of clusters is conducive to enhancing the balance of local-global information. Therefore, we set the number of clusters to 4×4 for both SpeCA and SpaCA in the subsequent experiments to obtain optimal classification accuracy.

C. Classification Results

Subsequently, we present a comparative analysis of SClusterFormer against classical and state-of-the-art methods as mentioned previously, focusing on classification accuracy. Three distinct HSI datasets were carefully chosen, each characterized by varying spatial dimensions, spectral complexities, and sampling rates. This diverse selection allowed for a comprehensive evaluation of the strengths and weaknesses of SClusterFormer in relation to existing methodologies. To guarantee the robustness and generalizability of our experimental findings, each experiment was diligently repeated ten times. The ultimate results were determined by computing the mean and standard deviation from these repetitions.

Table I presents the results for the Paiva University dataset, with a sampling rate set at 5%. This dataset featured a moderate spatial dimension. However, it boasted a lower spatial resolution and a greater number of spectral dimensions. Even under these conditions, our proposed SClusterFormer demonstrated notable competitiveness. Thanks to the relatively ample sample size, the deep learning-based and Transformer-based methods all achieved high classification accuracy. Notably, due to the MDC module's enhanced capacity for accurately extracting convolutional features from HSI salient targets in images with lower spatial resolution, our results exhibited a significant superiority.

Secondly, II presents the results for the WHU-Hi-HanChuan dataset, utilizing a 2% sampling rate as the set of experiments. This dataset encompassed a large spatial scale with intricate spectral-spatial features, resulting in an uneven distribution at the 2% sampling rate. Notably, our method achieved superior performance across all metrics. In comparison, the 2D-CNN, 3D-CNN, and HybridSN methods lagged behind SClusterFormer by margins of 6.43%, 6.85%, and 8.56% respectively in terms of AA. This was attributed to their predominant focus on local information, which led to a reduced capacity for balancing sample processing by disregarding global information. Significantly, it is worth emphasizing that AA of SClusterFormer surpassed other Transformer-based methods by margins ranging from 1.32% to 3.57%. This underscored the efficacy of cluster attention in reinforcing HSI information locally, enabling the model to fully exploit complex spectral-spatial features and enhancing the balance of sample distribution, especially on large-scale spatial data.

Table III presents the results of the quantitative experiments conducted on the HongHu dataset at a sample size of 100 for each category. In this scenario, the number of samples was uniformly distributed, yet our method achieved the highest accuracy among all compared methods. The SpeCA module in the upper branch of SClusterFormer effectively leveraged spectral information for modeling, fully harnessing its profound feature expression capabilities. In comparison to 2D-CNN, 3D-CNN, and HybridSN methods, our approach outperformed in OA as it capitalized on both local and non-local features. While the MSAF-KELM method exhibited a higher OA due to its robust multi-feature representation, it still trailed our proposed method. The AB-LSTM method, with its limited local information representation capability, achieved a comparatively lower accuracy. Both SpectralFormer, SSFTT and GAHT methods placed significant emphasis on spectral information, leading to competitive results. However, the MHSA component in these methods neglected the balance between local and global information, resulting in slightly lower accuracy compared to our approach.

Our method shows strong classification ability in the whole, but the classification accuracy decreases when dealing with very few scenes with low contrast or obscure spectral features. For example, Romaine lettuce, category No. 17 in the WHU-Hi-HongHu dataset, is more difficult to classify in the model due to the relatively sparse and regionally concentrated distribution of this category in the full map, and the high similarity between the spectral profile of the plant and other surrounding categories. SClusterFormer in these cases is not able to sufficiently differentiate the category due to the poor quality of the spectral-spatial cluster co-features, resulting in a relatively low classification accuracy. Nevertheless, this phenomenon only occurs in a few categories and is mainly affected by the distributional characteristics of the data itself. Future enhancement of the learning strategy by targeting spectral details of similar feature types may further improve the model's performance in such scenarios.

D. Visual Evaluation

Figs. 11-13 show the classification maps of our method with the comparison methods for the three datasets mentioned above. For the Pavia University dataset, SClusterFormer completely and clearly reproduced most of the categories, whereas most of the other methods included noise. In zoomed-in local maps, methods other than ours have difficulty distinguishing

TABLE I

CLASSIFICATION PERFORMANCE OF PAVIA UNIVERSITY DATA SET OBTAINED BY DIFFERENT METHODS (THE BEST RESULTS ARE SHOWN IN BOLD)

No.	Traditional Methods		Deep Learning Networks					Transformer-based Networks			
	RF [34]	SVM [16]	2D-CNN [25]	3D-CNN [35]	HybridSN [36]	AB-LSTM [37]	MSAF-KELM [38]	SpectralFormer [29]	SSFTT [30]	GAHT [32]	SClusterFormer
1	89.68±1.24	91.26±1.24	98.04±0.49	98.85±0.58	97.41±1.68	84.11±2.51	**99.55±0.33**	97.82±0.4	97.89±0.97	96.38±3.33	98.93±0.38
2	97.17±0.45	98.39±0.99	99.88±0.08	99.98±0.03	98.97±1.41	97.08±0.76	**100.00±0.00**	99.92±0.05	99.69±0.05	99.92±0.05	99.95±0.07
3	61.29±0.64	32.29±30.74	87.59±5.09	92.21±1.05	95.62±1.75	72.16±6.4	**99.12±0.73**	92.11±0.79	89.9±3.22	97.7±0.93	97.15±1.48
4	86.88±1.83	82.65±4.28	96.32±0.47	95.6±1.49	98.93±0.32	91.33±3.49	96.12±0.73	96.9±0.27	**99.05±0.22**	94.6±2.72	97.9±0.89
5	99.14±0.11	98.96±0.24	99.91±0.11	99.95±0.09	100.0±0.0	98.35±0.76	98.31±1.92	98.79±0.48	**100.00±0.00**	96.9±2.74	99.91±0.08
6	64.5±3.81	40.93±22.29	97.99±0.61	99.23±0.74	99.91±0.12	84.17±2.76	**100.00±0.00**	99.63±0.24	99.0±0.33	99.66±0.53	99.94±0.12
7	72.93±3.01	36.07±36.11	99.25±0.5	99.6±0.23	98.72±2.1	76.44±7.54	97.86±0.94	99.52±0.24	98.81±0.55	**99.92±0.12**	99.57±0.74
8	88.15±1.48	91.47±3.9	96.11±0.87	95.0±1.92	95.1±1.41	72.38±3.38	**99.84±0.09**	91.89±1.12	93.89±2.84	98.94±1.09	98.21±1.71
9	99.52±0.37	99.6±0.27	99.5±0.35	98.63±1.14	98.89±1.78	87.36±8.9	73.75±5.40	94.75±1.0	**99.64±0.29**	93.59±3.12	98.91±0.78
OA(%)	88.25±0.45	83.67±4.67	98.16±0.19	98.55±0.22	98.36±0.90	88.97±1.12	98.89±0.25	98.11±0.12	98.29±0.24	98.53±0.70	**99.32±0.17**
AA(%)	84.36±1.04	74.62±9.72	97.18±0.47	97.67±0.15	98.17±0.61	84.82±2.22	96.06±0.91	96.82±0.20	97.54±0.29	97.51±0.70	**98.94±0.31**
k×100	84.17±0.66	77.41±6.86	97.56±0.26	98.08±0.29	97.83±1.18	85.36±1.50	98.53±0.33	97.50±0.16	97.73±0.32	98.06±0.92	**99.10±0.23**

TABLE II

CLASSIFICATION PERFORMANCE OF WHU-HI-HANCHUAN DATA SET OBTAINED BY DIFFERENT METHODS (THE BEST RESULTS ARE SHOWN IN BOLD)

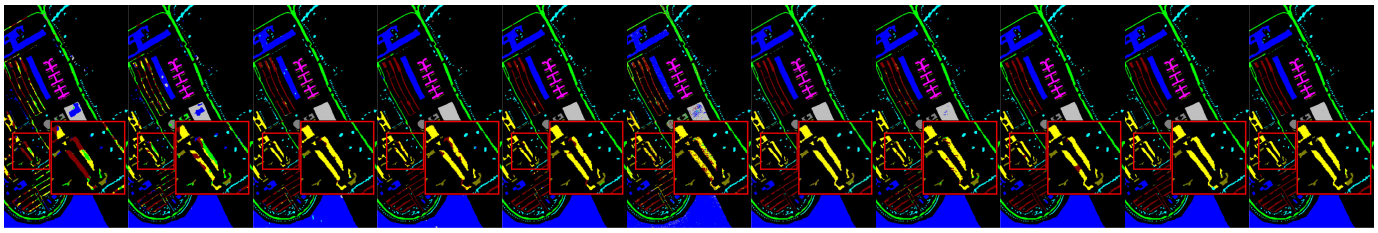
No.	Traditional Methods		Deep Learning Networks					Transformer-based Networks			
	RF [34]	SVM [16]	2D-CNN [25]	3D-CNN [35]	HybridSN [36]	AB-LSTM [37]	MSAF-KELM [38]	SpectralFormer [29]	SSFTT [30]	GAHT [32]	SClusterFormer
1	94.63±0.43	94.05±0.39	97.86±0.55	97.0±0.76	96.95±1.97	92.08±2.49	99.06±0.49	98.14±1.46	99.29±0.32	99.17±0.92	**99.66±0.14**
2	72.54±0.56	72.48±0.64	95.05±1.11	94.62±0.72	95.07±1.01	81.36±1.9	95.09±2.14	95.7±1.93	96.07±0.82	94.68±2.02	**97.45±1.32**
3	72.15±0.59	65.85±8.6	91.5±3.25	91.69±3.49	90.23±4.02	72.95±1.9	96.97±1.84	95.97±1.8	99.04±0.78	96.98±2.02	**99.34±0.38**
4	82.99±2.25	76.59±11.25	99.37±0.27	98.94±0.29	**98.93±0.59**	89.64±1.5	98.5±0.95	98.59±0.93	98.54±1.35	98.65±1.12	98.87±0.99
5	5.39±1.33	3.44±2.76	77.72±7.37	75.87±10.49	66.26±20.04	30.19±5.44	87.53±7.74	84.61±15.14	**95.0±2.34**	93.47±3.39	91.82±11.59
6	4.77±0.31	3.16±2.26	72.76±6.64	71.03±5.25	68.69±3.04	39.37±4.68	86.5±5.36	83.95±0.84	83.56±6.57	84.74±4.29	**90.38±4.14**
7	80.84±1.64	83.06±3.92	88.42±3.68	90.13±1.54	86.37±4.93	64.78±3.85	90.84±1.51	93.07±1.39	**95.58±2.28**	94.65±2.23	94.32±1.41
8	67.31±0.87	66.72±1.1	94.26±0.71	91.16±4.02	92.5±2.04	72.85±2.36	96.55±1.19	95.22±1.59	**97.0±0.87**	94.88±1.28	96.64±0.69
9	49.15±2.11	34.13±23.25	94.25±1.54	95.2±1.46	90.89±1.71	58.59±5.56	94.6±3.38	94.85±0.77	96.36±1.18	93.26±1.79	**96.97±1.21**
10	82.42±1.37	73.05±14.32	98.34±1.26	98.08±0.77	98.23±1.09	93.65±2.46	97.89±0.77	98.59±0.65	98.27±0.7	97.8±1.06	**98.59±0.54**
11	87.72±0.91	71.96±21.96	97.73±0.44	97.39±1.09	96.63±2.31	90.47±1.59	98.32±0.35	97.17±1.09	98.76±0.66	98.18±1.3	**98.87±0.32**
12	18.48±0.81	12.61±8.93	74.65±10.69	78.73±7.76	75.13±6.19	30.07±4.31	91.92±3.43	85.64±8.15	88.76±15.1	94.61±1.48	**99.12±0.36**
13	46.86±1.69	37.06±14.98	79.77±3.35	75.84±3.73	73.45±3.9	47.84±2.56	82.55±4.03	83.23±4.42	80.98±9.31	86.03±4.47	**87.86±1.54**
14	80.66±1.38	79.95±2.34	95.02±0.92	93.82±0.59	90.31±4.63	80.66±1.78	96.66±0.96	94.76±2.22	97.07±0.64	96.5±1.52	**97.1±0.87**
15	57.51±1.77	55.11±5.17	79.05±3.29	79.46±3.3	82.35±2.26	44.78±7.37	67.13±33.6	81.87±2.2	**93.08±1.64**	82.25±7.05	91.41±1.46
16	98.34±0.13	98.72±0.5	99.53±0.14	99.66±0.24	99.32±0.28	98.85±0.22	99.66±0.18	99.69±0.25	99.79±0.07	**99.86±0.06**	99.83±0.12
OA(%)	81.83±0.24	78.96±4.29	95.50±0.47	94.96±0.36	94.18±0.83	84.36±1.00	96.89±0.21	96.48±1.33	97.41±0.49	97.06±0.36	**98.11±0.09**
AA(%)	62.61±0.53	58±7.05	89.71±1.73	89.29±0.64	87.58±1.93	68.01±1.39	92.49±2.15	92.57±2.64	94.82±1.32	94.11±0.56	**96.14±0.74**
k×100	78.61±0.29	75.08±5.28	94.73±0.55	94.11±0.42	93.19±0.98	81.67±1.17	96.36±0.24	95.88±1.55	96.97±0.57	96.55±0.42	**97.79±0.11**

TABLE III

CLASSIFICATION PERFORMANCE OF WHU-HI-HONGHU DATA SET OBTAINED BY DIFFERENT METHODS (THE BEST RESULTS ARE SHOWN IN BOLD)

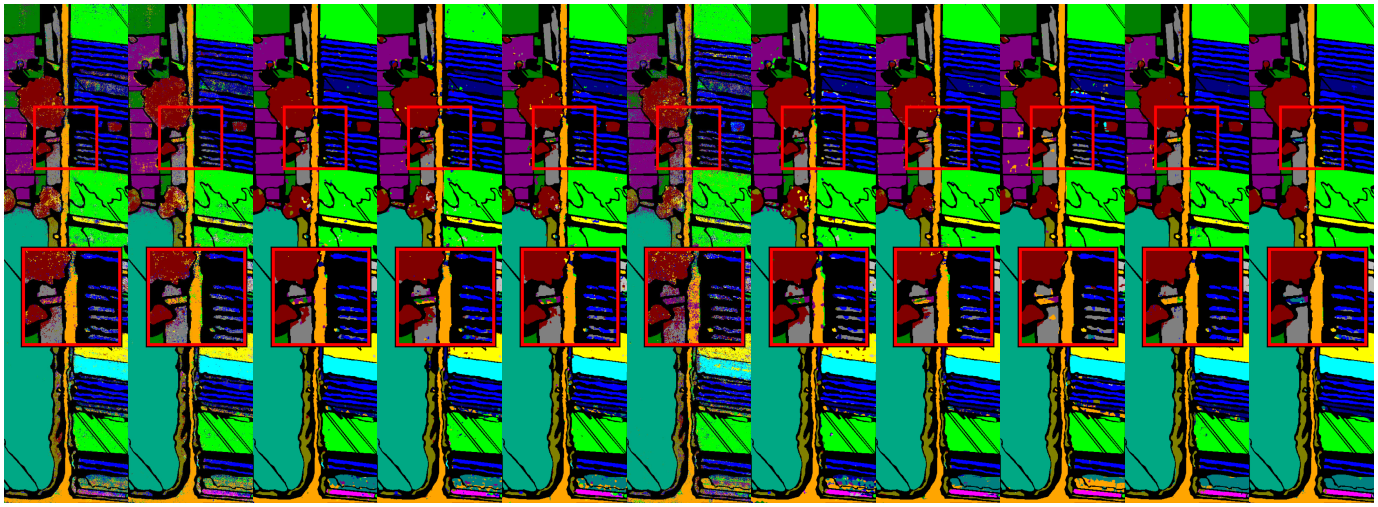
No.	Traditional Methods		Deep Learning Networks					Transformer-based Networks			
	RF [34]	SVM [16]	2D-CNN [25]	3D-CNN [35]	HybridSN [36]	AB-LSTM [37]	MSAF-KELM [38]	SpectralFormer [29]	SSFTT [30]	GAHT [32]	SClusterFormer
1	82.42±7.09	78.68±7.57	95.39±0.34	96.5±0.7	96.43±0.32	64.04±4.88	98.67±1.25	96.68±0.08	97.7±0.27	97.38±0.49	**98.8±0.36**
2	39.66±7.97	39.13±6.13	78.39±7.39	72.66±6.13	70.89±3.99	37.79±6.16	38.0±7.24	70.56±3.55	**78.54±3.15**	53.88±3.73	71.09±13.14
3	89.09±3.0	89.89±3.65	93.62±2.25	94.81±1.35	94.52±1.03	80.36±3.72	**96.47±3.45**	95.65±1.12	95.0±2.99	95.99±1.08	94.64±2.88
4	97.69±0.86	97.81±0.58	99.65±0.18	99.11±0.47	99.71±0.13	93.58±0.82	99.71±0.47	99.77±0.03	99.71±0.16	99.46±0.26	**99.94±0.09**
5	16.67±4.43	23.75±5.51	77.26±5.1	71.43±10.9	64.64±8.06	30.5±4.29	**90.31±3.45**	87.36±2.31	77.7±14.0	83.16±9.59	90.01±1.75
6	90.65±2.9	92.2±2.3	97.48±1.21	97.41±0.97	98.42±0.53	79.58±3.6	98.44±2.34	97.5±0.3	97.67±3.2	98.21±0.59	**99.24±0.18**
7	79.68±5.6	85.59±4.89	91.64±0.95	91.06±1.38	88.14±1.64	48.47±3.25	95.19±5.45	93.97±0.54	93.13±2.25	93.29±2.66	**96.92±1.35**
8	26.18±9.15	24.85±6.94	69.23±5.21	57.95±6.16	64.75±7.99	9.59±4.11	**92.18±5.78**	71.08±5.53	71.47±11.83	69.53±5.89	91.85±5.47
9	92.59±0.64	93.68±0.93	96.39±0.6	96.36±1.03	96.24±1.03	80.21±7.05	97.07±2.45	97.82±0.84	**97.98±0.58**	97.49±1.41	96.77±1.41
10	33.67±14.05	53.58±5.05	86.06±1.63	78.29±5.2	80.28±3.12	35.28±5.34	85.58±6.89	84.34±2.42	92.61±1.39	92.37±2.35	**94.8±2.93**
11	46.07±9.79	41.41±8.49	80.89±2.55	77.57±3.21	76.56±2.37	27.41±2.85	**94.38±3.84**	85.01±2.84	85.64±3.72	86.82±4.25	93.74±2.93
12	33.0±3.96	29.85±5.24	75.41±3.01	70.39±9.29	72.06±9.54	29.83±6.19	86.41±9.98	79.02±1.13	81.82±8.12	80.43±5.49	**89.0±4.11**
13	55.76±10.9	58.66±1.38	87.34±1.09	82.7±3.4	82.43±2.55	39.93±4.15	89.15±2.57	84.21±0.94	88.79±1.31	88.04±2.02	**93.25±2.44**
14	65.13±9.05	68.95±3.69	81.82±7.61	89.02±3.2	83.46±6.8	26.55±6.75	94.41±6.73	86.01±1.97	86.62±9.0	84.5±4.62	**95.09±1.31**
15	24.98±12.3	18.72±11.72	**88.57±4.87**	81.37±4.89	78.39±14.75	13.43±7.6	42.12±29.57	84.0±5.07	85.4±3.92	76.53±14.13	31.81±38.97
16	86.6±3.28	83.83±5.48	94.02±1.21	92.52±2.08	91.9±2.05	55.97±16.06	91.74±1.24	93.22±0.87	91.74±4.67	94.28±4.83	**99.32±0.52**
17	33.16±17.01	74.43±28.63	94.5±2.17	89.57±6.49	96.06±0.88	41.2±10.31	12.81±24.37	95.27±2.54	**98.75±0.37**	77.71±31.94	59.97±32.02
18	37.68±14.42	62.04±15.1	91.33±4.95	92.19±4.02	92.89±0.99	13.07±6.88	37.14±4.37	92.07±1.65	92.93±2.63	**93.45±2.93**	90.55±10.2
19	76.05±5.26	80.96±3.85	89.97±2.89	84.51±5.79	85.32±8.07	48.76±1.43	92.16±2.93	90.77±0.98	**93.04±1.87**	86.27±2.35	91.73±2.65
20	14.81±6.71	21.88±7.73	66.02±9.04	77.96±5.85	70.72±7.61	20.54±7.61	21.68±34.24	79.8±3.29	73.35±13.37	85.84±5.02	**90.73±3.05**
21	6.89±1.07	9.86±3.81	48.35±11.58	60.92±14.35	42.44±14.82	10.11±1.61	13.43±4.73	61.15±4.73	64.19±17.62	53.89±14.73	**90.59±7.6**
22	32.34±11.57	37.81±10.81	85.41±5.89	78.12±10.89	79.9±5.07	17.43±3.77	29.52±3.94	94.16±2.13	90.47±8.12	83.74±19.56	**97.01±2.68**
OA(%)	80.29±0.96	82.29±1.47	93.76±1.40	92.61±1.24	92.63±1.56	70.19±1.66	92.57±2.36	94.58±1.19	95.05±1.82	94.52±0.97	**96.75±0.69**
AA(%)	52.76±2.44	57.62±4.37	84.94±2.47	83.29±2.54	82.10±2.75	41.07±2.69	67.42±6.51	87.25±1.92	87.92±4.24	85.10±3.04	**88.95±2.33**
k×100	74.94±1.22	77.45±1.95	92.10±2.54	90.64±1.57	90.74±1.72	62.45±1.42	90.57±2.61	93.14±1.25	93.74±2.30	93.07±1.20	**95.89±0.87**

between Class 3 Gravel and Class 9 Shadows. Shadows are in dark-light conditions and thus have weaker spectral features. In contrast, SClusterFormer fully utilizes the features of HSI spectral-spatial correlation, and the classification map captures more accurate and smooth Shadows edges. For the large-scale dataset HanChuan, our method had the fewest noise



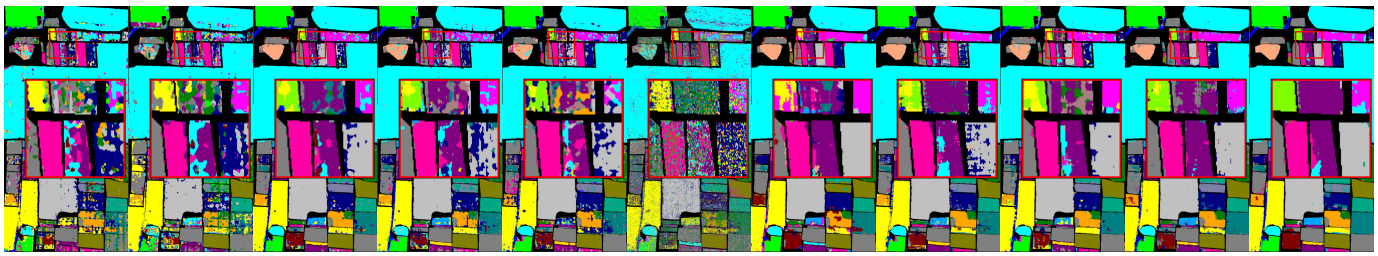
(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 11. Classification maps for Pavia University data set under different methods. (a) Random Forest. (b) SVM. (c) 2D-CNN. (d) 3D-CNN. (e) HybridSN. (f) AB-LSTM. (g) MSAF-KELM. (h) SpectralFormer. (i) SSFTT. (j) GAHT. (k) SClusterFormer.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 12. Classification maps for WHU-Hi-HanChuan data set under different methods. (a) Random Forest. (b) SVM. (c) 2D-CNN. (d) 3D-CNN. (e) HybridSN. (f) AB-LSTM. (g) MSAF-KELM. (h) SpectralFormer. (i) SSFTT. (j) GAHT. (k) SClusterFormer.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 13. Classification maps for WHU-Hi-HongHu data set under different methods. (a) Random Forest. (b) SVM. (c) 2D-CNN. (d) 3D-CNN. (e) HybridSN. (f) AB-LSTM. (g) MSAF-KELM. (h) SpectralFormer. (i) SSFTT. (j) GAHT. (k) SClusterFormer.

points and handled the boundary regions of the categories well. In the zoomed-in localized map, the spectral values of class 6 Watermelon and class 7 Greens may be in the same interval. Most of the methods are difficult to distinguish them by the band curves, thus a certain amount of class 6 noise is introduced within class 7. SClusterFormer benefits from an in-depth combinatorial analysis of the high and low-frequency information of the spectral bands, which accurately grasps the differences in the trends of the spectral curves, and therefore obtains classification maps that are closer to Ground Truth. For the HongHu dataset, our method was the closest to groundtruth, especially in a few of the larger categories on the left. These illustrate the effective context modeling capability of SClusterFormer. In the zoomed-in localized image, the upper class 11 Brassica parachinensis is difficult for most methods to learn its accurate discriminative features due to the low proportion of training samples relative to testing samples. Whereas, SClusterFormer gains clean and accurate classification results thanks to the balance of global-local information, which better distinguishes the spectral differences from the surrounding similar fields.

Despite SClusterFormer's minimal noise, misclassification mainly occurs at the edges of smaller area categories. This suggests the model's performance is impacted in boundary regions, especially where neighboring categories have similar spectral features. The complex transition spectral information at these edges introduces some perturbations into SClusterFormer's spectral clusters. Improving boundary processing or

TABLE IV
ABLATION EXPERIMENTAL RESULTS

Cases	Component					Indicators		
	MDC	FSA	SpeCA	SpaCA	CFPF	OA(%)	AA(%)	$k * 100$
1	×	✓	✓	✓	✓	97.04	96.38	96.70
2	✓	×	✓	✓	✓	98.67	97.09	98.53
3	✓	✓	×	✓	✓	98.76	98.64	98.64
4	✓	✓	✓	×	✓	98.77	98.27	98.61
5	✓	✓	✓	✓	×	98.32	96.95	98.10
6	×	×	✓	✓	✓	96.85	97.02	96.40
7	✓	✓	×	×	✓	98.04	97.70	98.06
8	✓	✓	✓	✓	✓	**99.32**	**98.94**	**99.10**

refining edge feature extraction could enhance accuracy in these areas.

In conclusion, the comparison of several state-of-the-art classifiers in visual evaluation still validates that the proposed SClusterFormer has excellent classification performance.

E. Ablation Study

1) Module-Wise Contribution Analysis: To meticulously elucidate the distinct contributions of each constituent element within the proposed SClusterFormer towards enhanced classification performance, we performed an ablation experiment employing a 5% sample rate on the Pavia University dataset. The framework was deconstructed into five discernible sections: MDC, FSA, SpeCA, SpaCA, and CFPF. Subsequently, a comprehensive evaluation was conducted utilizing performance metrics including OA, AA, and the Kappa coefficient. The outcomes of these ablation experiments are meticulously tabulated in Table IV for reference and analysis.

In Case 1, the DCN module was omitted, and the original HSI sliced patch features were directly fed into the network for classification. Notably, no deformable convolution was employed for processing salient features. As a consequence, the feature representation capability was diminished, leading to lower scores in the three classification accuracy metrics when compared to the performance achieved by the proposed SClusterFormer. This underscored the pivotal role of the deformable convolution module in enhancing the ultimate classification accuracy.

In Case 2, the FSA module was removed, resulting in the absence of spectral information weighting through frequency analysis. This caused a decrease in the expressiveness of the spectral information compared to our proposed method. As expected, all three classification accuracy metrics displayed lower values than those obtained with SClusterFormer. This highlights the importance of frequency-based spectral attention in preserving both high and low-frequency representation features, thereby enhancing the spectral modeling capacity of the subsequent SpeCA module.

In Case 3, the SpeCA module in the upper branch was removed in favor of MHSA. This substitution, while computing attention for all flat features, neglects to address the nuanced spectral information preserved in earlier. Consequently, in this instance, all three classification accuracy metrics experienced a decline compared to those achieved by the proposed SClusterFormer. This illustrates the upper branch's effectiveness in facilitating contextual information exchange through spectral cluster, thereby revitalizing the local information dynamics within the Transformer structure.

Case 4 removed the SpaCA module from the lower branch and implemented MHSA instead. Clearly, all three classification metrics registered lower values compared to those achieved by the proposed SClusterFormer. This demonstrates that the enhancement of contextual information through cluster attention is equally pertinent at the spatial feature level. Furthermore, it reaffirms our method's capacity to rectify the Transformer model's tendency to overlook local information, thereby bolstering the equilibrium between local and global considerations in HSI analysis.

In Case 5, the CFPF module was removed, leading to the direct classification of the junction output features from the two branches after summation. The three resulting metrics were also observed to be lower compared to those achieved by SClusterFormer. This observation indicates that the high-level and abstract classification features are more apt for pixel-level decision fusion. This, in turn, serves as a validation for our proposed CFPF module, which aims to maximize the preservation of the benefits derived from our dual-branch structure.

Case 6 removes the modules MDC and FSA from the feature extraction phase altogether, and the raw image patches will be fed into the Transformer for direct cluster attention computation. This configuration results in a 2.47% decrease in OA and a 1.92% decrease in AA, which emphasizes the importance of MDC and FSA in capturing critical information in large size spaces and high dimensional spectral trends. It also further illustrates that the direct application of unprocessed HSI cubes to the downstream network would introduce substantial feature errors.

Case 7 completely isolates the SpeCA and SpaCA modules and replaces them with the regular MHSA in both the upper and lower branches. Compared with our model, OA decreases by 1.28%, AA decreases by 1.24%, and $k \times 100$ decreases by 1.04. This demonstrates that the global-local balance and spectral-spatial balance adopted by the spectral-spatial cluster attention can better adapt to the data of the complex and large-scale HSI remote sensing scenarios compared with the MHSA, which plays an essential role in the whole architecture.

2) Comparative Analysis of Attention Mechanism: Meanwhile, we conducted experiments on the WHU-Hi-HanChuan dataset with 2% sampling rate to compare the effects of our proposed cluster attention and MHSA on the classification effect. We also performed experiments on the upper branch separately to verify the validity of the two-branch results. As shown in Table V, we replaced the attention of SClusterFormer with MHSA, and we can see that the accuracy of the method using cluster attention was better than that of MHSA for both the two-branch structure and the single-branch structure. In addition, we find that the results of the two-branch structure are all higher than those of the single-branch structure, which justifies the two-branch results adopted by our SClusterFormer.

Table V equally demonstrates the running time and model size of the network using MHSA and cluster-based attention.

TABLE V
ATTENTION ABLATION EXPERIMENTAL RESULTS(OPTIMAL RESULTS ARE BOLDED)

Methods	Accuracy Indicators						
	OA(%)	AA(%)	$k * 100$	Train Time (s)	Test Time (s)	Parameters (M)	FLOPS (G)
MHSA-Transformer-Upper Branch	97.16±0.71	94.31±0.42	96.15±0.47	113.82	1.37	1.09	20.15
MHSA-Transformer	97.42±0.50	94.92±0.34	96.25±0.32	145.42	1.64	2.48	38.81
SClusterFormer-Upper Branch	97.32±0.59	95.21±0.76	96.26±0.46	111.65	1.32	0.75	12.71
SClusterFormer	**98.11±0.09**	**96.14±0.74**	**97.79±0.11**	131.12	1.41	1.67	24.26

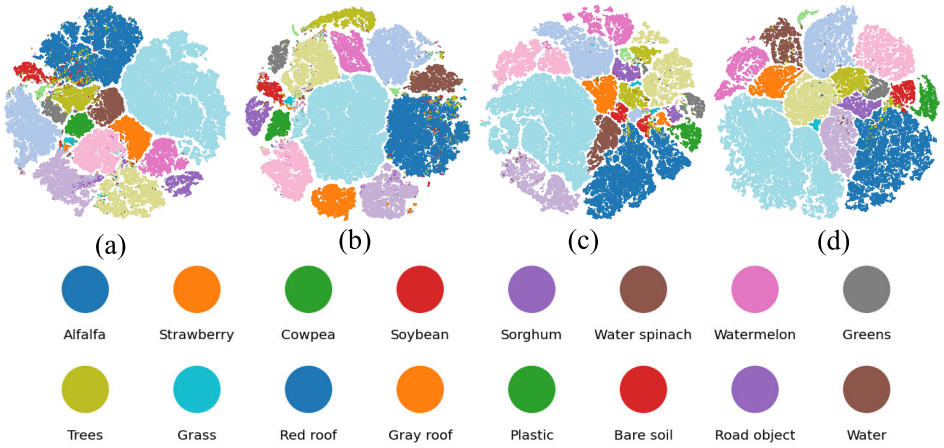


Fig. 14. Demonstration of the impact of the attention module and branching structure in SClusterFormer on the WHU-Hi-HanChuan dataset. (a) T-SNE visualization of MHSA-Transformer-Upper Branch. (b) T-SNE visualization of MHSA-Transformer. (c) T-SNE visualization of SClusterFormer-Upper Branch. (d) T-SNE visualization of SClusterFormer.

TABLE VI
THE ACCURACY AND PARAMETER COMPLEXITY OF SEVERAL BASELINE METHODS IN THE WHU-HI-HANCHUAN DATASET

Indicators	Baseline Methods							
	3D-CNN	DCN	ViT	SA-ViT	HT	SA-HT	SClusterFormer-Upper Branch	SClusterFormer
OA(%)	93.67±0.21	93.78±0.26	92.17±0.56	93.53±0.38	95.82±0.21	95.87±0.18	97.32±0.29	**98.11±0.09**
AA(%)	85.90±0.73	90.01±0.70	81.96±1.37	82.73±0.53	91.82±0.87	92.12±0.81	95.21±0.76	**96.14±0.74**
k×100	92.59±0.24	93.58±0.30	90.82±0.65	91.43±0.45	94.28±0.24	94.43±0.21	96.26±0.26	**97.79±0.11**
Parameters (M)	1.55	1.14	0.63	0.65	1.65	1.67	0.75	1.67
FLOPs (G)	7.96	0.55	1.35	1.41	24.57	24.61	12.71	24.26

We can learn that the training and testing time using the cluster-based attention approach were lower than the MHSA method for both two-branch and one-branch networks. In addition, the number of parameters and floating point operations (FLOPs) decreased by 32.67% and 37.5% compared to the MHSA method, which indicates that our proposed cluster-based attention reduces the running cost of the model while guaranteeing the classification performance.

As shown in Fig. 14, we also performed t-stochastic neighbor embedding (t-SNE) visualization operation on the output features of the above experiments. Each point in the figure represents a sample, and different colors indicate different categories. Again, we can see that the features obtained by cluster attention are much clearer than those of MHSA, and there were fewer cases of overlapping coverage of sample categories. Similarly, the feature maps of the two-branch results are all better than the single-branch structure, which again proves our point above.

F. Comparison of Baseline Methods

To better demonstrate the trade-off between performance improvement and model complexity, we designed a series of baseline comparison experiments. We compared SClusterFormer with several simpler models, including three-dimensional convolutional neural networks (3D-CNN), deformable convolutional networks (DCN), vision Transformers (ViT), vision Transformers with spectral attention (SA-ViT), hierarchical Transformers (HT), and hierarchical Transformers with spectral attention (SA-HT). These methods were used to compare the classification accuracy with the computational cost on the WHU-Hi-HanChuan dataset, where a random training sample rate of 2% was set. The comparison offers us a clearer perspective on how our model achieves a balance between complexity and performance gains, especially in scenarios with limited computational resources.

The results demonstrate that our proposed model, SClusterFormer, achieves the highest performance in terms of OA, AA, and kappa, surpassing all baseline methods. Despite having a larger parameter count and higher computational cost (FLOPs), SClusterFormer offers a significant improvement in classification accuracy compared to simpler models like 3D-CNN and vision Transformers. Additionally, the SClusterFormer-Upper Branch model, which reduces complexity, still maintains competitive accuracy while balancing resource usage. This highlights the effectiveness of our model's design in improving performance while managing computational demands. This further demonstrates that

the impressive performance of SClusterFormer does not come from a simple combination of a hierarchical Transformer and a deformable convolutional network. The accurate modeling of spectral-spatial cluster attention achieves a significant improvement in classification accuracy with limited computational effort.

IV. Conclusion

In this paper, a novel deformable convolution-enhanced hierarchical Transformer with spectral-spatial cluster attention (termed SClusterFormer) is proposed for effective HSI classification. The method extracts HSI and EMAP multiscale deformable convolution features through a two-branch network, and designs a frequency-based spectral attention module for spectral information enhancement. The spectral and spatial cluster attention modules are designed to enhance the aggregation processing of local details and discriminative distinction of global structures in the Transformer. Finally, a cross-feature pixel-level fusion strategy is used to achieve two-branch fusion. On three datasets with large spatial scales and numerous spectral bands, SClusterFormer was compared with a range of classical and state-of-the-art methods, achieving the highest accuracy and delivering optimal classification maps. This success is attributed to the fine-grained analysis of spectral band frequencies and the balanced integration of global and local information through spectral-spatial clusters.

In the future, we will consider extending the two-branch network to a multi-source branch network. It will be combined with LiDAR data to overcome the current challenges posed by spectral similarity in similar categories (e.g., the same genus of greenery or roads) and transitory perturbation spectra in edge regions.

References

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

[2] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2017.

[3] M. A. Calin, S. V. Parasca, D. Savastru, and D. Manea, "Hyperspectral imaging in the medical field: Present and future," *Appl. Spectrosc. Rev.*, vol. 49, no. 6, pp. 435–447, Aug. 2014.

[4] M. Pallocci et al., "Forensic applications of hyperspectral imaging technique: A narrative review," *Medico-Legal J.*, vol. 90, no. 4, pp. 216–220, Dec. 2022.

[5] C. K. Muro, K. C. Doty, J. Bueno, L. Halámková, and I. K. Lednev, "Vibrational spectroscopy: Recent developments to revolutionize forensic science," *Anal. Chem.*, vol. 87, no. 1, pp. 306–327, Jan. 2015.

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

[7] L. Fang, Y. Jiang, Y. Yan, J. Yue, and Y. Deng, "Hyperspectral image instance segmentation using spectral–spatial feature pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5502613.

[8] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2018.

[9] M. Lv, W. Li, T. Chen, J. Zhou, and R. Tao, "Discriminant tensor-based manifold embedding for medical hyperspectral imagery," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 9, pp. 3517–3528, Sep. 2021.

[10] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

[11] J. Yue, L. Fang, and M. He, "Spectral–spatial latent reconstruction for open-set hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 5227–5241, 2022.

[12] L. Wu, L. Fang, J. Yue, B. Zhang, P. Ghamisi, and M. He, "Deep bilateral filtering network for point-supervised semantic segmentation in remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 7419–7434, 2022.

[13] J. Liu, Z. Wu, L. Xiao, J. Sun, and H. Yan, "A truncated matrix decomposition for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 29, pp. 8028–8042, 2020.

[14] X. Ou, M. Wu, B. Tu, G. Zhang, and W. Li, "Multi-objective unsupervised band selection method for hyperspectral images classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1952–1965, 2023.

[15] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

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

[17] W. Huang, Y. Huang, H. Wang, Y. Liu, and H. J. Shim, "Local binary patterns and superpixel-based multiple kernels for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4550–4563, 2020.

[18] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[19] M. Dalla Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 542–546, May 2011.

[20] H. Su, Y. Yu, Z. Wu, and Q. Du, "Random subspace-based k-nearest class collaborative representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6840–6853, Aug. 2021.

[21] F. Tong and Y. Zhang, "Spectral–spatial and cascaded multilayer random forests for tree species classification in airborne hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4411711.

[22] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6308–6325, 2020.

[23] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[24] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, Jan. 2015.

[25] 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.

[26] 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.

[27] U. A. Bhatti et al., "Local similarity-based spatial–spectral fusion hyperspectral image classification with deep CNN and Gabor filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514215.

[28] X. Ma et al., "A lightweight hybrid convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513714.

[29] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.

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

[31] Y. Fang, Q. Ye, L. Sun, Y. Zheng, and Z. Wu, "Multiattention joint convolution feature representation with lightweight transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513814.
[32] 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.
[33] Y. Xu et al., "Spatial–Spectral 1DSwin transformer with groupwise feature tokenization for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5516616.
[34] J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh, "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 492–501, Mar. 2005.
[35] 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.
[36] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3D–2D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2019.
[37] S. Mei, X. Li, X. Liu, H. Cai, and Q. Du, "Hyperspectral image classification using attention-based bidirectional long short-term memory network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5509612.
[38] L. Sun, Y. Fang, Y. Chen, W. Huang, Z. Wu, and B. Jeon, "Multistructure KELM with attention fusion strategy for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539217.

Yu Fang was born in Jiangsu, China, in 1998. He received the M.E. degree in computer technology from Nanjing University of Information Science and Technology (NUIST), Nanjing, Jiangsu, China, in 2024. He is currently pursuing the Ph.D. degree with Nanjing University of Science and Technology (NJUST), Nanjing. His research interests include hyperspectral image processing and deep learning.

Le Sun (Senior Member, IEEE) was born in Jiangsu, China, in 1987. He received the B.S. degree in information and computing science and the Ph.D. degree in computer science from Nanjing University of Science and Technology (NJUST), Nanjing, Jiangsu, China, in 2009 and in 2014, respectively. From 2015 to 2018, he was a Postdoctoral Researcher with the School of Electronic and Electrical Engineering, Sungkyunkwan University, Seoul, South Korea. He is currently a Full Professor with the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing. His research interests include hyperspectral image processing (including unmixing, classification, and restoration), sparse representation, compressive sensing, and deep learning.

Yuhui Zheng (Member, IEEE) was born in Shanxi, China, in 1982. He received the B.S. degree in chemistry and the Ph.D. degree in computer science from Nanjing University of Science and Technology (NJUST), Nanjing, Jiangsu, in 2004 and 2009, respectively. From 2014 to 2015, he was a Visiting Scholar with the Digital Media Laboratory, School of Electronic and Electrical Engineering, Sungkyunkwan University, Seoul, South Korea. He is currently a Full Professor with the School of Computer and Science, Nanjing University of Information Science and Technology, Nanjing. His research interests include image processing, pattern recognition, and remote-sensing information systems.

Zebin Wu (Senior Member, IEEE) was born in Zhejiang, China, in 1981. He received the B.S. and Ph.D. degrees from the School of Computer Science and Engineering, Nanjing University of Science and Technology (NJUST), Nanjing, Jiangsu, China, in 2003 and 2008, respectively. He is currently a Full Professor with the School of Computer Science and Engineering, NJUST. His research interests include virtual reality and system simulation, remote sensing information processing, and distributed computing.