Main–Sub Transformer With Spectral–Spatial Separable Convolution for Hyperspectral Image Classification

Jingpeng Gao, *Member, IEEE*, Xiangyu Ji, *Graduate Student Member, IEEE*, Geng Chen, and Ruitong Guo

***Abstract*—Due to their spatial and spectral information, hyperspectral images are frequently used in various scientific and industrial fields. Recent developments in hyperspectral image classification have revolved around the use of convolutional neural networks and transformers, which are capable of modeling local and global data. However, most of the backbone networks of existing methods are based on 3-D convolution, which has high complexity in network structure. Moreover, local information and global information are extracted through different modules, and the coupling relationship between the two types of information is weak. To address the above-mentioned issues, we propose a method named main–sub transformer network with spectral–spatial separable convolution method (MST-SSSNet), which includes two key modules: the spectral–spatial separable convolution (SSSC) module and the main–sub transformer encoder (MST) module. The SSSC module uses the proposed spectral–spatial separable convolution, reducing network parameters and efficiently extracting local features. The MST module adds the designed subtransformer in front of the conventional transformer encoder (main transformer). It assists the main-transformer encoder to establish global correlation by learning local information. The WHU-Hi dataset can be used as a benchmark dataset for precise crop classification and hyperspectral image classification research. MST-SSSNet is shown to deliver better classification performance than current state-of-the-art methods on the datasets.**

***Index Terms*—Convolution neural networks (CNNs), hyperspectral image (HSI) classification, main–sub transformer encoder (MST), spectral–spatial separable convolution (SSSC).**

I. Introduction

HYPERSPECTRAL images (HSI) contain rich information [1], which has two spatial dimensions and one spectral dimension. Compared with ordinary RGB images, it has richer spectral information [2] and can be used in precision agriculture [3], modern medical detection [4], military security [5], and other fields [6], [7], [8]. The process of HSI classification encompasses the identification of each pixel within a scene and its categorization into predefined classes [9]. HSI classification is a basic technology in HSI processing and has been a hot research topic in the remote sensing field [10], [11], [12].

Traditional HSI classification techniques typically leverage the abundant spectral features present in HSI, including support vector machine (SVM) [7], [14], k-nearest neighbor method [15], [16], and other methods [17], [18], [19]. In [20], a method considered spatial information for HSI classification. The research by Sun et al. [21] introduced a multiscale spectral–spatial kernel approach, employing adjacent superpixels. This method takes into account both the spectral and spatial information of the data, further improving the classification performance. Although the above-mentioned methods achieved good results at the time, most of them relied on manually designed features and were shallow models that could not extract high-level features of HSI images. This limits the improvement of classification performance.

The emergence of deep learning techniques has brought considerable focus to the utilization of deep models in the realm of HSI classification [22]. Stacked autoencoder and deep belief networks were used as traditional depth models in HSI classification [23], [24]. Nevertheless, their approach involved the utilization of numerous fully connected layers, each containing a substantial volume of trainable weights. The labeling cost of HSI is high, and insufficient data to train the model means that it often leads to serious overfitting problems.

However, convolutional neural networks (CNNs) based methods have the disadvantage of weak modeling ability of global feature dependencies, which hinders further improvement of classification performance.

In recent times, CNNs have become a focal point of attention in numerous disciplines, primarily due to their outstanding ability to model local features [25], [26], [27], and have also been explored a lot in HSI classification. 1D-CNN [28], 2D-CNN [29], and 3D-CNN [30] were all explored in HSI classification. Roy et al. [31] combined 3-D convolutional layers and 2-D convolutional layers to enhance spatial feature extraction and proposed HybridSN. However, CNN-based methods have the disadvantage of weak modeling ability of global feature dependencies, which hinders further improvement of classification performance.

Beyond the scope of CNNs, researchers have delved into numerous alternative network architectures for HSI classification. Mou et al. [32] first used RNN for HSI classification. Graph

Manuscript received 22 October 2023; revised 24 November 2023, 5 December 2023, and 7 December 2023; accepted 11 December 2023. Date of publication 14 December 2023; date of current version 10 January 2024. This work was supported by the Basic Research Project Group Project under Grant KY10800220035. *(Corresponding author: Xiangyu Ji.)*

The authors are with the College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China (e-mail: gaojingpeng@hrbeu.edu.cn; jixiangyu@hrbeu.edu.cn; chengengic@hrbeu.edu.cn; 983402243@hrbeu.edu.cn).

The code will be downloaded to https://github.com/fengqinshou/MST-SSSNet.

Digital Object Identifier 10.1109/JSTARS.2023.3342983

convolutional networks have been widely explored in HSI classification [33], [34], [35]. Zhu et al. [36] designed a multiscale long and short graph convolution, which makes full use of texture structures of different sizes and captures local and global spectral information at the same time. In addition, capsule networks [37], [38] and generative adversarial networks (GANs) [39], [40] have also been explored in HSI classification but they all have their own shortcomings, which can be summarized as follows. RNN may experience problems of vanishing gradient and exploding gradient. The risk of overfitting in GCN is heightened by their substantial parameter count. The training cycle of a capsule network is long and there is a problem of capsule congestion. There is a problem of pattern collapse in GANs.

In the most recent developments, the vision transformer (ViT) [41] has attained remarkable success within the realm of computer vision. Hong et al. [42] proposed spectralFormer, which learns spectral features from shallow to deep through spectral grouping embedding and cross-layer adaptive fusion. However, it puts too much emphasis on the learning and modeling of spectral sequences, ignoring the spatial characteristics. Huang et al. [43] proposed a supervised contrastive spectral–spatial masked transformer (SC-SS-MTr). However, transformers have always had the disadvantage of weak local modeling capabilities. CNN has been widely studied for its local modeling ability to solve this problem [44], [45]. Sun et al. [46] introduced a transformer with feature tokenization using the CNN feature extraction module. Roy et al. [47] designed a morphFormer network with morphological operations.

While the above-mentioned methods-based transformers and CNNs have been extensively explored in the HSI classification, there are still several shortcomings. The transformer has high requirements for training data and computing resources [48] and the combination with CNN makes this problem more serious. This not only makes the model easy to overfit but also makes the algorithm difficult to deploy on the HSI classification device, limiting the application of the algorithm. Moreover, local information and global information are extracted by different modules, respectively, and the coupling relationship between the two types of information is weak. Specifically, in previous methods, the global correlations established through the transformer were mostly directly based on feature maps extracted from local information by CNN. Not introducing local information to guide the establishment of global correlations may lead to incorrect category information being introduced into the model. Such an outcome could cause imprecise extraction of features, thereby constraining the potential improvement in classification performance. In response to the issues outlined above, we present a solution termed main–sub transformer with spectral–spatial separable convolution (MST-SSSNet) designed specifically for HSI classification. It enables a more effective utilization of both local and global spectral–spatial information, effectively integrating these two types of information. Our method includes two key modules: spectral–spatial separable convolution (SSSC) module and main–sub transformer encoder (MST) module. We use the SSSC module for feature extraction to capture shallow spectral–spatial features and use the MST module to learn local information to establish the correlation of global information.

For the SSSC module, the previous methods were mostly based on 3-D convolution design feature extraction modules, which have the advantage of extracting spectral–spatial features. However, this will result in modules having higher algorithm complexity. Some models also use 1-D and 2-D convolutions to extract spectral and spatial features, respectively, which reduce the requirement for computing resources. However, it cannot capture the potential correlation between spectral and spatial information.

We innovatively propose SSSC to build a local feature extraction module. The idea of this convolution is to extract spatial–spectral features through two orthogonal 2-D convolutions instead of the 3-D convolution commonly used in previous methods. The two spatial dimensions and the spectral dimension form two orthogonal planes, respectively. Two 2-D convolutions perform convolution operations through these two planes respectively to replace the 3-D convolution. Therefore, spectral–spatial feature information can be effectively extracted with fewer parameters.

For the MST module, unlike previous transformer-based methods, the main–sub transformer proposed is not a simple two-layer or double-branch structure, it is designed as a two-step global feature construction. The first step is to learn local features through the subtransformer, and the second step is to introduce the learned local features into the main transformer to jointly establish global information with the original data. The pivotal contributions of this study can be outlined as follows.

1) We propose a method named MST-SSSNet for HSI classification, which uses the designed SSSC to effectively extract spectral–spatial features and then uses local information to model global information.
2) We design a lightweight SSSC module that replaces the 3-D convolution with our SSSC. We innovatively use an orthogonal structured 2-D convolution instead of 3-D convolution to achieve efficient feature extraction.
3) We design an MST module including a subtransformer and a main transformer. Unlike cascaded or dual-branch structures, the designed MST module assists the main transformer in establishing global information by dividing the image into different scales and learning local information through the subtransformer.

II. Related Work

In recent years, deep learning has developed rapidly. Compared with traditional hand-designed feature extractors, deep learning methods usually have stronger feature modeling capabilities and robustness [22]. Therefore, HSI classification methods based on deep learning have been widely explored. In this section, we will introduce the HSI classification method based on CNN, the HSI classification method based on transformer, and other deep learning methods.

A. Method Based on CNN

CNNs have been explored a lot in HSI classification [31], [49], [50], [51]. Hu et al. [28] involved the effective extraction of spectral features through the use of 1D-CNN. Zhao and Du

[29] used the principal component analysis (PCA) to reduce the dimensionality of HSI data and then extracted spatial features through 2D-CNN. In [30], spectral and spatial features were extracted concurrently using a 3D-CNN. Furthermore, regularization techniques were applied to enhance the system's generalization prowess. In [49], a method combining 3-D and 2-D is proposed, which aims to extract spectral spatial features and use them for classification. Roy et al. [31] combined 3-D convolutional layers of different scales to extract features and used 2-D convolutional layers to enhance spatial feature extraction. It achieved good results. In [50], a global framework classification pattern has been designed to alleviate the problem of difficult extraction of global information caused by traditional patch-based dataset partitioning. Yu et al. [51] introduced the feedback attention mechanism proposed in this article into CNN networks and designed a dense spatial–spectral CNN structure for HSI classification, called FADCNN, which improved classification performance. In addition to the above-mentioned work, some methods consider introducing CNN into other backbone networks to enhance classification performance [33], [34], [35], [59], [46]. These work with better performance than separate networks.

B. Method Based on Transformer

Transformers have received widespread attention for their excellent long-distance modeling capabilities. It also demonstrates excellent potential in HSI classification [52], [53], [54], [55]. He et al. [56] used transformers and proposed BERT to HSI classification. Hong et al. [42] used transformer encoders and designed a cross-layer fusion structure to create spectralFormer. Cao et al. [57] proposed a transformer-based MAE using contrastive learning, attempting to combine these two methods to further improve performance. Huang et al. [43] designed a spectral–spatial masked transformer (SS-MTr) and through contrastive and supervised learning proposed an SC-SS-MTr. Some other methods have been used to improve transformers. Yu et al. [58] designed a lightweight classification network using an image-level framework and CNN and transformer. In [44], a multiscale network was designed for more accurate feature extraction, and the transformer was improved for HSI image classification. Huang et al. [44] introduced active learning into the transformer and designed a learning strategy combining superpixel segmentation while improving the transformer using Outlook attention. However, a transformer has the disadvantage of poor local modeling ability. Some works have solved this problem by combining CNN with a transformer, which has been proven to be effective. He et al. [59] used a VGG-like network to extract the spatial features of HSI data and then modeled the spectral information through a transformer. Sun et al. [46] used a CNN network consisting of 3-D and 2-D convolutional layers for feature extraction and learned advanced semantic information through a transformer. Roy et al. [47] combined the attention mechanism with morphological operations to improve the interaction and proposed morphFormer. Yang et al. [60] designed a multilevel feature fusion network for class prediction using interactive CNN and transformer. This network can extract category features of different perception fields and depths for better prediction.

C. Method Based on Other Networks

1) Method Based on RNN: Zhang et al. [61] scanned HSI into a sequence of pixels and each pixel and its spectral information is a step of the model. In [62], the RNN model was simplified and designed into an efficient model that can be extended. Zhou et al. [63] studied the effectiveness of multiple scanning strategies to generate features. This strategy is proven to have significant improvements in RNNs.

2) Method Based on GCN: In [64], locality preserving low-pass graph convolutional embedding autoencoder is proposed, and self-training clustering mechanism and joint optimization loss are introduced to achieve mutual benefit. Ding et al. [33] combined multiple filters through defined degree scales to better process HSI information. Zhang et al. [65] proposed an adaptive receptive field graph neural framework that can alleviate excessive smoothness in the model and reduce computational complexity.

3) Method Based on GAN: Zhang et al. [39] designed a semisupervised HSI data framework based on one-dimensional GAN for HSI classification. Sun et al. [40] proposed an auxiliary classifier based on the gradient penalty Wasserstein GAN (ACWGAN-GP). Feng et al. [66] introduced contrastive learning into GAN and designed a pair of coarse-grained GAN networks.

III. Proposed Method

A. Overall Framework of the Proposed Method

Fig. 1 illustrates the overall framework of our MST-SSSNet method. In this section, we introduce the MST-SSSNet method from three steps: preprocessing, spectral–spatial separable convolutional module, and main–sub transformer module.

In the preprocessing part, the original HSI data are processed by PCA to reduce the redundancy existing in the data. This step is crucial to improving the overall algorithm speed. The data after dimension reduction need to be block extracted to generate a dataset for model learning, and the dataset is divided into a training set and a test set.

The data of the training set are used as the input of the SSSC module, which is a two-branch structure designed by the SSSC, and finally uses a 2-D convolution to enhance the extraction of spatial features, and finally outputs the feature map.

The feature map first generates a patch token through transformation. Then the patch token generates subtokens and main tokens, respectively. The main tokens are the input of the main transformer and the subtokens are the input of the subtransformer. The subtoken learns local detail features through the subtransformer and then fuses them with the main token to help the main token establish global features in the main transformer. Finally, the classification structure is obtained through a linear layer and softmax function.

B. Preprocessing

HSIs typically comprise a substantial number of spectral bands, and there is high information redundancy between adjacent bands. These issues increase the complexity of computation and storage and may lead to overfitting issues. As an effective method for dimensionality reduction, PCA streamlines data

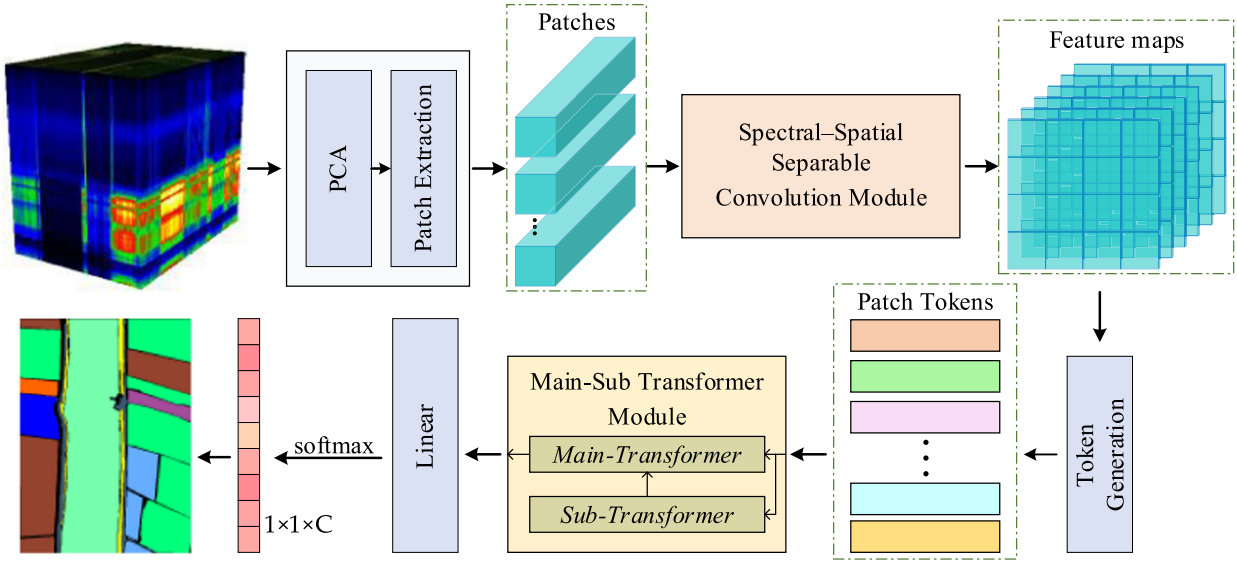

Fig. 1. Overall framework of the proposed MST-SSSNet.

complexity and minimizes redundancy among spectral bands. It changes the feature spatial of pixels to better distinguish spectral signals of different substances. Thus, PCA is applied in the handling of HSI data. The HSI data, having undergone PCA, are recorded as $\mathrm{I}_{\mathrm{PCA}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{D}}$. H and W denote the spatial size, and C represents the center pixel to extract a patch $\mathrm{P} \in \mathbb{R}^{S\times S\times D}$ from $I_{\mathbf{PCA}}$. S is the size of the patch. The true label is ascertained based on the label assigned to the central pixel. Edge pixels need a padding operation before being extracted, with a filling width of $(S-1)/2$. After extracting all patches, we separate the remaining samples into a training set and a test set.

C. Spectral–Spatial Separable Convolution Module

To address the above issues, inspired by spatial separable convolution [67], we designed SSSC to replace 3-D convolution. Richer spectral–spatial features are extracted with equivalent parameter quantities.

Fig. 2 illustrates the comparison of details between 3-D convolution and SSSC. X and Y denote spatial dimensions. L represents the spectral dimension. The input cube data are used 2-D convolution on the Y–L plane for convolution. It is called the Y–L convolution. Then convolution is performed on the X–L plane. It is called the X–L convolution. The above is used to replace the 3-D convolution. The convolution of data with a size of $7\times7\times7$ is taken as an example. The $7\times7\times7$ sized blue block represents the original data, and the 3×3 sized blocks of different colors represent Y–L convolution kernels and X–L convolution kernels. We calculate the parameter quantities for 3-D convolution and SSSC. Selecting a convolution kernel with a size of $3\times3\times3$ for 3-D convolution requires $175\times27=4725$ multiplication operations. SSSC selects kernel sizes $3\times1\times3$ and $3\times3\times1$. It takes $175\times9+75\times9=2250$ times. The procedure significantly diminishes the parameter count, leading to enhanced computational efficiency. Compared to networks designed based on 3-D convolution, we by using the SSSC

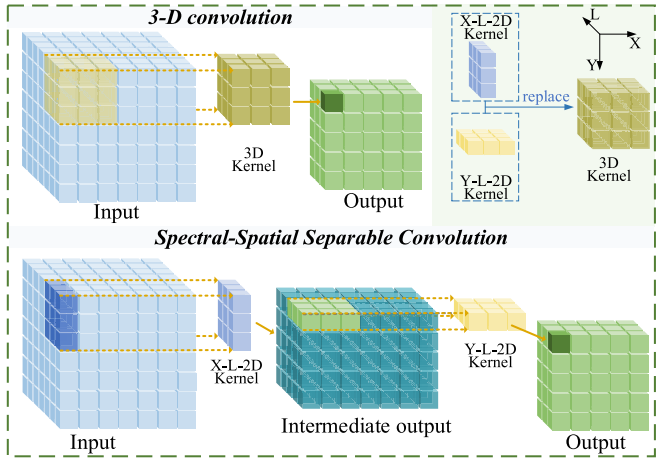

Fig. 2. Comparison of details between 3-D convolution and SSSC.

can design deeper networks, thereby enhancing the ability to capture features. It should be noted that the number of times we use a pair of SSSCs to extract features in the frequency band dimension is two. Within the context of parameter reduction, we have simultaneously refined our capacity for spectral feature extraction. Consequently, employing SSSC enables the extraction of profoundly efficient feature representations, enriching the process of network learning To elaborate further, in SSSC, the calculated values of the jth feature cube of X–L convolution and Y–L convolution in the ith layer at the spatial position can be given as

$$v_{ij}^{xyz}=\Phi\left(b_{ij}+\sum_{m}\sum_{p=0}^{X_i-1}\sum_{r=0}^{L_i-1}w_{ijm}^{pr}v_{(i-1)m}^{(x+p)y(z+r)}\right) \tag{1}$$

$$v_{ij}^{xyz}=\Phi\left(b_{ij}+\sum_{m}\sum_{q=0}^{Y_i-1}\sum_{r=0}^{L_i-1}w_{ijm}^{qr}v_{(i-1)m}^{x(y+q)(z+r)}\right) \tag{2}$$

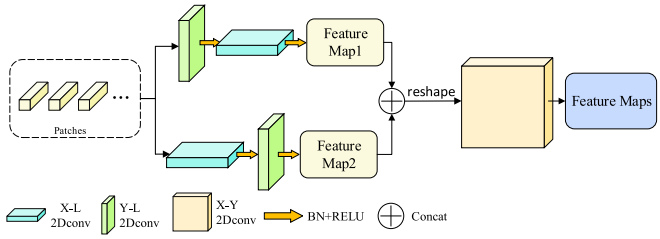

Fig. 3. Illustration of the SSSC module.

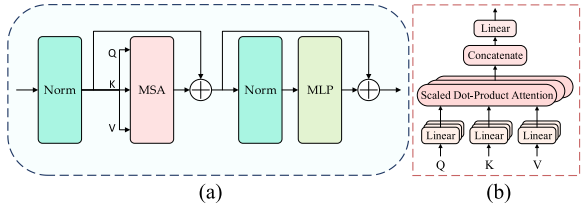

Fig. 4. Details for a transformer encoder and MSA. (a) Transformer encoder. (b) MSA.

where $\Phi(\cdot)$ denotes the activation function and b_{ij} represents the bias. m is the feature map in the $(i-1)$th layer that is connected to the jth feature map. Y_i, X_i, L_i, respectively, represent the height, width, and channel number of the SSSC kernel. L_i stands for spectral dimension. In the mth feature cube, the parameters w_{ijm}^{pr} and w_{ijm}^{qr} correspond to the weights linked with the position (p,q,r).

The SSSC module is graphically presented in Fig. 3. The activation function will result in spatial asymmetry in SSSC. This asymmetry will affect the network's preference for learning features. Therefore, we design a dual-branch network to eliminate this preference. The SSSC module processes data in two stages. In the first stage, two pairs of SSSCs in different orders perform feature extraction on the input patches. The first-stage feature mapping formula can be described as follows:

$$X_{P1} = \mathrm{RELU}(\mathrm{BN}(\mathrm{ConvYL}(\mathrm{RELU}(\mathrm{BN}(\mathrm{Conv}XL(X))))) \tag{3}$$

$$X_{P2} = \mathrm{RELU}(\mathrm{BN}(\mathrm{Conv}XL(\mathrm{RELU}(\mathrm{BN}(\mathrm{Conv}YL(X)))))) \tag{4}$$

where X denotes the input. X_{P1} and X_{P2} represent the feature maps.$\mathrm{RELU}(\cdot)$ is the activation function, $\mathrm{BN}(\cdot)$ denotes the BatchNorm operation.Conv$XL(\cdot)$ and Conv$YL(\cdot)$ represent the different direction SSSCs. In the second stage, by reshaping and fusing the feature maps from the two branches, a new feature map is derived. Then the feature map undergoes standard 2-D convolution operations, enhancing the network's spatial feature extraction ability. The formula for the second stage is given as follows:

$$X_{P3} = \mathrm{RELU}(\mathrm{BN}(\mathrm{Conv2D}(\mathrm{reshape}(\mathrm{X}_{P1}) \oplus \mathrm{reshape}(\mathrm{X}_{P2})))) \tag{5}$$

where the SSSC module output is defined as $X_{P3} \in \mathbb{R}^{m\times n\times d}$.$m$ is the height and n represents the weight. The number of channels is represented as d. $\oplus$ represents the concatenate function. The extracted features will provide a good feature representation for subsequent module processing.

D. Token Generator and Main–Sub Transformer Module

1) Token Generator: SSSC module has extracted good local spectral–spatial features $X_{P3} \in \mathbb{R}^{a\times a\times d}$, the spatial size is represented as a, and d signifies the number of channels in the given context. The feature token is defined as $\mathrm{X_T} \in \mathbb{R}^{\mathrm{w}\times\mathrm{d}}$, where w represents the token number. It can be obtained by the following formula:

$$X_{\mathrm{flat}} = \mathrm{Flatten}(X_{P3}) \tag{6}$$

$$X_{\mathrm{T}} = \mathrm{softmax}(W_a X_{\mathrm{flat}})^{\mathrm{T}} X_{\mathrm{flat}} \tag{7}$$

where W_a represents a learnable weight matrix. The feature map is projected into tokens through the above-mentioned steps [45] for subsequent operations.

2) Main–Sub Transformer Module: In Fig. 4(a), both the main-transformer encoder and the subtransformer encoder share an identical architecture. This configuration includes two normalization layers (LN): a multihead self-attention (MSA) block and a multilayer perceptron (MLP) layer. Residual skip connections are applied prior to the MSA block and the MLP layer. The specifics of the MSA are illustrated in Fig. 4(b). The entire process can be mathematically expressed as follows:

$$\mathrm{SA} = \mathrm{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d_K}}}\right)\mathrm{V} \tag{8}$$

$$\mathrm{MSA}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \mathrm{Concat}(\mathrm{SA}_1,\mathrm{SA}_2,\ldots,\mathrm{SA_h})\mathrm{W} \tag{9}$$

where the matrices Q, K, and V are learnable weight matrices. SA represents self-attention. The incorporation of SA in the MSA block effectively captures correlations among feature sequences. Within the MSA block, multiple groups of weight matrices are employed to map Q, K, and V, employing a uniform operation process to calculate multihead attention values. Subsequently, the results from each head attention are concatenated. "h" denotes the number of heads, and W signifies the parameter matrix. The MLP comprises two fully connected layers separated by a nonlinear activation function known as Gaussian error linear units. In order to establish the correlation between local and global information, we designed a main–sub transformer. When people understand a sentence, they first understand the meaning of the words or phrases; second, they establish the context of the entire sentence through their meanings, thereby understanding the entire sentence. This idea also applies to the model's understanding of HSI. When establishing global correlation, irrelevant information is often introduced, which affects the model's feature learning. Entering local information to assist in the establishment of global correlation will greatly improve this problem. Specifically, we divide the input patch tokens into more fine-scale subtokens. Its detailed features are then learned through a subtransformer. Unlike general dual-scale or multiscale methods, we do not cascade two or more transformers at different scales. Instead, the detailed features learned by the subtransformer are integrated into the main token.

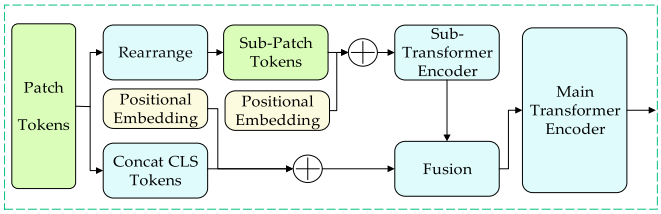

Fig. 5. Illustration of the main–sub transformer module.

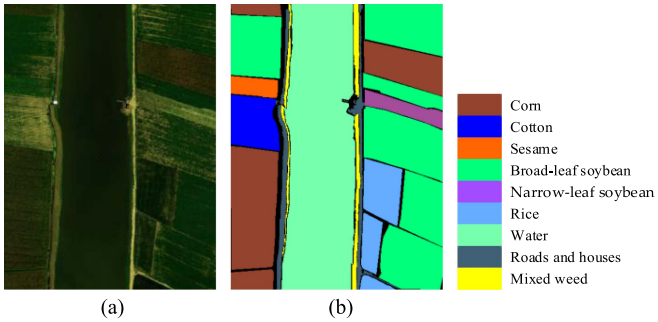

Fig. 6. LongKou dataset. (a) False-color composite image. (b) Ground truth map.

Finally, the main transformer is used to learn global correlations while retaining the detailed features. The details of the main–sub transformer module are shown in Fig. 5.

Each token as input represents $X_{\mathrm{T}} = [\mathrm{T}_1, \mathrm{T}_2, \ldots, \mathrm{T}_{\mathrm{w}}]$. We divide the token into n subtokens by reshaping the data form. After adding trainable position coding, it is expressed as $X'_{\mathrm{T}} = [\mathrm{T}_{11}, \mathrm{T}_{12}, \ldots, \mathrm{T}_{1\mathrm{n}}, \mathrm{T}_{21}, \ldots, \mathrm{T}_{\mathrm{wn}}] + \mathrm{PE}_{\mathrm{sub}}$. A subtransformer encoder is used to learn local details. Then the subtoken is converted into the same data form $X_{\mathrm{t}} = [\mathrm{T'}_1, \mathrm{T'}_1, \ldots, \mathrm{T'}_{\mathrm{w}}]$ as the token. After processing a block consisting of two normalization layers and a linear layer, it is added to the patch token. Finally, it is concatenated with a learnable class token $\mathrm{T}_0^{\mathrm{cls}}$ to get the main token. The comprehensive process outlined above is encapsulated by the following equations:

$$X_{\mathrm{t}} = \mathrm{Norm}(\mathrm{Linear}(\mathrm{Norm}(X'_{\mathrm{T}}))) \tag{10}$$

$$X_{\mathrm{m}} = X_{\mathrm{t}} + X_{\mathrm{T}} \tag{11}$$

$$X_{\mathrm{main}} = \mathrm{Concat}(\mathrm{T}_0^{\mathrm{cls}}, X_{\mathrm{m}}). \tag{12}$$

The main token is the input of the main-transformer encoder. Finally, we only use $\mathrm{T}_0^{\mathrm{cls}}$ as the input to the linear layer for the classification. Via the linear layer, the probability of the input belonging to a specific class will get through the softmax function. The label corresponding to the highest probability signifies the class of the given sample.

IV. Experiment and Analysis

This section introduces the WHU-Hi dataset used in the experiments, including LongKou, HanChuan, and HongHu [68], [69]. Furthermore, we present the experimental settings, including evaluation indicators, configuration, and parameter analysis. Then, the ablation experiment is introduced. Finally, we show and analyze the results.

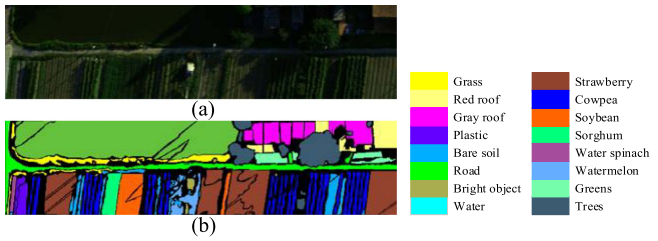

Fig. 7. HanChuan dataset. (a) False-color composite image. (b) Ground truth map.

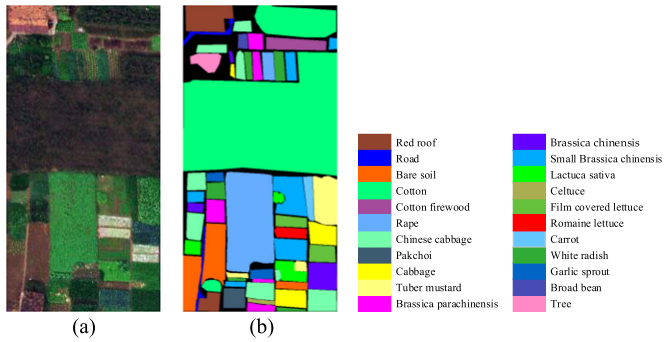

Fig. 8. HongHu dataset. (a) False-color composite image. (b) Ground truth map.

A. Data Description

To evaluate the effectiveness of our MST-SSSNet, rigorous experiments were executed utilizing the WHU-Hi dataset, a UAV-borne HSI resource provided by the RSIDEA research group at Wuhan University. This dataset is publicly available and can be downloaded from the RSIDEA team's homepage. This dataset encompasses three hyperspectral datasets: WHU-Hi-LongKou, WHU-Hi-HanChuan, and WHU-Hi-HongHu. The False-color composite image and ground truth maps of the three datasets are shown in Figs. 6–8.

The LongKou dataset was procured through an 8-mm focal length headwall nanohyperspectral imaging sensor mounted on a DJI Matrice 600 Pro, conducting aerial surveys over Longkou Town in China in the year 2018. Within this study region, there exist nine specific classes, comprising six diverse crop species. The dataset consists of 550 $\times$ 400 pixels and ranges from 400 to 1000 nm, including 270 bands.

The HanChuan dataset was acquired utilizing a 17-mm focal length headwall nanohyperspectral imaging sensor installed on a Leica Aibot X6, conducting surveys over HanChuan, China, in the year 2016. There are 16 classes including 7 crop species: watermelon, water spinach, greens, strawberry, cowpea, greens, and sorghum. The dataset consists of 1217 $\times$ 303 pixels ranging from 400 to 1000 nm and includes 274 bands. Notably, because the time to collect the HanChuan dataset was in the afternoon, there are many areas that are shadow-covered in the image.

The HongHu dataset was captured using a 17-mm focal length headwall nano-hyperspectral imaging sensor mounted on a DJI Matrice 600 Pro, conducting surveys over HongHu City in China in the year 2017. The experimental area comprises a diverse agricultural landscape featuring 22 distinct classes, with varying cultivars of the same crop cultivated within the region. The size

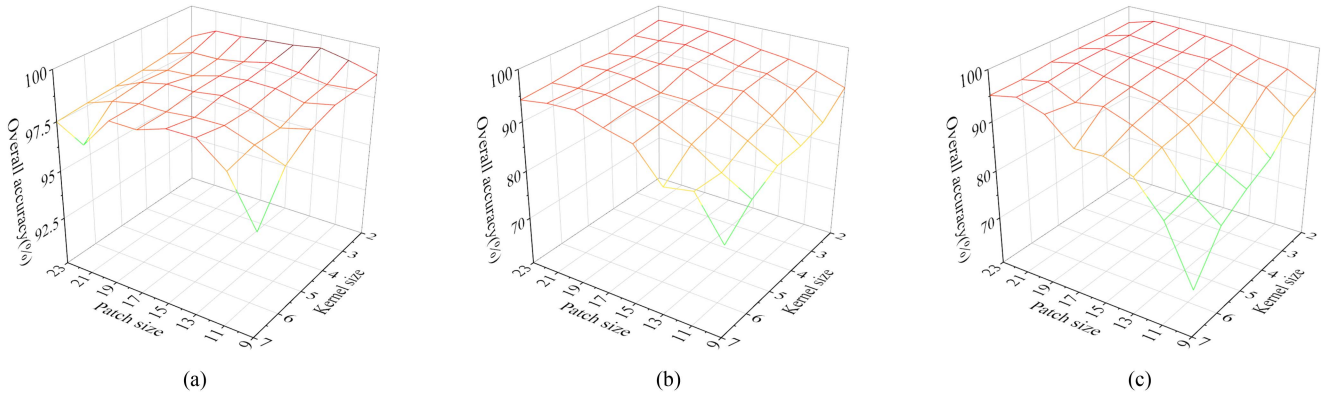

Fig. 9. Impact between different SSSC kernel sizes and patch sizes for the OA. (a) LongKou. (b) HanChuan. (c) HongHu.

of the image is 940 × 475 pixels and ranges from 400 to 1000 nm, comprising 270 bands.

B. Experimental Setting

1) Evaluation Indicators: In our experimental framework, we adopted overall accuracy (OA), average accuracy (AA), and Kappa coefficient (K) as assessment metrics [70]. OA represents the comprehensive accuracy across all samples, while AA signifies the mean accuracy computed for each individual class. Kappa coefficient gauges the alignment between the classification outcomes and the true underlying classes. The mathematical expressions for these metrics are detailed as follows:

$$\mathrm{OA} = \left(\frac{1}{n}\sum_{k}\left(\frac{\text{True positive} + \text{True Negative}}{\text{Total number of pixels}}\right)\right) \tag{13}$$

$$\mathrm{Recall} = (\text{True positive} + \text{False negative}) \tag{14}$$

$$\mathrm{AA} = \frac{\sum_{i=1}^{n}\mathrm{Recall}_i}{n} \times 100\% \tag{15}$$

$$K = \frac{N\sum_{i=1}^{n} x_{ii} - \sum_{i=1}^{n}\left(x_{i+} \times x_{+i}\right)}{N^2 - \sum_{i=1}^{n}\left(x_{i+} \times x_{+i}\right)}. \tag{16}$$

2) Configuration: The verification experiments for the proposed methodology were conducted within the PyTorch computational framework, using an 11th Gen Intel(R) Core(TM) i9-11900K processor clocked at 3.50 GHz, and an NVIDIA GeForce RTX 3090Ti 24-GB GPU server. The optimization process is initiated with the Adam optimizer, and the learning rate is set to 0.001. During batch training, each batch was configured to contain 64. The original spectral band number for the LongKou and HongHu datasets is 270, while the original band number for the HanChuan dataset is 274. The number of bands after PCA operation is set to 30. The size of the input image is the same as the size of the original dataset. The LongKou consists of 550 × 400 pixels. The HanChuan consists of 1217 × 303 pixels. The HongHu consists of 1217 × 303 pixels. A total of 100 training epochs were applied to each dataset. Ten experiments were rigorously carried out for every method, and the most optimal result among them was meticulously selected as the definitive outcome.

3) Patch Size and SSSC Kernel Size Analysis: We analyzed the impact of input patch size and SSSC kernel size on classification accuracy. In Fig. 9 the impact of patch size and kernel size on classification accuracy is shown.

In general, the accuracy of classification tends to diminish as the kernel size increases. In particular, we choose [2], [7] all kernel sizes instead of only odd ones. The result shows that odd kernel sizes have no obvious advantages over even kernel sizes. The reason for this may be that the feature has captured the central feature because of the two orthogonal convolutions in the spatial dimension. For the channel dimension, the centermost channel has the same importance as the other channels, so there is no obvious difference between odd and even convolution kernels. The best kernel size of the three datasets is 2.

For patch size, the accuracy of classification exhibits an initial rise, followed by a subsequent decline as the patch size increases. The patch sizes with the size of [9], [23] are tested in experiments. In the local range, it accords with the characteristics of convex function. For three datasets, when small patch size and large kernel size are combined, it usually leads to a serious decrease in accuracy. That is because too small a patch and too large a kernel can result in too few convolutions on a patch and cannot accurately extract features.

For the LongKou dataset, classification accuracy is not sensitive to changes in two parameters. The optimal patch size is 15. In HanChuan and HongHu, the influence of the two parameters on classification accuracy exhibits similarity. The optimal patch size is 21. The experimental results indicate that selecting the appropriate patch size and kernel size can better extract features. Following an evaluation of the performance across the three datasets, a patch size of 21 was chosen as the optimal selection.

C. Ablation Studies

To fully validate the effectiveness of our method, ablation experiments with different component combinations were conducted on the HongHu dataset. Five combinations were considered. We assess the effect of distinct components on the overarching model through their impact on classification accuracy. The results of all experiments are meticulously documented in

TABLE I
ABLATION EXPERIMENT OF THE PROPOSED METHOD ON THE HONGHU DATASET

Methods	OA(%)	AA(%)	Kappa(%)	Train time(s)	Test time(s)
SSSC + Linear	85.11	53.24	80.80	39.58	110.06
only MST	93.86	82.67	92.24	27.24	85.35
3Dconv + MST	97.79	93.18	97.21	55.75	169.40
SSSC + TE	97.15	92.47	96.40	43.41	141.29
SSSC + MST	**97.84**	**94.51**	**97.27**	50.22	130.48

Table I, with the most superior classification outcome emphatically presented in bold typeface.

The effectiveness experiment of the MST module is performed by replacing it with a linear layer and ordinary transformer encoder (TE). About the linear layer, the feature map output by the SSSC module is reshaped and connected to the linear layer. In the linear layer, in_features is set to 1600 (the same size as the reshapped feature map), and out_features is set to 22 (the same number of categories as the dataset). Among them, the combination of the SSSC module and linear layer has the worst effect because the features lack the establishment of long-range correlation. The MST module is superior to ordinary transformer encoders, especially in terms of significantly improving AA. This indicates that the MST module can reduce the introduction of irrelevant information by establishing long-range correlation through local information. Thereby the classification accuracy is improved.

To validate the efficacy of the SSSC module, two distinct sets of experiments were crafted. One set of experiments replaced the SSSC module with 3-D convolutional layers. The depth of the 3-D convolutional layer is 1, the number of convolutional kernels is 8, and the size is $3 \times 3 \times 3$. The other set of experiments only used the MST module after PCA operation. The MST module operating in isolation demonstrates inferior classification accuracy attributed to its absence of local feature extraction capabilities. The combination of the SSSC module and the MST module is better than the combination of the 3-D convolution and MST module in classification performance and calculation time, which shows that the proposed spectral–spatial separable volume of the product can extract richer features with fewer parameters. In summary, the analysis of comprehensive experimental results further confirms the effectiveness of our model.

D. Analysis of Results

In this section, three datasets will be selected for the experiment. The training samples within each dataset have been allocated at an average of 100 samples per class.

For the LongKou dataset, we randomly selected 900 samples as the training set. In the case of the HanChuan dataset, 1600 samples were meticulously chosen to form the training set. For the HongHu dataset, 2200 sample training sets were selected. The samples not chosen for training are specifically allocated as test sets to evaluate the model's efficacy.

1) Comparison Methods: To validate the proposed methodology, a selection of representative baseline methods as well as the most advanced backbone methods have been chosen. We divide them into nontransformer methods (such as SVM [13], 1DConv [28], 2DConv [29], 3DConv [30], and HybridSN [31]) and transformer-based methods (such as ViT [38], spectralFormer [40], SSFTT [45], and SC-SS-MTr [41]). The introduction to these methods is as follows.

SVM is one of the first machine-learning methods used in HSI classification. It represents the classic machine learning method. The 1-D-CNN is a method including a 1-D convolutional layer, maximum pooling layer, and fully connected layer. The 2-D-CNN is a method that includes three convolutional layers and two fully connected layers. HybridSN consists of a convolutional part, a fully 2-D convolutional layer, and two fully connected layers. The 3-D-CNN is a 3-D convolution-based method, including three 3-D connected parts. The convolutional part includes three 3-D convolutional layers of different scales and one 2-D convolutional layer. The linear layer part consists of two linear layers. MST-SSSNet surpasses current state-of-the-art methods across three datasets based on three comprehensive evaluation indicators. The best performer among CNN-based approaches is HyBridSN. Thanks to the design of the convolutional kernel size in the network, HybridSN pays attention to more abundant spectral features and achieves better results compared with other CNN-class methods. This shows that extracting rich spectral features can better distinguish different classes. In addition to the proposed methods, SSFTT and SC-SS-MTr perform better than other transformer-based methods. SSFTT benefits from the excellent basic architecture combined with CNN and transformer and extracts local and global features at the same time to achieve better classification performance. SC-SS-MTr enhances the performance of the transformer and achieves competitive classification results through supervised learning, discriminant feature learning, and mask. The proposed MST-SSSNet uses the designed SSSC to extract spectral–spatial features with fewer parameters so that the network can be designed deeper and the local information can be extracted more fully. Therefore, it has the best classification results. For the LongKou dataset, the OA, AA, and Kappa coefficients of our method reached 98.97%, 96.16%, and 98.64%, respectively. On this dataset, HybridSN and most transformer-based methods have achieved competitive classification results. We posit that this can be attributed to the limited number of classes in the LongKou and the concentrated distribution of these categories, leading to a reduced level of classification complexity. Datasets have low requirements for model classification ability. However, our method still achieves the best classification performance. For HanChuan, the proposed method reached 97.15%, 93.02%, and 96.67% in three evaluation indicators, and the best classification accuracy was achieved in 12 out of 16 classes. Compared with other methods, OA, AA, and Kappa coefficients of the best-performing methods increased by at least 0.89%, 4.51%, and 2.22%. For the HongHu dataset, the OA, AA, and kappa coefficients of the proposed method

TABLE II
COMPARATIVE EXPERIMENTAL RESULTS ON THE LONGKOU DATASET

Class	SVM	1D-CNN	2D-CNN	3D-CNN	HybridSN	ViT	SpectralFormer	SSFTT	SC-SS-MTr	MST-SSSNet
1	98.17	87.74	99.59	98.29	96.43	99.20	99.60	**99.99**	99.14	99.94
2	91.07	98.61	91.16	98.55	99.11	**99.79**	98.19	99.15	99.03	99.30
3	93.21	93.24	90.94	94.13	91.29	94.16	99.34	98.38	**99.97**	97.98
4	88.03	85.88	98.91	98.79	99.27	**99.66**	97.64	98.72	98.21	98.62
5	93.26	94.06	89.96	98.83	97.59	97.23	85.50	98.28	**98.96**	97.87
6	98.88	87.79	95.27	97.53	97.75	99.52	99.12	99.83	99.83	**99.85**
7	99.98	93.98	99.12	98.83	**99.92**	99.15	99.83	99.76	96.74	**99.92**
8	90.90	65.92	83.65	79.73	87.98	91.24	93.52	89.85	**97.02**	91.62
9	87.73	74.56	96.14	98.25	97.40	92.23	**98.43**	94.33	97.33	93.76
OA(%)	94.69	88.75	97.64	97.78	98.76	98.82	98.44	98.90	98.47	**98.97**
AA(%)	93.47	85.02	91.29	92.35	96.35	96.05	94.56	95.87	95.99	**96.16**
K ×100	93.10	74.01	96.90	97.09	98.37	98.45	97.95	98.55	97.37	**98.64**

Bold numbers represent the best results for the corresponding class.

reach 97.84%, 97.27%, and 94.51% in OA, AA, and Kappa coefficients, respectively, and the best classification accuracy is achieved in 12 out of 22 classes. Compared with the best performance of other methods, OA, AA, and Kappa coefficients can be improved by at least 1.08%, 2.11%, and 1.27%.

In HanChuan and HongHu, the accuracy of only two and one classes of our method is lower than 90% respectively, while the accuracy of the other classes is relatively stable. HanChuan's class13 is a difficult class to classify. Although the classification accuracy of our method is lower than 90% (88.32%), it is still the highest accuracy in this class. Other methods have more than five classes with a classification accuracy of less than 90%. This is because when establishing global correlation, combining local features can reduce the influence of irrelevant information so that the proposed method is less affected by class features and has good robustness.

Based on the performance of the three datasets, the proposed methods have achieved the best classification performance. This proves that the global correlation with strong coupling with local information is closer to the core features of the class. And the SSSC module can extract excellent local detail features to provide a good feature representation for subsequent modules. The proposed method can effectively improve the classification performance of the model by local information to establish global correlation.

Vit used a transformer for image processing for the first time and achieved excellent results. It proves that the excellent performance of the transformer is not limited to the field of NLP. SpectralFormer is specifically designed to emphasize the extraction of spectral information. Remarkably, it achieves exceptional classification results without the utilization of convolutional or recurrent units. The improvement of SSFTT on the tokenizer fully explores the ability of the transformer to handle HSIs. Its feature extraction module consists of a 3-D convolutional layer and a 2-D convolutional layer. After being tokenized, it is connected to the transformer encoder and achieves excellent results in image classification. SC-SS-MTr used a spectral–spatial masked transformer, which gains competitive classification results. It uses supervised learning and contrastive learning for better generalization. The parameters and experimental details of all the above methods are set in accordance with the reference papers.

2) Quantitative Results and Analysis: Tables II–IV list the comparison of various classification indicators of various methods under three datasets. The best outcomes are highlighted in bold black for clear visibility. The outcomes unequivocally demonstrate the superior performance of our MST-SSSNet, surpassing current state-of-the-art methods across three datasets based on three comprehensive evaluation indicators.

3) Visual Evaluation: To visually discern the performance disparity between the proposed method and other models, the classification results are plotted as visual classification maps. The visual classification maps of the three datasets of LongKou, HanChuan, and HongHu are shown in Figs. 10–12.

For the LongKou dataset, this is relatively easy to distinguish dataset, except for 1D-CNN and 2D-CNN, which failed to make use of the spatial or spectral features of HSI, resulting in unsatisfactory classification results; other models have achieved good results. However, for some of the difficult areas, such as the left side of the map where the blue extreme meets the top and bottom of the other classes, most methods make a mistake in demarcating the boundaries. Even if the classification accuracy is similar, such errors are often more serious. In addition, we circle an area with a red border to enlarge it for display. In this area, the classification, especially the boundary recognition, is difficult due to the variety of classes. In addition, other methods make it easy to divide some gray areas into blue classes, because the gray samples whose map coordinates are connected with blue affect the learning of the gray class. The proposed method successfully classifies the gray part correctly, and the recognition of the class boundary is more accurate.

For the HanChuan dataset, the classes of this dataset are interleaved, making classification difficult. The classification results of other methods obviously contain more pronounced noise, such as the region near the purple class in the upper right corner. Compared with other methods, the proposed MST-SSSNet gives smoother classification results. We have circled

TABLE III
COMPARATIVE EXPERIMENTAL RESULTS ON HANCHUAN DATASET

Class	SVM	1D-CNN	2D-CNN	3D-CNN	HybridSN	ViT	SpectralFormer	SSFTT	SC-SS-MTr	MST-SSSNet
1	76.95	58.65	87.87	86.89	91.88	94.24	89.12	95.19	97.36	**97.87**
2	54.04	81.04	79.41	97.52	92.41	94.20	92.31	96.37	96.47	**97.68**
3	81.73	41.98	65.76	78.74	86.99	95.07	77.18	92.07	92.72	**97.18**
4	93.72	79.38	85.80	97.80	98.13	98.15	95.94	97.30	93.38	**99.58**
5	80.27	51.65	56.32	55.79	75.59	**99.59**	78.95	95.08	87.38	90.12
6	45.48	32.38	38.72	87.32	61.18	81.60	59.22	85.40	78.01	**93.35**
7	85.75	41.98	62.20	80.71	78.35	81.53	79.64	88.77	84.99	**94.91**
8	56.63	52.10	70.16	82.80	84.96	79.84	86.59	88.19	91.52	**94.60**
9	58.49	47.93	63.84	77.67	77.79	92.87	69.72	90.55	91.73	**93.68**
10	89.65	59.40	97.32	**99.20**	97.18	97.99	94.77	97.60	98.80	97.78
11	95.32	73.14	81.58	77.57	91.00	87.63	79.85	94.05	93.64	**97.76**
12	62.67	35.50	40.09	75.33	64.68	76.06	67.96	83.05	77.94	**91.44**
13	64.07	20.25	55.01	62.54	76.92	80.58	77.62	84.63	87.94	**88.32**
14	71.79	78.95	80.14	84.58	85.90	95.44	82.47	92.86	92.27	**96.74**
15	77.12	51.56	68.50	**92.53**	78.40	90.80	79.87	91.60	94.01	74.60
16	94.92	92.39	98.43	99.70	99.42	99.71	98.72	**99.82**	99.65	99.45
OA(%)	79.02	70.52	84.67	88.90	91.03	93.24	88.74	94.96	95.26	**97.15**
AA(%)	94.29	49.82	61.57	72.90	79.20	84.36	72.47	88.51	88.03	**93.02**
K ×100	95.74	65.09	81.96	86.97	89.48	92.07	86.78	94.09	94.45	**96.67**

Bold numbers represent the best results for the corresponding class.

TABLE IV
COMPARATIVE EXPERIMENTAL RESULTS ON THE HONGHU DATASET

Class	SVM	1D-CNN	2D-CNN	3D-CNN	HybridSN	ViT	SpectralFormer	SSFTT	SC-SS-MTr	MST-SSSNet
1	88.53	90.55	96.86	98.14	97.53	95.51	98.41	97.36	98.40	**98.84**
2	88.62	58.74	65.04	55.13	81.38	77.13	75.94	79.24	72.20	**88.98**
3	74.36	77.98	83.45	**96.03**	87.80	95.17	84.40	91.45	94.15	94.36
4	78.32	85.29	96.52	98.20	98.58	99.05	98.58	98.41	99.31	**99.80**
5	73.25	52.18	85.93	97.47	84.27	86.86	69.23	95.59	92.68	**98.25**
6	80.07	79.34	90.77	94.70	97.68	96.80	95.27	97.70	97.84	**99.20**
7	57.43	59.82	67.93	75.20	88.29	**94.71**	85.85	88.62	90.58	93.97
8	41.02	34.61	26.75	82.04	81.57	78.77	58.31	84.72	**96.79**	95.93
9	91.26	89.53	96.09	**98.63**	96.10	96.97	97.71	97.12	96.49	94.24
10	54.23	67.65	89.80	76.50	90.55	82.01	76.14	95.96	97.37	**98.25**
11	40.49	43.51	68.38	60.95	87.24	81.82	80.27	86.35	94.36	**96.58**
12	54.97	48.04	28.69	74.46	71.18	85.70	60.45	64.54	91.43	**99.09**
13	52.89	62.02	70.23	70.64	85.30	88.80	74.07	97.09	**97.64**	96.72
14	62.81	80.87	62.54	91.33	96.12	95.84	85.99	89.90	96.30	**96.87**
15	86.14	63.87	93.19	**95.94**	81.81	80.93	93.64	75.61	85.13	91.19
16	82.80	94.53	84.88	97.39	96.48	96.33	83.16	86.58	**98.95**	96.91
17	67.42	77.15	45.46	79.10	79.03	65.71	56.50	89.13	92.71	**96.35**
18	71.06	57.02	76.16	86.60	88.14	87.96	75.24	90.65	**92.40**	92.08
19	75.30	61.72	85.18	**97.25**	91.75	92.62	77.37	90.74	**97.25**	93.65
20	80.24	25.14	39.12	77.21	89.54	89.84	78.28	77.48	**89.44**	88.91
21	68.90	25.31	21.29	76.86	74.77	84.78	64.19	84.69	93.38	**98.16**
22	72.77	19.15	55.76	64.45	90.62	87.13	68.41	89.56	86.16	**96.01**
OA(%)	73.02	78.11	86.16	89.71	93.79	94.54	89.60	94.20	96.76	**97.84**
AA(%)	70.13	50.64	63.13	76.60	83.21	87.33	75.65	87.23	92.40	**94.51**
K ×100	67.38	71.32	82.35	87.01	92.13	93.09	86.83	92.65	95.90	**97.27**

Bold numbers represent the best results for the corresponding class.

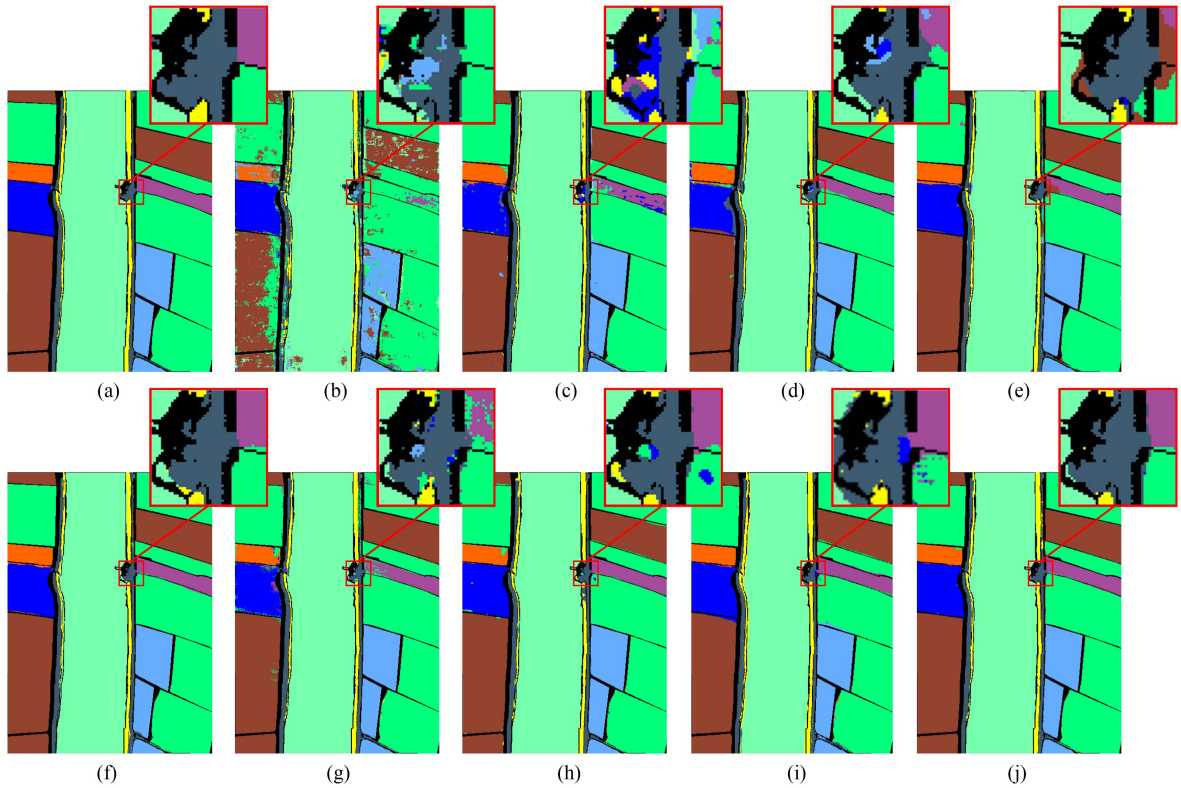

Fig. 10. Visualization of the experimental results based on the LongKou dataset. (a) Ground truth. (b) 1D-CNN. (c) 2D-CNN. (d) 3D-CNN. (e) HybridSN. (f) ViT. (g) SpectralFormer. (h) SSFTT. (i) SC-SS-MTr. (j) MST-SSSNet(ours).

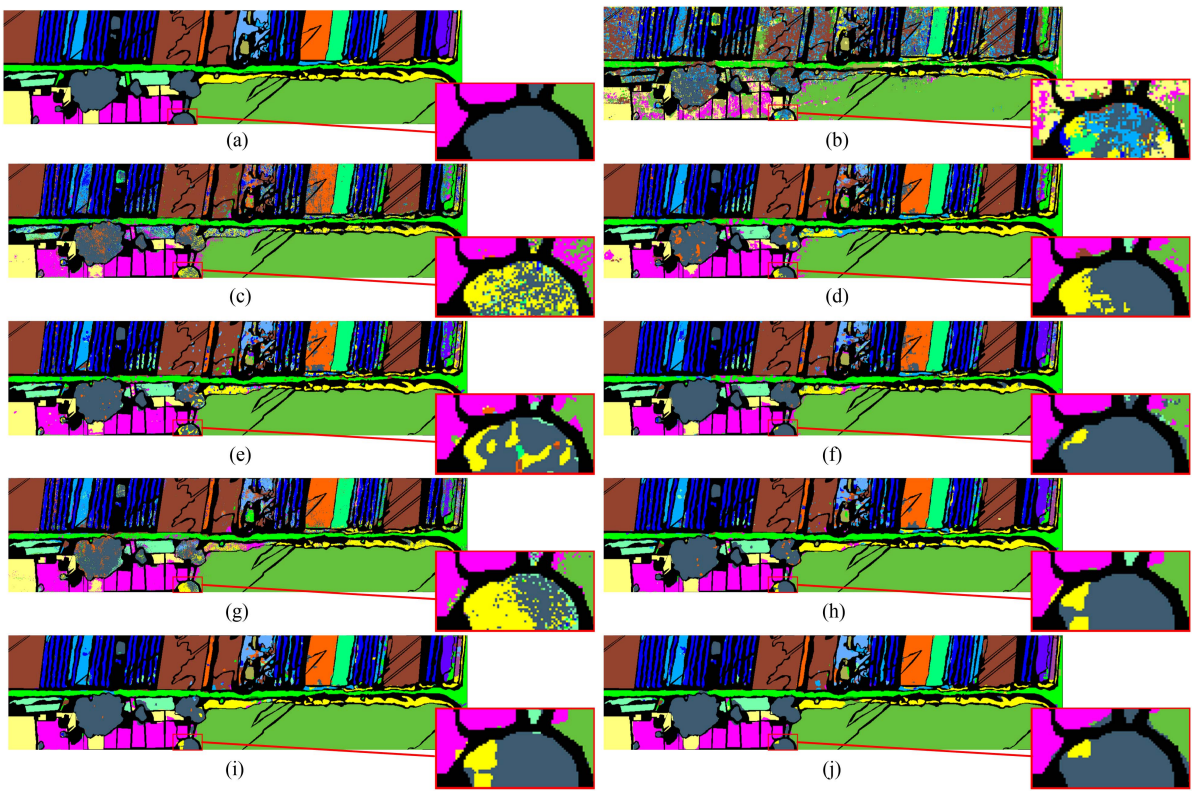

Fig. 11. Visualization of the experimental results based on the HanChuan dataset. (a) Ground truth. (b) 1D-CNN. (c) 2D-CNN. (d) 3D-CNN. (e) HybridSN. (f) ViT. (g) SpectralFormer. (h) SSFTT. (i) SC-SS-MTr. (j) MST-SSSNet(ours).

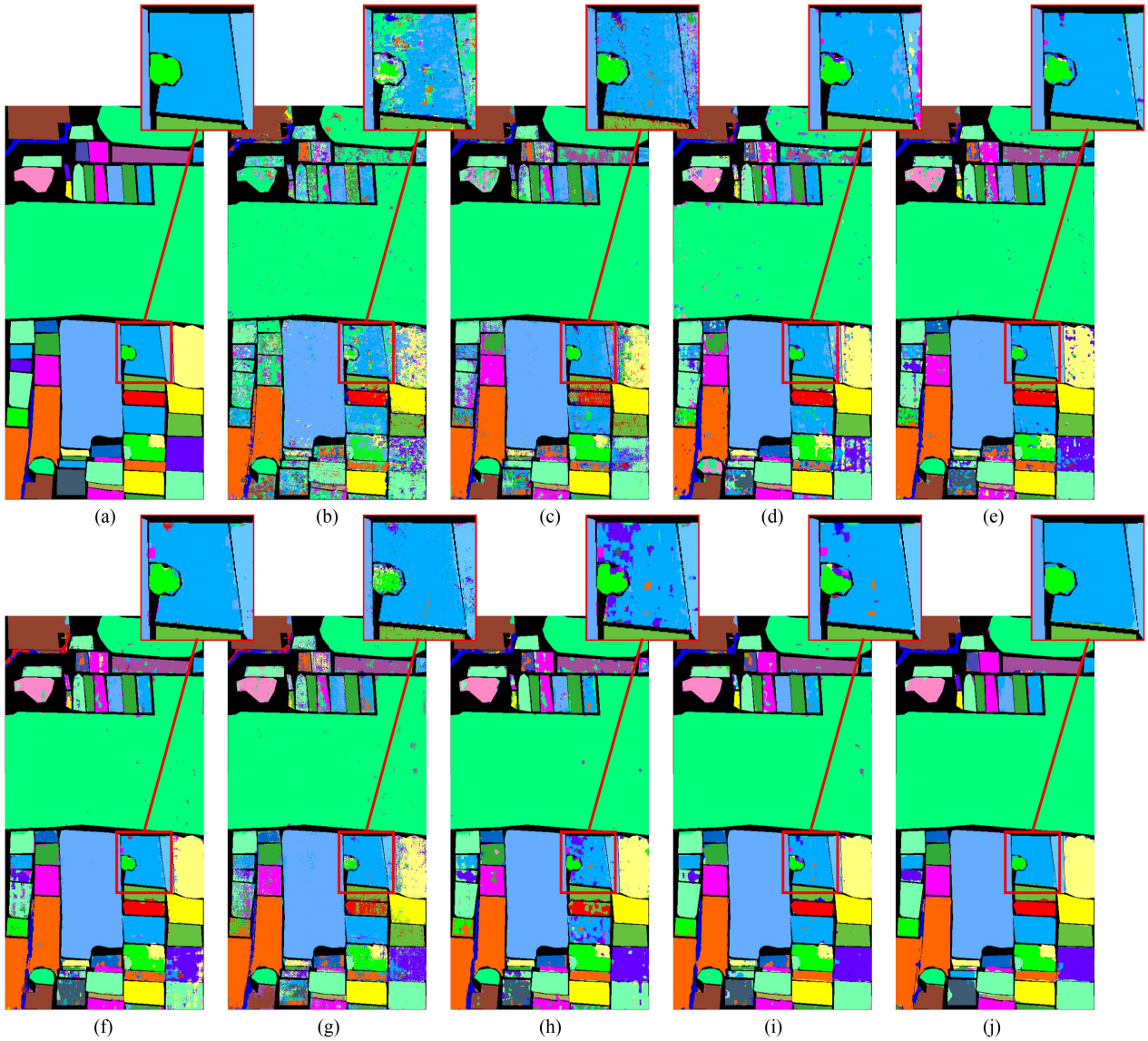

Fig. 12. Visualization of the experimental results based on the HongHu dataset. (a) Ground truth. (b) 1D-CNN. (c) 2D-CNN. (d) 3D-CNN. (e) HybridSN. (f) ViT. (g) SpectralFormer. (h) SSFTT. (i) SC-SS-MTr; (j) MST-SSSNet(ours).

an area where our method is more accurate in distinguishing between the boundaries of the semicircle and the interior. In addition, almost all methods misidentify the gray corner above the semicircle. It proves that our method is more reliable in distinguishing details.

For the HongHu dataset, this dataset has the characteristics of the previous two datasets. There are large areas of the same class, and there are areas where multiple classes gather. At the same time, this dataset has the largest number of samples. Since the partitioning of the training set is random, such a partitioning method is more challenging for the model. In order to pursue the improvement of classification accuracy, the model is easier to overfit a large number of classes. Other methods are more accurate for large areas of light green in the middle, but the area below obviously achieves poor results. The classification results are also circled in an area where it is clear that the proposed method gets results closer to the ground truth map and has better antioverfitting ability for most classes.

The analysis of visual classification results from three datasets demonstrates that the proposed method exhibits relatively stable classification performance in areas with simple and complex categories. However, for boundary areas that are difficult to distinguish but often very important, the proposed method has achieved significant advantages over other methods. This stems from the SSSC module overcoming the drawback of most feature extraction modules focusing too much on spatial or spectral features, as well as the design of the MST module to establish global correlations based on local features. Therefore, the best classification results were achieved.

4) Influence of Sample Number: In addition, we selected 20%, 40%, 60%, 80%, and 100% of the original training samples from three datasets to compare 2D-CNN, 3D-CNN, HybridSN,

TABLE V
TRAIN TIME AND TEST TIME OF DIFFERENT METHODS IN THE LONGKOU DATASET

Methods	2D-CNN	3D-CNN	HybridSN	SSFTT	SC-SS-MTr	MST-SSSNet
Train time(s)	15.32	44.37	32.64	9.20	152.11	15.69
Test time(s)	21.77	53.62	16.95	9.18	217.63	10.72

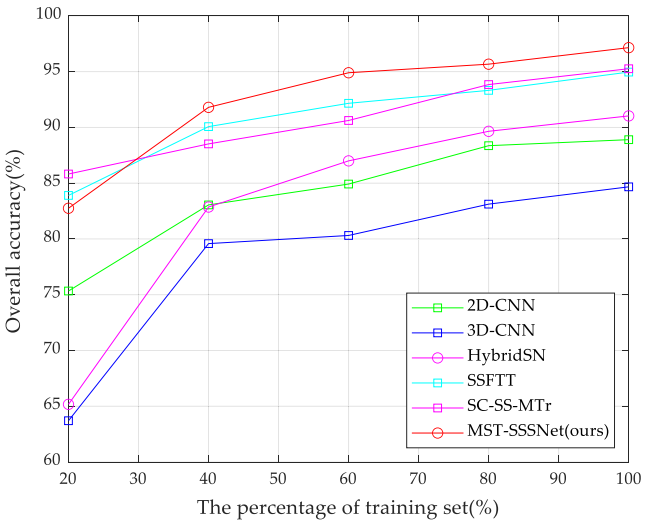

Fig. 13. OA results on the HongHu dataset with a varying number of training samples.

SSFTT, and SC-SS-MTr with our method. The classification outcomes are depicted in Fig. 13. It is evident that as the utilized training samples increase, the overall classification performance, represented by OA, exhibits a gradual improvement. Among them, the proposed MST-SSSNet method performs best in classification performance when the training set proportion is about 30% to 100%, and its performance is equivalent to that of SSFTT and SC-SS-MTr methods when the training set proportion is below 30%, indicating that the effectiveness of our method endures even in situations characterized by inadequate sample sizes.

5) Time Cost Comparison: The train and test time among 2D-CNN, 3D-CNN, HybridSN, SSFTT, SC-SS-MTr, and our method on the LongKou dataset is given in Table V. The SSFTT model has the fastest calculation speed, and the proposed method and 2D-CNN are slightly higher than SSFTT, although SSFTT adopts a 3-D convolution and transformer structure that consumes computational resources. However, due to limited fitting ability, a large patch size is not required. In the original text, the patch size is set to 13, which is much smaller than other methods including the proposed method. This resulted in the method achieving the best training and testing time. However, a too small patch size limits the improvement of classification performance and prevents learning more global category information. The proposed SSSC can reduce parameters and, thus, reduce training time. However, the design of the two transformers in the proposed method, especially the subtransformer, requires multiple calculations of attention and residual connections, which increases the calculation time. After considering the calculation time and classification accuracy, it is worthwhile to increase computational resources. HyBridSN and 3D-CNN both have high parameter count due to their multilayer 3-D convolution, resulting in longer calculation time. SC-SS-MTr has a heavy computational burden due to its two stages and the use of masked transformers. Overall, the proposed MST-SSSNet achieves better classification performance with minimal computing resources. Thus, it has excellent classification efficiency.

V. DISCUSSION

Computational complexity is an important indicator for evaluating models. In this section, we discussed the computational complexity of the proposed module.

We analyzed the computational complexity of the proposed SSSC and the MST module. For the SSSC module, the biggest difference from other feature extraction modules is the use of the proposed SSSC. Let us assume that the total number of pixels in the input 3-D feature map is P, K is the convolutional kernel size, and C_l represents the number of output channels in the lth layer. The algorithm complexity of SSSC can be expressed as

$$O\left(P \cdot K^2 \cdot 2 \cdot \sum_{i=0}^{l-1} C_i \cdot C_{i+1}\right). \tag{17}$$

The complexity of most existing algorithms using 3-D convolution is

$$O\left(P \cdot K^3 \cdot \sum_{i=0}^{l-1} C_i \cdot C_{i+1}\right). \tag{18}$$

It can be seen that the proposed SSSC outperforms 3-D convolution in terms of algorithm complexity.

For the MST module, the main computational resource is consumed in the computation of multihead attention. If C represents the number of channels, N_q and N_k represent the number of elements in query and key, respectively. The computational complexity of MSA can be expressed as

$$O\left(N_q C^2 + vC^2 + N_q N_k C\right). \tag{19}$$

Let the token generator generate n tokens of length d. Each patch token is divided into s^2 subtokens. The computational complexity of MSA in the MST module can further be expressed as

$$O\left(n^2 d + (sn)^2 d/n\right) = O\left(5n^2 d\right). \tag{20}$$

Therefore, the subtransformer consumes the main computing resources. For the entire MST, the computational complexity depends on the number and length of tokens generated.

VI. Conclusion

This article proposes an MST-SSSNet HSI classification method, which achieves excellent classification performance through the SSSC module and MST module. The SSSC module is a lightweight feature extraction module constructed from the proposed SSSC, which can accurately and effectively extract local spectral–spatial features. The MST module uses subtransformers to learn local details to help the main transformer establish global correlation, thereby enhancing the coupling between local information and global information. On three representative datasets, we compared state-of-the-art methods. Based on the results of experiments, our MST-SSSNet has the best classification performance compared with other methods. In the future, the next step is to study MST-SSSNet-based unsupervised classification methods in response to the scarcity of hyperspectral samples. And we hope to introduce the physical features of spectral bands and prior knowledge of HSI into the model to improve classification performance. In addition, we will conduct more in-depth research on the problem that PCA may not fully utilize HSI information.

Acknowledgment

The authors would like to thank Y. Zhong's team at Wuhan University, China, for collecting the Wuhan HSI dataset, as well as the researchers who shared their source codes in the community.

Reference

[1] M. Wang et al., "Tensor decompositions for hyperspectral data processing in remote sensing: A comprehensive review," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 1, pp. 26–72, Mar. 2023, doi: 10.1109/MGRS.2022.3227063.

[2] N. Audebert, B. L. Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019, doi: 10.1109/MGRS.2019.2912563.

[3] C. M. Gevaert, J. Suomalainen, J. Tang, and L. Kooistra, "Generation of spectral–temporal response surfaces by combining multispectral satellite and hyperspectral UAV imagery for precision agriculture applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3140–3146, Jun. 2015, doi: 10.1109/JSTARS.2015.2406339.

[4] Q. Hao et al., "Fusing multiple deep models for in vivo human brain hyperspectral image classification to identify glioblastoma tumor," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 4007314, doi: 10.1109/TIM.2021.3117634.

[5] M. Shimoni, R. Haelterman, and C. Perneel, "Hyperspectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019, doi: 10.1109/MGRS.2019.2902525.

[6] S. Feng, S. Tang, C. Zhao, and Y. Cui, "A hyperspectral anomaly detection method based on low-rank and sparse decomposition with density peak guided collaborative representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5501513, doi: 10.1109/TGRS.2021.3054736.

[7] J. Wang, Z. Li, J. Yang, S. Liu, J. Zhang, and S. Li, "A multilevel spatial and spectral feature extraction network for marine oil spill monitoring using airborne hyperspectral image," *Remote Sens.*, vol. 15, no. 5, Feb. 2023, Art. no. 1302, doi: 10.3390/rs15051302.

[8] A. Hamedianfar, K. Laakso, M. Middleton, T. Törmänen, J. Köykkä, and J. Torppa, "Leveraging high-resolution long-wave infrared hyperspectral laboratory imaging data for mineral identification using machine learning methods," *Remote Sens.*, vol. 15, no. 19, Oct. 2023, Art. no. 4806, doi: 10.3390/rs15194806.

[9] J. Atli Benediktsson and P. Ghamisi, *Spectral-Spatial Classification of Hyperspectral Remote Sensing Images*. Norwood, MA, USA: Artech House, 2015.

[10] 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, doi: 10.1109/MGRS.2013.2244672.

[11] P. Ghamisi et al., "New frontiers in spectral-spatial hyperspectral image classification: The latest advances based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 3, pp. 10–43, Sep. 2018, doi: 10.1109/MGRS.2018.2854840.

[12] C. Zhao, B. Qin, S. Feng, W. Zhu, L. Zhang, and J. Ren, "An unsupervised domain adaptation method towards multi-level features and decision boundaries for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5546216, doi: 10.1109/TGRS.2022.3230378.

[13] 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, doi: 10.1109/TGRS.2004.831865.

[14] Q. Ye et al., "L1-norm distance minimization-based fast robust twin support vector plane clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4494–4503, Sep. 2018, doi: 10.1109/TNNLS.2017.2749428.

[15] L. Samaniego, A. Bardossy, and K. Schulz, "Supervised classification of remotely sensed imagery using a modified k-NN technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2112–2125, Jul. 2008, doi: 10.1109/TGRS.2008.916629.

[16] W. Li, Q. Du, F. Zhang, and W. Hu, "Collaborative-representation-based nearest neighbor classifier for hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 2, pp. 389–393, Feb. 2015, doi: 10.1109/LGRS.2014.2343956.

[17] Y. Gu, J. Chanussot, X. Jia, and J. A. Benediktsson, "Multiple kernel learning for hyperspectral image classification: A review," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6547–6565, Nov. 2017, doi: 10.1109/TGRS.2017.2729882.

[18] Y. E. SahIn, S. Arisoy, and K. Kayabol, "Anomaly detection with Bayesian Gauss background model in hyperspectral images," in *Proc. 26th Signal Process. Commun. Appl. Conf.*, 2018, pp. 1–4, doi: 10.1109/SIU.2018.8404293.

[19] W. Liu, J. E. Fowler, and C. Zhao, "Spatial logistic regression for support-vector classification of hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 3, pp. 439–443, Mar. 2017, doi: 10.1109/LGRS.2017.2648515.

[20] 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, doi: 10.1109/TGRS.2008.922034.

[21] L. Sun, C. Ma, Y. Chen, H. J. Shim, Z. Wu, and B. Jeon, "Adjacent superpixel-based multiscale spatial-spectral kernel for hyperspectral classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1905–1919, Jun. 2019, doi: 10.1109/JSTARS.2019.2915588.

[22] B. Rasti et al., "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020, doi: 10.1109/MGRS.2020.2979764.

[23] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014, doi: 10.1109/JSTARS.2014.2329330.

[24] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015, doi: 10.1109/JSTARS.2015.2388577.

[25] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016, doi: 10.1109/TPAMI.2015.2439281.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[27] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, May 2016, doi: 10.1109/TMI.2016.2535865.

[28] 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, 2015.

[29] 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.
[30] 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.
[31] 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.
[32] 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.
[33] Y. Ding et al., "AF2GNN: Graph convolution with adaptive filters and aggregator fusion for hyperspectral image classification," *Inf. Sci.*, vol. 602, pp. 201–219, 2022.
[34] Y. Ding et al., "Multi-feature fusion: Graph neural network and CNN combining for hyperspectral image classification," *Neurocomputing*, vol. 501, pp. 246–257, 2022.
[35] Y. Ding et al., "Multi-scale receptive fields: Graph attention neural network for hyperspectral image classification," *Expert Syst. Appl.*, vol. 223, 2023, Art. no. 119858.
[36] W. Zhu, C. Zhao, S. Feng, and B. Qin, "Multiscale short and long range graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535815, doi: 10.1109/TGRS.2022.3199467.
[37] R. Lei et al., "Multiscale feature aggregation capsule neural network for hyperspectral remote sensing image classification," *Remote Sens.*, vol. 14, no. 7, Mar. 2022, Art. no. 1652, doi: 10.3390/rs14071652.
[38] M. E. Paoletti, S. Moreno-Álvarez, and J. M. Haut, "Multiple attention-guided capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5520420, doi: 10.1109/TGRS.2021.3135506.
[39] Y. Zhang, 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.
[40] C. Sun, X. Zhang, H. Meng, X. Cao, and J. Zhang, "AC-WGAN-GP: Generating labeled samples for improving hyperspectral image classification with small-samples," *Remote Sens.*, vol. 14, no. 19, Oct. 2022, Art. no. 4910, doi: 10.3390/rs14194910.
[41] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representions (ICLR)*, 2021, pp. 1–21.
[42] D. Hong et al., "Spectral former: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615, doi: 10.1109/TGRS.2021.3130716.
[43] L. Huang, Y. Chen, and X. He, "Spectral–spatial masked transformer with supervised and contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5508718, doi: 10.1109/TGRS.2023.3264235.
[44] X. Huang, Y. Zhou, X. Yang, X. Zhu, and K. Wang, "SS-TMNet: Spatial–spectral transformer network with multi-scale convolution for hyperspectral image classification," *Remote Sens.*, vol. 15, no. 5, Feb. 2023, Art. no. 1206, doi: 10.3390/rs15051206.
[45] X. Qiao, S. K. Roy, and W. Huang, "Multiscale neighborhood attention transformer with optimized spatial pattern for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5523815, doi: 10.1109/TGRS.2023.3314550.
[46] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5522214, doi: 10.1109/TGRS.2022.3144158.
[47] 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.
[48] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.
[49] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial–spectral fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2485–2501, 2020, doi: 10.1109/JSTARS.2020.2983224.
[50] H. Yu, H. Zhang, Y. Liu, K. Zheng, Z. Xu, and C. Xiao, "Dual-channel convolution network with image-based global learning framework for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6005705, doi: 10.1109/LGRS.2021.3139358.
[51] C. Yu, R. Han, M. Song, C. Liu, and C. I. Chang, "Feedback attention-based dense CNN for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5501916, doi: 10.1109/TGRS.2021.3058549.
[52] S. Hao, Y. Xia, and Y. Ye, "Generative adversarial network with transformer for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5510205, doi: 10.1109/LGRS.2023.3322139.
[53] J. Bai et al., "Hyperspectral image classification based on multibranch attention transformer networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535317, doi: 10.1109/TGRS.2022.3196661.
[54] C. Zhao et al., "Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning," *IEEE Trans. Image Process.*, vol. 32, pp. 3606–3621, 2023, doi: 10.1109/TIP.2023.3287738.
[55] W. Zhou, S.-I. Kamata, H. Wang, and X. Xue, "Multiscanning-based RNN–Transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5512319, doi: 10.1109/TGRS.2023.3277014.
[56] 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.
[57] X. Cao, H. Lin, S. Guo, T. Xiong, and L. Jiao, "Transformer-based masked autoencoder with contrastive loss for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5524312, doi: 10.1109/TGRS.2023.3315678.
[58] H. Yu, Z. Xu, K. Zheng, D. Hong, H. Yang, and M. Song, "MSTNet: A multilevel spectral–spatial transformer network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5532513, doi: 10.1109/TGRS.2022.3186400.
[59] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, Jan. 2021, Art. no. 498, doi: 10.3390/rs13030498.
[60] H. Yang, H. Yu, K. Zheng, J. Hu, T. Tao, and Q. Zhang, "Hyperspectral image classification based on interactive transformer and CNN with multilevel feature fusion network," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5507905, doi: 10.1109/LGRS.2023.3303008.
[61] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, and H. Zhou, "Spatial sequential recurrent neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4141–4155, Nov. 2018.
[62] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Scalable recurrent neural network for hyperspectral image classification," *J. Supercomputing*, vol. 76, no. 11, pp. 8866–8882, Feb. 2020.
[63] W. Zhou, S. I. Kamata, Z. Luo, and H. Wang, "Multiscanning strategy based recurrent neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5521018.
[64] Y. Ding et al., "Self-supervised locality preserving low-pass graph convolutional embedding for large-scale hyperspectral image clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536016.
[65] Z. Zhang et al., "Multireceptive field: An adaptive path aggregation graph neural framework for hyperspectral image classification," *Expert Syst. Appl.*, vol. 217, Art. no. 119508, 2023.
[66] J. Feng, Z. Gao, R. Shang, X. Zhang, and L. Jiao, "Multi-complementary generative adversarial networks with contrastive learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5520018, doi: 10.1109/TGRS.2023.3304836.
[67] S. Christian, V. Vincent, I. Sergey, S. Jonathon, and W. Zbigniew, "Rethinking the inception architecture for computer vision," 2015, *arXiv:1512.00567v3*.
[68] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," *Remote Sens. Environ.*, vol. 250, 2020, Art. no. 112012.
[69] Y. Zhong et al., "Mini-UAV-borne hyperspectral remote sensing: From observation and processing to applications," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 4, pp. 46–62, Dec. 2018.
[70] L. Fang, N. He, S. Li, P. Ghamisi, and J. A. Benediktsson, "Extinction profiles fusion for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1803–1815, Mar. 2018, doi: 10.1109/TGRS.2017.2768479.

Jingpeng Gao (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical information engineering from Harbin Engineering University, Harbin, China, in 2002, 2007, and 2014, respectively.

Since 2002, he has been with Harbin Engineering University, and became a Lecturer in 2007, and a Master Tutor in 2015. From 2015 to 2017, he was with the State Key Laboratory of Computational Mathematical and Experimental Physics, Beijing Institute of Space Long March Vehicle, as a Postdoctoral Researcher. His research interests include machine learning, radar target recognition, and hyperspectral image processing.

Xiangyu Ji (Graduate Student Member, IEEE) was born in 2001. He received the B.S. degree in electronic engineering from Xidian University, Xi'an, China, in 2023. He is currently working toward the M.S. degree in electronic information with the College of Information and Communication Engineering, Harbin Engineering University, Harbin, China.

His research interests include deep learning and hyperspectral image processing.

Geng Chen received the B.S. degree in electronic and information engineering from the Shandong University of Science and Technology, Qingdao, China, in 2022. He is currently working toward the M.S. degree in electronic and information engineering with Harbin Engineering University, Harbin, China.

His main research interests include machine learning and hyperspectral image processing.

Ruitong Guo was born in 2000. She received the B.S. degree in electronic engineering from Northeast Petroleum University, DaQing, China, in 2022. She is currently working toward the M.S. degree in information and communication engineering with Harbin Engineering University, Harbin, China.

Her research interests include remote sensing image processing and machine learning.