A Spectral-Spatial Fusion Transformer Network for Hyperspectral Image Classification

Diling Liao, Cuiping Shi, *Member, IEEE*, Liguo Wang, *Member, IEEE*

***Abstract*—In the past, deep learning (DL) technologies have been widely used in hyperspectral image classification tasks. Among them, convolutional neural networks (CNNs) use fixed size receptive field (RF) to obtain spectral and spatial features of hyperspectral images (HSIs), showing great feature extraction capabilities, which are one of the most popular DL frameworks. However, the convolution using local extraction and global parameter sharing mechanism pays more attention to spatial content information, which changes the spectral sequence information in the learned features. In addition, CNN is difficult to describe the long-distance correlation between HSI pixels and bands. To solve these problems, a spectral-spatial fusion Transformer network (S^2FTNet) is proposed for the classification of hyperspectral images. Specifically, S^2FTNet adopts the Transformer framework to build a spatial Transformer module (SpaFormer) and a spectral Transformer module (SpeFormer) to capture image spatial and spectral long-distance dependencies. In addition, an adaptive spectral-spatial fusion mechanism (AS^2FM) is proposed to effectively fuse the obtained advanced high-level semantic features. Finally, a large number of experiments were carried out on four datasets, Indian Pines, Pavia, Salinas and WHU-Hi-LongKou, which verified that the proposed S^2FTNet can provide better classification performance than other the state-of-the-art networks.**

***Index Terms*—Deep learning, long-distance dependence, fusion, hyperspectral image.**

I. Introduction

Hyperspectral Images (HSIs) are captured by airborne imaging spectrometer and carries a lot of spectral and spatial information. In recent years, HSIs have played an important role in many fields, including health care [1], military [2], earth exploration [3], environmental protection [4], etc. Among them, hyperspectral image classification is an important stage of hyperspectral image processing, and is one of the hot spots of image research. Specifically, hyperspectral image classification is to classify images pixel by pixel by learning prior knowledge [5] - [7].

In the early stage of research, classification methods paid more attention to the spectral feature extraction of images, and many classical methods appeared, such as Support Vector Machines (SVM) [8], Principal Component Analysis (PCA) [9] and Composite Kernels [10]. Although the above traditional methods can obtain the basic features of the image, the classification performance is not satisfactory. In addition, these methods have many disadvantages. For example, too much dependence on knowledge in professional fields, low generalization ability, and weak representation ability of acquired features. Therefore, the deep learning (DL) technology is becoming more and more popular in computer vision tasks (such as classification [11] - [13], detection [14] [15], segmentation [16], etc.), because it can not only get rid of the constraints of manual, but also adaptively learn high-level semantic information.

In recent years, many excellent frameworks have emerged for DL technology, including convolutional neural networks (CNNs) [17], generative adversarial networks (GANs) [18] [19], recurrent neural networks (RNNs) [20] [21], graph convolutional networks (GCNs) [22] [23], capsule network (CapsNet) [24], and vision Transformer (ViT) [25].

Among them, CNNs are one of the most popular DL methods, which improve the discriminative ability of features through local connection and global parameter sharing mechanism. Unlike other ordinary images, HSI contains rich spectral and spatial features, and the construction of CNN network can easily extract these two features of HSI. In [26], Hu *et al.* used 1D-CNN to classify HSI pixel by pixel, and verified that 1D-CNN is suitable for hyperspectral image classification tasks. In addition, the image has rich spatial information. In order to integrate the spatial information of the image, [27] proposed 2D-CNN, which uses adjacent pixels around the central classification pixel as training samples to perform classification tasks, improving the classification performance. However, only using 2D-CNN is not enough to extract spectral-spatial joint features of images. Therefore, Hamida *et al.* [28] cut the HSI into multiple 3D cubes and constructed 3D CNN to extract the spectral-spatial joint features of the image, verifying that the method can effectively improve the classification performance. Similarly, Roy *et al.* [29] designed a spectral-spatial hybrid network based on 3D-CNN and 2D-CNN, and proved its effectiveness. In [30], Shang *et al.* proposed a classification method based on multi-scale cross-branch response and second-order channel attention (MCRSCA), which takes into

Manuscript received January 01, 2023. This work was supported in part by the National Natural Science Foundation of China (42271409 and 62071084), in part by the Heilongjiang Science Foundation Project of China under Grant LH2021D022, and in part by the Fundamental Research Funds in Heilongjiang Provincial Universities of China under Grant 145209149.

Cuiping Shi is with the Department of Communication Engineering, Qiqihar university, Qiqihar 161000, China. (e-mail: shicuiping@qqhru.edu.cn).

Diling Liao is with the Department of Communication Engineering, Qiqihar university, Qiqihar 161000, China (e-mail: 2020910228@qqhru.edu.cn)

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

Corresponding author: Cuiping Shi (shicuiping@qqhru.edu.cn).

account the inherent spatial structure information of ground objects and avoids the loss of spatial details. With the gradual increase of 3D-CNN network depth, gradient disappearance and gradient explosion will occur [31], and the classification accuracy will gradually decrease. In order to solve this problem, Zhong *et al.* [32] introduced ResNet [33] structure into the designed spatial 3D CNN module and spectral 3D CNN module, and extracted rich spatial and spectral features. In addition, Paoletti *et al.* [34] proposed a depth pyramid residual network for spectral-spatial hyperspectral image classification by making better use of the potential of available information on each unit. In order to further improve the classification performance and alleviate the problem of over fitting, attention mechanism has been widely concerned and successfully applied in hyperspectral image classification [35] - [38]. For example, He *et al*. [39] proposed a dual global-local attention network (DGLANet). In order to reduce the spatial and spectral redundancy information of pixels, Mei *et al.* proposed a network based on bidirectional long short-term memory (Bi-LSTM) in [40], which designed a spectral-spatial attention mechanism and emphasized effective information. In addition, lightweight classification methods based on CNN are also popular. For example, a lightweight network [41] is constructed by using 3-D depthwise convolution, which reduces model parameters and computational overhead. In [42], Meng *et al.* proposed a lightweight spectral-spatial convolution module (LS^2CM) as an alternative to convolution layer. In [43], Kang *et al.* proposed a spectral spatial classification framework based on edge preserving filtering (EPF). Zhong *et al.* [44] designed an iterative edge preserving filtering (IEPF) method based on EPF and further improved classification performance. In addition, they also embedded an iterative strategy into Spectral Spatial (SS) classifiers and designed a new hyperspectral image classification method that combines multiple SS classifiers [45].

In the past, Transformer has received extensive attention in the field of natural language processing (NLP). It is worth noting that Transformer has recently been introduced into computer vision and successfully applied to image classification tasks [25]. Since the spectrum of HSIs are sequence data and usually contain hundreds of wavebands, He *et al.* [46] proposed a spatial-spectral Transformer (SST) network by combining transfer learning with the Transformer framework, and proved that the Transformer can construct the correlation of spectral sequences. Similarly, Hong *et al.* [47] reconsidered Transformer from the perspective of spectral sequence attributes, proposed a spectral Transformer (SF) network, and confirmed that it has more significant advantages than classical ViT and advanced backbone networks. In general, CNN based network access to high-level semantic features is relatively limited. Therefore, Sun *et al.* [48] proposed a spectral-spatial feature tokenization Transformer (SSFTT) network to capture spectral-spatial features and advanced semantic features. Similarly, Zhong *et al.* [49] proposed a new spectral-spatial Transformer network (SSTN) to overcome the weak ability of CNNs to learn long-distance dependencies. In [50], Huang *et al.* proposed a new 3D-swin Transformer-based hierarchical contrastive learning (3DSwinT-HCL) method based on 3D swin Transformer. This method uses Transformer to effectively make up the shortcomings of CNNs lack of receptive field and inability to capture the order attribute of data. In order to solve the problem that the network is easily interfered by irrelevant information around the target pixel in the training phase, which leads to inaccurate feature extraction, Bai *et al.* [51] proposed a hyperspectral image classification method based on the multi branch attention Transformer network. In [52], Zou *et al.* proposed the local enhanced spectral-spatial Transformer (LESSForm) method, which alleviates the problem that Transformer based classification methods usually generate inaccurate tag embedding from a single spectral or spatial dimension of the original HSI. Inspired by the bottleneck Transformer of computer vision, Song *et al.* [53] proposed a bottleneck spatial-spectral Transformer (BS2T) network, which uses Transformer to make the extracted features more spatial location aware and spectral aware. In [54], Mei *et al.* proposed a Group-Aware Hierarchical Transformer (GAHT) to solve the problem of over dispersion of features extracted by multi head self-attention (MHSA) in the Transformer.

Although the above DL methods have been widely used in hyperspectral image classification, there are still some challenges. On the one hand, CNNs using the mechanism of local extraction and global parameter sharing pay more attention to spatial content information, thus distorting the spectrum sequence information in the learning features [52]. On the other hand, CNNs are difficult to describe the long-distance correlation between HSI pixels and bands. On the contrary, Transformer can not only effectively extract long-distance dependence, but well maintain spectral sequence information. However, although existing Transformer methods can bring advantages to modeling long-distance features, the extraction of long-distance features using MHSA is not sufficient. Therefore, this paper proposed a hyperspectral image classification method based on spectral-spatial fusion Transformer network (S^2FTNet). In particular, we propose a multi head dual self-attention (MHD-SA) and use it to replace MHSA in the Transformer module, constructing a spatial Transformer module (SpaFormer). Then, the MHD-SA and convolution are combined to construct a spectral Transformer module (SpeFormer). These two modules are respectively used to enhance the capture ability of long-distance features in spatial and spectral bands. In addition, an adaptive spectral-spatial fusion mechanism (AS^2FM) is proposed to effectively combine the obtained spectral-spatial high-level semantic features.

The main contributions of this paper are as follows:

1) In order to enhance the long-distance dependency of features and improve the representation ability of features, a Transformer block based on multi head double self-attention (MHD-SA) is proposed. Then, three improved Transformer blocks are constructed in parallel as a spatial Transformer module (SpaFormer) to extract the long-distance dependence of images with different spatial dimensions.

2) In order to increase the receptive field of spectral

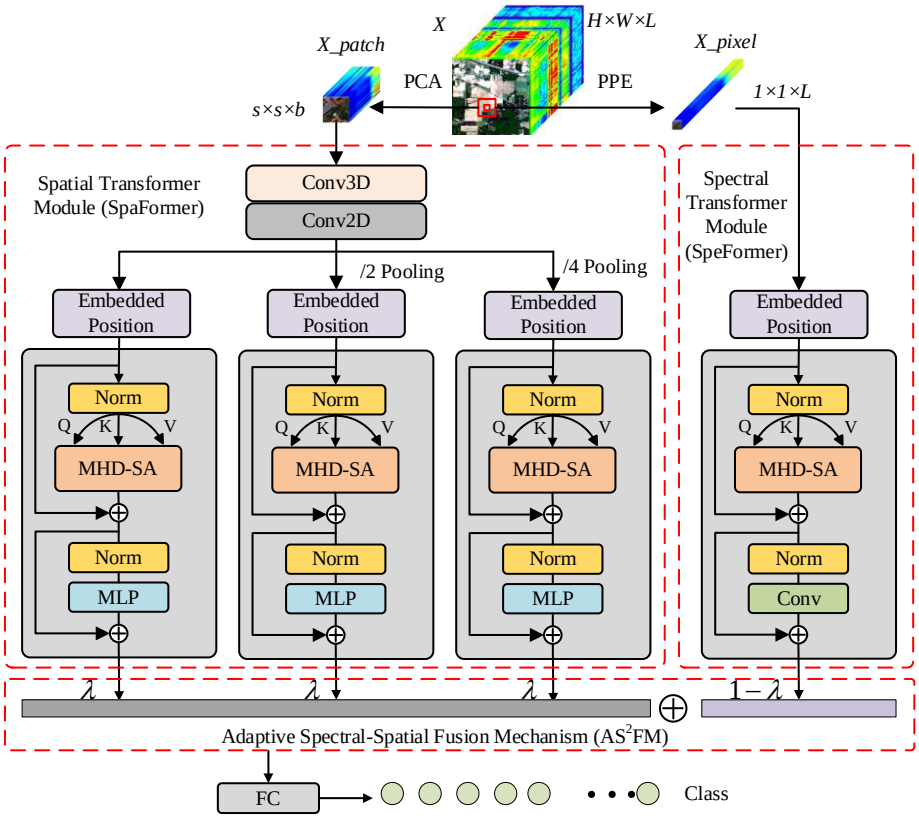


Fig. 1. S²FTNet overall network framework.

extraction and learn more spectral sequence information, a spectral Transformer module (SpeFormer) is designed. It uses convolution to replace the traditional Transformer's MLP, and combines it with the proposed MHD-SA.

3) Considering the different importance of high-level semantic features extracted by spatial branches and spectral branches, in order to combine them more effectively, an adaptive spectral-spatial fusion mechanism (AS²FM) is proposed.

4) Based on Transformer and CNN, we proposed a spectral-spatial fusion Transformer network (S²FTNet), which uses a dual branch structure to extract spectral and spatial features respectively, and combines the features obtained from the two branches with adaptive fusion mechanism. Extensive experiments have proved that our method has better performance and potential compared with some state-of-the-art CNN-based and Transformer networks.

The rest of this paper is arranged as follows. In the second part, the network structure of S²FTNet proposed in this paper is introduced in detail. In the third part, the parameter analysis of the model, quantitative analysis of comparative experiments and visual evaluation are provided. The fourth part gives the conclusion and prospect of the paper

II. Methodology

The method S²FTNet proposed in this paper includes three main modules, SpaFormer, SpeFormer, and AS²FM respectively. The overall network framework is shown in Fig. 1. Suppose input HSI data is $X \in \mathbb{R}^{H\times W\times L}$. Where, W and H represent the width and height of the image, L represents the number of bands of the image, and the corresponding label set $Y_i \in \{1,2,...,Class\}$. In order to facilitate feature extraction, X is first processed by edge filling strategy. Then, the new data obtained after filling is extracted in two ways. One is to extract the adjacent edge blocks of the pixel with the pixel to be classified as the center and reduce the spectral dimension by principal component analysis (PCA) to obtain data $X\_patch \in \mathbb{R}^{s\times s\times b}$. The other is pixel by pixel extraction (PPE) to obtain data $X\_pixel \in \mathbb{R}^{1\times 1\times L}$. Where, $s\times s$ represents the image space size after segmentation, and b represents the number of spectral bands after PCA dimensionality reduction. Next, the two processed data are used as the input data of SpaFormer and SpeFormer modules respectively, and the advanced semantic features extracted by the two modules are fused through an adaptive fuse mechanism. Finally, the fused feature vectors are transferred to the classifier for classification.

Then, the three main modules of S²FTNet proposed in this paper are introduced in detail.

A. Spatial Transformer Module (SpaFormer)

In recent years, CNNs are one of the most classical deep learning frameworks, and are widely used in hyperspectral image classification tasks. Convolution (Conv) of CNN uses a mechanism of local connection and global parameter sharing, so that more attention is paid to the local features of the image during the feature extraction process. In contrast to Conv, Transformer can build long-distance dependencies, making up

the shortcomings of Conv in feature extraction. Therefore, the SpaFormer uses the above two frameworks for modeling, the structure is shown in Fig. 1. Next, this section will introduce the proposed SpaFormer module in detail.

First, the input image data X_patch passes through two Conv blocks, namely 3D Convolution (Conv3D) and 2D Convolution (Conv2D), and each Conv block contains convolution layer, batch normalization layer and nonlinear activation layer. Specifically, X_patch extracts the spectral-spatial joint information of the image through Conv3D, and the calculation process is

$$F_{3D} = f\left(\delta_1\left(X\_patch \Theta w^{3D} + b^{3D}\right)\right) \tag{1}$$

In Formula (1), w^{3D} represents the weight offset of 3-D Conv, b^{3D} represents the offset term, and F_{3D} represents the output of Conv3D. Θ is a 3-D Conv operator, δ_1 is a 3-D batch normalization operation, and $f(\cdot)$ is a nonlinear activation function ReLU. In order to further extract image spatial information, the module introduces Conv2D after Conv3D. The calculation principle of Conv2D is similar to that of Conv3D, and the formula is

$$F_{2D} = f\left(\delta_2\left(F_{3D} \odot w^{2D} + b^{2D}\right)\right) \tag{2}$$

In Formula (2), w^{2D} represents the weight offset of 2-D Conv, F_{2D} represents the offset term, and b^{2D} represents the output of Conv2D. $\odot$ is a 2-D Conv operator, and δ_2 is a 2-D batch normalization operation. The module first extracts the spectral-spatial joint and spatial features of the image by designing Conv3D and Conv2D, which provided complete shallow information for extracting high-level semantic features.

Then, three improved Transformer blocks are used for parallel connection to build the SpaFormer module, which is used to explore the long-distance dependency of images. As can be seen from Fig. 1, each Transformer block contains multiple components, including position embedding (PE), two layers of normalization (Norm), multi head double self-attention (MHD-SA) and multilayer perceptron (MLP).

To strengthen the correlation between positions, the Transformer block first introduced PE. To put it simply, all tokens $T = [T_1, T_2, ..., T_w]$ are connected to the learnable classification token T_0, and the location information PE_{pos} is attached to all tokens, that is

$$T_{PE} = [T_0, T_1, T_2, ..., T_w] + PE_{pos} \tag{3}$$

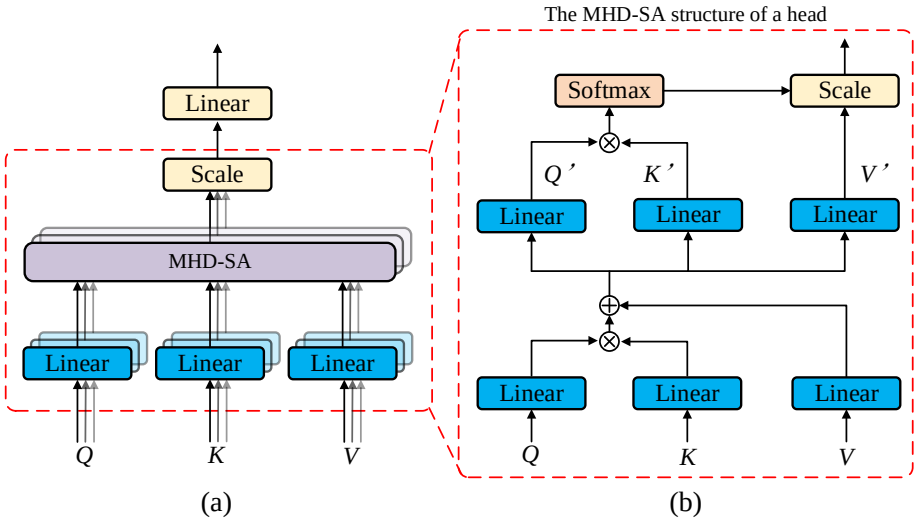


Fig. 2. The overall structure of MHD-SA. (a) Multi head structure, (b) single head structure.

The proposed MHD-SA is the most important component of the entire Transformer, and its structure is shown in Fig. 2 (a). At the same time, for the convenience of illustration, the single head structure of MHD-SA is shown in Fig. 2 (b). MHD-SA usually contains three feature inputs, namely Query (Q), Key (K) and Value (V), and Q, K and V are obtained by linear mapping of three predefined weight matrices W_Q, W_K and W_V. The self-attention score of single headed double self-attention (DSA) is calculated by Q and K, and then the score is weighted into V, that is

$$SA = soft\max\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{4}$$

$$DSA = soft\max\left(\frac{L_Q(SA)L_K(SA)}{\sqrt{d_{L_K}}}\right)L_V(SA) \tag{5}$$

In the above formula, SA represents the self-attention value,

$L_Q(\cdot)$, $L_K(\cdot)$ and $L_V(\cdot)$ represent the features obtained by SA through linear mapping. d_K and d_{L_K} represent the feature dimensions of K and L_K respectively. Generally, Transformer contains multiple head self-attention, so the MHD-SA can be represented as

$$MHD-SA = Concat\left(DSA_1, DSA_2, ..., DSA_h\right)W \tag{6}$$

where $Concat(\cdot)$ represents the cascade function, h represents the number of headers, and W represents the weight parameter.

Finally, MLP is introduced after MHD-SA to alleviate the problem of gradient explosion and gradient disappearance. The MLP structure contains two full connection layers, and a gaussian error linear unit (GELU) is embedded between the two full connection layers.

It is worth noting that SpaFormer contains three improved Transformer blocks. Although the three Transformer blocks have the same structure, the input data is different. It can be seen from Fig. 1 that the space size $s \times s$ of the input data of the three blocks performs $pooling = false$, $pooling = 2$, and $pooling = 4$ operations respectively, and the output space size is $[s / pooling] \times [s / pooling]$, while $[\cdot]$ represents the upper rounding symbol. With different space sizes, Transformer blocks can be used to explore long-distance dependencies of different spaces, which can enrich the diversity of features.

To sum up, the spatial branch contains two Conv blocks and SpaFormer modules. First, the spectral-spatial joint and spatial features of the shallow layer are extracted through two Conv blocks to provide complete shallow information. Then, Three Transformer blocks are utilized to explore long-distance dependencies of data with different input spatial sizes, which enriches the diversity of features.

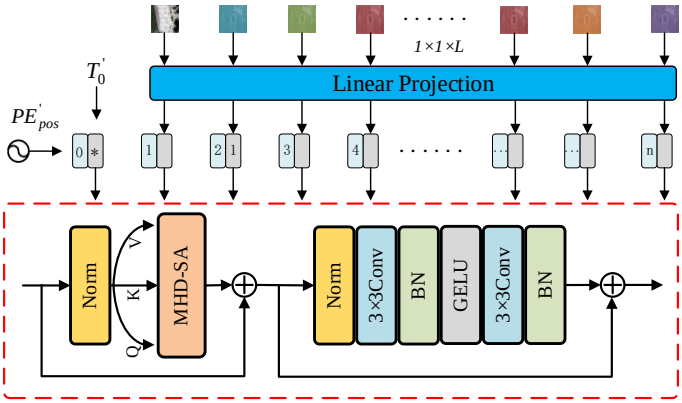


Fig. 3. The overall structure of SpeFormer.

B. Spectral Transformer Module (SpeFormer)

HSI not only has rich spatial information, but also contains hundreds of spectral bands. Extracting rich spectral features of images and taking full account of spectral sequence can improve the discrimination ability of features and classification performance. Therefore, inspired by [48], this paper proposed a spectral Transformer module (SpeFormer). The overall structure is shown in Fig. 3.

It can be seen that the input data size is $\mathbb{R}^{1\times1\times L}$ and L is the number of spectral bands of HSI. First, the input is dimensionally reduced by linear mapping and cascaded with learnable token $T_0^{'}$. Then, the results are embedded in position, and the obtained feature tensor $T_{PE}^{'}$ contains position and spectral order information. The calculation process is similar to the SpaFormer, that is

$$T_{PE}^{'} = \left[T_0^{'}, T_1^{'}, T_2^{'}, ..., T_w^{'}\right] + PE_{pos}^{'} \tag{7}$$

Then, Transformer block based on Conv is introduced, which fully considers the correlation between spectral sequences and can obtain the long-distance dependence between spectral. The traditional MLP of Transformer includes two fully connected layers (FC). Although the two layers of FC can extract spectral nonlinear features to a certain extent, it still lacks consideration of local spectral correlation. According to [55], the linear transformation at different positions in two FCs of the Transformer block is the same, but they use different parameters from one layer to another, which can be replaced by two 1×1 Conv. Therefore, in order to further explore the local spectral correlation and increase the convolution receptive field, SpeFormer uses two 3×3 Conv blocks (including a Conv layer and a batch normalization layer BN) to replace FC in the traditional MLP block. This improved method can effectively increase the RF of spectral information extraction, while avoiding the destruction of spectral order. Therefore, the improved Transformer block includes two layer normalization, a multi-head double self-attention (MHD-SA), two Conv blocks and a GELU. This process can be expressed as

$$SpeFormer = \delta_2\left(f_2\left(g\left(\delta_1\left(f_1\left(MHD-SA\right)\right)\right)\right)\right) \tag{8}$$

In Formula (8), $f(\cdot)$ represents Conv function, $\delta(\cdot)$ represents BN function, and *SpeFormer* represents the output result of improved Transformer block.

To sum up, SpeFormer is used to explore the long-distance dependence between spectral bands of images, fully considering spectral sequence, which provides more assistance for long-distance features of spectral bands.

C. Adaptive Spectral Spatial Fusion Mechanism (AS^{2}FM)

In this paper, the proposed S^{2}FTNet selects cross entropy as the loss function, and optimizes the network through back propagation. Where, the expression of cross entropy loss function is

$$Loss = \frac{1}{C}\sum_{a=1}^{r}\left[-y_a^{'}\log\left(y_a\right) - \left(1-y_a^{'}\right)\log\left(1-y_a\right)\right] \tag{9}$$

In Formula (9), $y_a^{'}$ and y_a represent real object labels and

model prediction labels respectively, C represents the total number of categories in the dataset, and *Loss* represents the average loss value of each mini-batch.

S^{2}FTNet includes two branches, spatial Transformer branch and spectral Transformer branch, respectively. Then, the high-level semantic features obtained from these two branches will be combined and sent to the classifier. In this section, we will introduce in detail how to effectively combine the features extracted from these two branches. Usually, two features are cascaded as follows

$$F = Concat\left(F_{Spa}, F_{Spe}\right) \tag{10}$$

However, considering that the two important degrees of the features extracted from the two branches are different, we introduce the balance factor λ for score weighting. That is

$$F = Concat\left(\lambda F_{Spa}, (1-\lambda) F_{Spe}\right) \tag{11}$$

In the back propagation process, the balance factor update can be expressed as

$$\lambda = \lambda_0 - \eta \frac{\partial}{\partial \lambda} Loss \tag{12}$$

where λ_0 is the random initial value of the balance factor, and η is the learning rate. By adaptively determining the proportion of these two parts, the model has stronger data representation ability than feature weighted addition.

Table I
The implementation process of S^{2}FTNet.

Algorithm 1 S^{2}FTNet implementation process

Input: HSI image data $X \in \mathbb{R}^{H\times W\times L}$, label is $Y \in \mathbb{R}^{H\times W}$, PCA parameter $b = 30$, space size $s = 13$.

Output: The classification accuracy and visual classification map of the marked samples.

1:The HSI data is edge filled and neighborhood and pixel by pixel cubes are extracted respectively. The data obtained are then processed by PCA to obtain $X_1 \in \mathbb{R}^{s\times s\times b}$ and $X_2 \in \mathbb{R}^{1\times 1\times L}$.

2:Set the GSD optimizer and learning rate $r = 0.005$, and select the batch size $b = 64$ and training iterations as $e = 200$.

3: **for** $e = 1$ to 200 do

4: Select X_1 as the input data of the SpeFormer, and execute Conv3D and Conv2D.

5: The space size $s\times s$ is operated by $pooling = false$, $pooling = 2$ and $pooling = 4$ respectively to get three data.

6: A spatial Tranformer block that performs paralleling.

7: Select X_2 as the input data of the SpeFormer, and execute the improved SpeFormer block.

8: The outputs of the two branches are weighted by adaptive scores using balance factors λ.

9: Use the Softmax function to identify the label.

10:**end**

11:Save the parameters of the optimal model, and obtain the classification accuracy of the marker samples and the visual classification map of the ground object categories.

D. Algorithm implementation process

In this section, we give the implementation process of the proposed network S^{2}FTNet, as shown in Table I. Take Pavia dataset as an example, that is, input data $X_1 \in \mathbb{R}^{13\times 13\times 30}$. X performs edge filling, cuts and extracts cube by pixel, respectively, to obtain processed data $X_1 \in \mathbb{R}^{13\times 13\times 30}$ and $X_2 \in \mathbb{R}^{1\times 1\times 103}$. In the SpaFormer branch, first select X_1 as the input data, and execute Conv3D and Conv2D. Among them, Conv3D and Conv2D respectively select 8 convolution kernels with a size of 7×7×7 and 64 convolution kernels with a size of 7×7. Then, $pooling = false$, $pooling = 2$ and $pooling = 4$ operations are performed on the input image data space size $s\times s$. The space of the three images is 13×13 becomes 13×13、7×7 and 4×4, respectively. Then, in order to adapt to the improved Transformer blocks, they are reshaped and used as the input of three blocks. In the SpeFormer branch, first select X_2 as the input data to reduce the complexity, and select dim=64 to linearly map the spectral dimensions of the data. Then, the linear mapping result is executed into a position embedded and improved Transformer block. It is worth noting that the advanced semantics extracted from the two branches are adaptively weighted by introducing the balance factor λ. Finally, Softmax function is used for classification.

III. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the effectiveness of the proposed method, a series of experiments are conducted. The experiments include network ablation experiments, parameter optimization, quantitative comparison and visualization of classification results.

A. Dataset description

In this paper, three classical datasets and a newer dataset are selected for all experiments, Indian Pines, Pavia, Salinas, and WHU-Hi-LongKou datasets respectively. Next, in this section, we will detail the category information of each dataset and the number of training samples for the proposed method, specific information is shown in Table II.

Indian Pines dataset: It was captured by airborne imaging spectrometer AVIRIS from an Indian Pine tree in Indiana in 1992. Among them, there are 16 land cover categories, mainly including corn, grass, soybean, and woods. The space size of the image is 145×145, the spatial resolution is about 20m, and the imaging wavelength range is 0.4-2.5 μm. It also contains 220 continuous spectral bands. In addition to the 104-108, 150-163 and 220 absorption bands, the remaining 200 bands

Table II
Detailed category information of four datasets.

Indian Pines				Salinas			
Class	Class name	Train	Test	Class	Class name	Train	Test
1	Alfalfa	4	42	1	Brocoil-green-weeds_1	100	1909
2	Corn-notill	142	1286	2	Brocoil-green-weeds_2	186	3540
3	Corn-mintill	82	748	3	Fallow	98	1878
4	Corn	23	214	4	Fallow-rough-plow	69	1325
5	Grass-pasture	48	435	5	Fallow-smooth	133	2545
6	Grass-trees	72	658	6	Stubble	197	3762
7	Grass-pasture-mowed	3	25	7	Celery	178	3401
8	Hay-windrowed	47	431	8	Grapes-untrained	563	10708
9	Oats	3	17	9	Soil-vinyard-develop	310	5893
10	Soybean-notill	97	875	10	Corn-senesced-green-weeds	163	3115
11	Soybean-mintill	245	2210	11	Lettuce-romaine-4wk	53	1015
12	Soybean-clean	59	534	12	Lettuce-romaine-5wk	96	1831
13	Wheat	20	185	13	Lettuce-romaine-6wk	45	871
14	Woods	126	1139	14	Lettuce-romaine-7wk	53	1017
15	Bldg-Grass-Tree-Drivers	38	348	15	Vinyard-untrained	363	6905
16	Stone-Steel-Towers	9	84	16	Vinyard-vertical-trellis	90	1717
/	Total	1018	9231	/	Total	2697	51432
Pavia				WHU-Hi-LongKou			
Class	Class name	Train	Test	Class	Class name	Train	Test
1	Asphalt	331	6300	1	Corn	172	34339
2	Meadows	932	17717	2	Cotton	41	8333
3	Gravel	104	1995	3	Sesame	15	3016
4	Trees	153	2911	4	Broad-leaf soybean	316	62896
5	Painted metal sheets	67	1278	5	Narrow-leaf soybean	20	4131
6	Bare Soil	251	4778	6	Rice	59	11795
7	Bitumen	66	1264	7	Water	335	66721
8	Self-blocking bricks	184	3498	8	Roads and houses	35	7089
9	Shadows	47	900	9	Mixed weed	26	5203
/	Total	2135	40641	/	Total	1019	203523

were used for experiments.

Pavia dataset: It was captured by the airborne imaging spectrometer ROSIS-03 over the University Pavia, Italy, in 2003. The space size of the image is 610×340, with a spatial resolution of 1.3m and 115 continuous spectral bands. Similarly, because individual bands cannot be reflected by water, there are only 103 bands left. Compared with Indian dataset, Pavia contains fewer land cover categories, including trees, asphalt roads, bricks, meadows, etc.

Salinas dataset: It was captured by the imaging spectrometer AVIRIS over Salinas Valley, California, USA. The space size a total of 111104 pixels. In addition to background pixels, pixels remain for classification tasks. These pixels contain a total of 16 categories, including fallow, celery, etc.

WHU-Hi-LongKou dataset: it is collected from Longkou Town, Hubei Province, China by the 8mm focus (HNH) imaging sensor carried on the DJI Matrix 600 Pro (DJI M600 Pro) UAV platform. The space size is 550×400, the spatial resolution is about 0.463m, the wavelength range is 0.4um~1um, and 270 spectral bands are included. The number of land cover categories included in WHU-Hi-LongKou is the same as that in Pavia dataset, which is a simple crop scenario. The main categories include Water, Broad leaf soybean, Corn, Rice and Cotton.

B. Experimental setup

All experiments in this section are implemented on the platform of Intel (R) Core (TM) i9-9900K CPU, NVIDIA GeForce RTX 2080Ti GPU and 128G random access memory, and the language framework is Python. In addition, in order to better evaluate the classification performance of the model, we choose three common evaluation indicators: Overall Accuracy (OA), Average Accuracy (AA) and Kappa coefficient. Among them, OA represents the ratio of the number of accurately classified samples to the total number of samples, AA represents the average of the classification accuracy of each category, and Kappa is a measure of robustness.

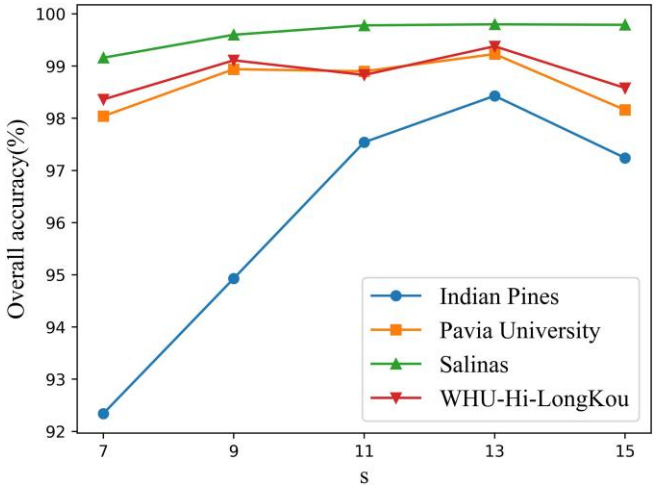


Fig. 4. Impact of different input space sizes on OA.

The network constructed by combining CNN and Transformer is more inclined to spatial information of global context. In order to analyze the impact of different input space sizes s on the final classification performance, we selected

7-15 input space sizes for experiments on four datasets. The adjacent space size interval is 2. The experimental results are shown in Fig. 4. It can be seen from Fig. 4 that Indian Pines dataset is highly sensitive to different input space sizes. The classification accuracy OA of Pavia and WHU-Hi-LongKou datasets shows a trend of increasing first and then decreasing. For Salinas dataset, with the increase of input space size s, OA increases first and then tends to be stable. It is worth noting that when $s=13$, the four datasets have achieved the highest overall accuracy OA. Therefore, $s=13$ is selected as the input space size of the proposed network.

In addition, different learning rates and batch sizes have a greater impact on the performance of the model. In order to explore the optimal learning rate and batch size of the proposed network, some relevant experiments were carried out, and the experimental results are shown in Fig. 5. Fig. 5 (a)-(d) show the results of experiments on Indian Pines, Pavia, Salinas and WHU-Hi-LongKou datasets respectively. Among them, different contour colors represent different ranges of OA values, and red to blue represent a gradual decrease in OA values. It can be found that the OA value of the same dataset is more sensitive to different learning rates and batch sizes of the model. Especially for the Indian Pines and WHU-Hi-LongKou datasets, due to the small number of training data samples used in the training process, the learning rate has a significant impact on them.

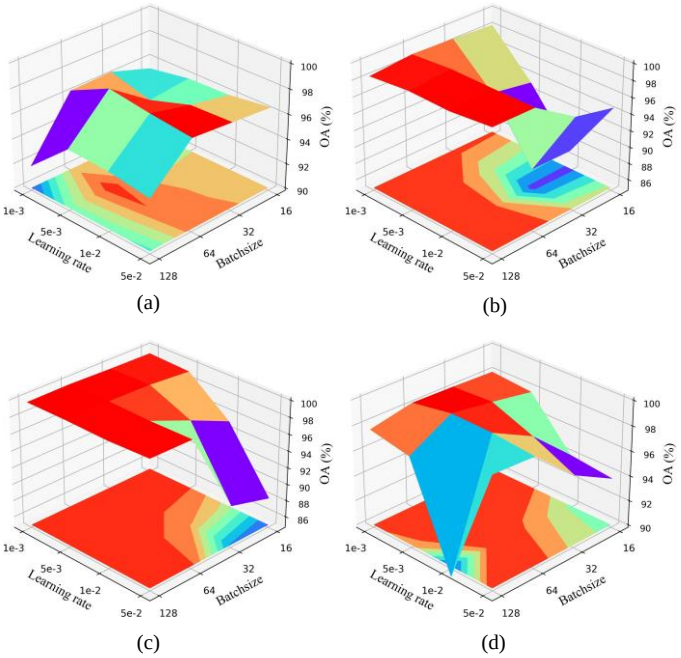


Fig. 5. Effect of different learning rates and batch sizes on performance accuracy OA. (a) Experimental results on Indian Pines dataset, (b) Experimental results on Pavia dataset, (c) Experimental results on Salinas dataset and (d) Experimental results on WHU-Hi-LongKou dataset.

Specifically, for the Indian Pines dataset, as shown in Fig. 5 (a), the optimal learning rate and batch size are $5e-4$ and 64, respectively. For Pavia data, as shown in Fig. 5 (b), when the batch size is 64 or 128, the learning rate has little impact on the performance OA. Similarly, for Salinas dataset, as shown in Fig. 5 (c), when the learning rate is large and the batch size is large, better OA values can often be obtained. For the WHU-Hi-LongKou dataset, as shown in Fig. 5 (d), when the batch size is 64, the selected learning rate can achieve better classification results. Therefore, through the parameter experiment of the model, it can be found that the best learning rate and batch size of the classification network proposed in this paper are $5e-3$ and 64.

C. Ablation experiments

In the proposed method, the network mainly includes four parts, Conv2D&3D, SpaFormer, SpeFormer, and AS^{2}FM respectively. In order to better verify the impact of each part on the classification performance OA value. We conducted ablation experiments on them in four datasets, and the experimental results are shown in Table III. Among them, "√" indicates that the module is available, and "-" indicates that the module is not used. There are five cases in total. It can be seen from the table that the case 1 only includes Conv2D and Conv3D, and the overall accuracy OA value obtained is low. In the case 2 and the case 3, SpaFormer and SpeFormer are added on the basis of Conv2D&Conv3D respectively. It can be found that the accuracy OA is worth improving greatly. Generally, the features extracted from the two branches will be combined in a cascade (Cat) manner, as in case 4. In order to better combine these two features, we introduce a balance factor to fuse the features obtained from the two branches. The experiment shows that the OA value of Case 5 is higher than that of Case 4 on the four datasets, which fully proved the effectiveness of this adaptive combination method.

Table III
Impact of different modules on network OA value (%)

Case	Conv2D&3D	SpaFormer	SpeFormer	Cat	AS^{2}FM	Indian Pines	Pavia	Salinas	WHU-Hi-LongKou
1	√	-	-	-	-	79.38	96.96	98.64	94.12
2	√	√	-	-	-	96.80	98.43	99.27	98.89
3	√	-	√	-	-	97.30	98.03	99.43	98.71
4	√	√	√	√	-	97.85	98.97	99.76	98.92
5	√	√	√	-	√	**98.50**	**99.38**	**99.80**	**99.39**

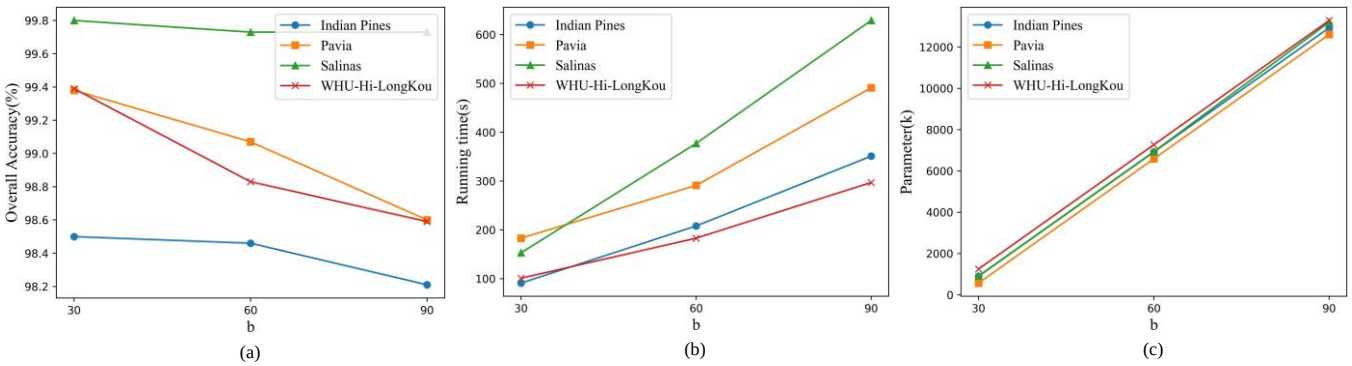


Fig. 6. Comparison of different b of PCA. (a) The impact of different b on OA. (b) The impact of different b on running time. (c) The impact of different b on parameter.

In addition, we also conducted experiments on the impact of different b of PCA (b =30, b =60, and b =90) on classification performance. The experimental results are shown in Fig. 6, and Fig. 6 (a)-(c) show the impact of different b on the overall accuracy (%), running time (s), and parameter (k). From Fig. 6 (a), it can be seen that different b have little impact on the Salinas dataset, and they slowly decrease as b increases in the other three datasets. We infer that this is due to the dimensionality disaster caused by the high-dimensional characteristics of HSIs and the inclusion of redundant features, which leads to a small reduction in classification accuracy. As shown in Fig. 6 (b) and (c), it can be seen that with the increase of b , the running time and parameter increase exponentially. Therefore, in our proposed method, we choose b =30 as the optimal dimensionality reduction parameter for PCA.

D. Analysis of experimental results

In order to verify the superiority of the proposed classification network, we have selected a classifier (ISVM) [45] and a variety of the state-of-the-art networks based on CNN and Transformer, including 2DCNN [27], 3DCNN [28], Hybrid-SN [29], PyResNet [34], LiteDepthwiseNet [41], MCRSCA [30], ViT [25], SF [47], SSFTT [48], SSTN [49] and GAHT [54].

1) Quantitative analysis

The classification accuracy of OA, AA, Kappa and each category of all methods on the four datasets are shown in Table IV-Table VII. The best classification results are bold. As can be seen from the table, CNN based methods have achieved relatively good classification results due to their strong ability to extract context features. However, due to the limited advanced global features obtained by CNN, it is easy to fall into the performance bottleneck. In addition, although Transformer based methods show great potential by building long-distance dependencies, the classification performance of networks built only using Transformer frameworks is not satisfactory, such as ViT and SF. However, the classification network constructed by combining CNN and Transformer framework has achieved good classification results, such as SSFTT, SSTN, GAHT, and the proposed method. It is worth noting that ISVM based on classifier design has also obtained competitive classification results.

In general, the classification accuracy of the proposed classification method is better than that of other comparison methods on the four datasets. This result not only benefits from the proposed method S^{2}FTNet, which combines the advantages of CNN and Transformer, but also benefits from the effective fusion of the extracted spectral-spatial high-level semantic features. More specifically, compared with the best CNN method among the comparison methods (MCRSCA), the OA value of S^{2}FTNet is 0.38%, 0.39%, 2.78%, and 1.04% higher on the Indian Pines, Pavia, Salinas, and WHU-Hi-LongKou datasets, respectively. Compared with the best Transformer method in the comparison method (SSFTT), the OA value of S^{2}FTNet is 1.07%、0.23%、0.39%, and 0.40% higher on the Indian Pines, Pavia, Salinas and WHU-Hi-LongKou datasets, respectively. Compared with ISVM classifiers, the OA values of S^{2}FTNet on four datasets are 0.90%, 3.36%, 0.03%, and 0.42% higher, respectively. It is worth noting that our method achieves 100% accuracy for individual categories in some datasets. For example, category 1 (Alfalfa), category 3 (Corn-mintill), category 7 (Grass pace moved), category 8 (Hay windowed), and category 9 (Oats) on the Indian Pines dataset. Category 5 (Painted metal sheets), Category 6 (Bare Soil) and Category 7 (Bitumen) on Pavia dataset. Category 1 (Brocool-green-weeds_1), Category 7 (Celery), Category 10 (Corn-senced-green-weeds), Category 11 (Lettuce-remaine-4wk), Category 12 (Lettuce-remaine-5wk), and Category 13 (Lettuce-remaine-6wk) on the Salinas dataset.

Table IV

Classification accuracy of OA, AA, Kappa and various categories of all methods on Indian Pines dataset. The best classification results are bold.

Methods	Classifier	CNNs						Transformers					Proposed
	ISVM [45]	2DCNN [27]	3DCNN [28]	Hybrid-SN [29]	PyResNet [34]	LiteDepthwiseNet [41]	MCRSCA [30]	ViT [25]	SF [47]	SSFTT [48]	SSTN [49]	GAHT [54]	
OA (%)	97.60	82.04	81.15	94.31	92.86	96.28	98.12	79.73	88.54	97.43	95.43	83.00	**98.50**
AA (%)	98.46	89.09	87.15	94.32	92.15	95.03	96.31	83.36	91.81	93.85	84.54	86.22	**97.65**
κ ×100	97.26	79.26	78.26	93.51	91.87	95.75	97.86	76.75	86.88	97.07	94.78	80.49	**98.30**
1	97.83	100	100	96.91	98.25	92.51	90.54	99.00	**100**	85.71	39.41	85.56	**100**

2	91.04	76.01	72.54	90.87	93.47	96.60	98.17	72.83	82.89	98.13	96.91	80.11	**98.52**
3	99.40	77.24	69.29	92.05	87.13	96.24	98.10	69.39	84.39	96.66	95.55	76.80	**100**
4	99.58	96.07	95.83	91.32	**99.77**	95.42	96.79	82.71	91.44	96.73	96.46	88.60	99.01
5	**99.59**	93.85	95.23	98.73	95.33	95.43	97.33	85.30	93.50	98.39	94.99	91.29	98.78
6	**99.86**	79.85	81.47	97.87	96.91	98.16	98.24	85.22	91.74	99.09	96.98	88.04	99.68
7	96.43	80.00	60.00	91.32	69.90	81.41	91.36	92.86	**100**	**100**	31.12	91.67	**100**
8	**100**	97.04	95.89	98.34	97.05	97.08	99.84	92.06	95.34	98.38	96.71	90.06	**100**
9	**100**	**100**	**100**	95.54	85.19	92.46	86.88	79.50	**100**	58.82	40.00	87.03	**100**
10	97.84	83.60	87.14	91.61	89.02	94.35	97.53	76.73	87.01	**99.31**	88.37	79.48	96.86
11	96.95	75.16	75.66	95.23	93.95	96.42	98.56	77.39	88.19	98.10	96.29	78.98	**99.04**
12	**99.49**	82.73	77.70	92.74	85.23	94.88	95.81	68.13	79.56	89.70	92.87	74.75	99.01
13	**100**	**100**	99.86	99.52	98.36	99.63	98.54	93.56	95.37	99.46	98.79	93.59	92.00
14	99.72	92.99	93.23	96.79	96.83	97.55	**99.76**	92.06	93.92	98.68	98.79	92.67	98.14
15	97.70	90.86	91.10	94.74	91.83	94.59	96.31	82.41	91.14	93.97	94.17	81.98	**98.78**
16	**100**	**100**	99	85.57	96.16	97.75	97.16	84.62	94.33	90.47	95.24	98.94	82.50

Table V

Classification accuracy of OA, AA, Kappa and various categories of all methods on Pavia dataset. The best classification results are bold.

Methods	Classifier	CNNs						Transformers					Proposed
	ISVM [45]	2DCNN [27]	3DCNN [28]	Hybrid-SN [29]	PyResNet [34]	LiteDepthwiseNet [41]	MCRSCA [30]	ViT [25]	SF [47]	SSFTT [48]	SSTN [49]	GAHT [54]	
OA (%)	96.02	94.55	93.69	97.99	97.72	98.86	98.99	94.35	95.89	99.15	97.20	94.68	**99.38**
AA (%)	97.70	93.55	93.38	97.49	97.00	98.73	98.38	92.15	93.64	98.62	96.75	94.40	**98.89**
κ	94.80	92.74	91.56	97.33	96.98	98.49	98.66	92.48	94.55	98.87	96.27	92.93	**99.18**
1	97.39	91.07	88.64	97.19	96.63	99.50	98.37	90.74	93.23	**99.67**	95.89	91.86	99.56
2	93.53	97.18	96.27	99.22	99.38	99.89	99.87	97.57	98.96	**99.99**	98.83	97.03	99.96
3	94.76	83.47	83.55	97.66	93.46	**99.22**	95.52	82.37	82.55	98.59	92.36	88.98	94.14
4	99.09	99.31	98.99	99.08	97.81	99.41	98.42	99.14	**99.83**	93.71	97.21	98.65	98.04
5	99.85	99.84	99.44	98.18	99.11	99.79	99.97	96.87	99.73	99.84	96.13	**100**	**100**
6	97.97	95.81	95.83	98.68	99.17	99.86	99.36	94.43	97.79	99.54	99.95	91.94	**100**
7	99.70	89.88	91.04	97.36	99.39	99.95	95.96	79.37	80.51	99.53	99.08	92.16	**100**
8	98.29	86.76	87.87	92.16	91.07	91.49	98.04	89.59	90.31	98.00	92.89	88.92	**98.94**
9	98.73	98.66	98.76	97.87	96.94	99.51	99.94	99.30	99.88	98.67	98.41	**100**	99.33

Table VI

Classification accuracy of OA, AA, Kappa and various categories of all methods on Salinas dataset. The best classification results are bold.

Methods	Classifier	CNNs						Transformers					Proposed
	ISVM [45]	2DCNN [27]	3DCNN [28]	Hybrid-SN [29]	PyResNet [34]	LiteDepthwiseNet [41]	MCRSCA [30]	ViT [25]	SF [47]	SSFTT [48]	SSTN [49]	GAHT [54]	
OA (%)	99.67	96.01	96.62	98.99	96.57	96.88	97.09	97.87	97.72	99.41	94.03	95.56	**99.80**
AA (%)	99.41	98.02	98.22	99.29	97.27	97.97	98.30	99.43	98.85	99.37	98.08	97.29	**99.74**
κ ×100	99.63	95.55	96.24	98.88	96.18	96.52	96.75	97.55	97.46	99.34	93.40	95.05	**99.78**
1	**100**	99.81	99.72	99.86	87.44	**100**	99.20	97.87	99.40	99.95	99.37	99.83	**100**
2	99.60	99.62	99.31	99.98	99.88	**99.99**	99.08	99.43	99.97	99.92	99.86	99.85	99.75
3	**100**	99.39	98.51	99.82	80.83	96.37	99.41	97.55	98.75	99.89	98.52	97.05	99.31
4	97.85	99.51	99.42	99.63	99.82	98.34	99.30	98.98	99.79	**99.85**	99.03	97.07	98.49
5	99.74	99.10	99.02	99.45	**99.86**	99.42	98.77	98.68	99.13	98.66	98.85	98.80	99.76
6	99.90	99.95	99.96	99.87	96.92	**99.98**	99.93	99.92	99.93	99.79	99.91	99.89	99.92
7	99.64	99.62	99.76	99.81	99.86	99.82	98.87	99.84	99.96	99.97	99.92	99.01	**100**
8	**99.87**	90.84	92.18	98.15	98.42	94.98	94.74	92.45	94.86	98.55	94.63	91.91	99.76
9	99.98	99.73	99.93	99.88	**100**	99.73	99.84	99.39	99.93	99.88	99.45	99.93	99.98
10	99.42	95.84	96.64	99.29	99.27	96.15	97.75	98.16	98.25	99.13	99.64	95.71	**100**
11	99.44	97.98	98.69	99.05	98.97	98.09	97.54	94.14	99.11	**100**	98.80	97.39	**100**
12	**100**	99.58	99.59	99.87	99.43	99.38	99.83	99.08	99.75	**100**	99.91	98.74	**100**
13	98.91	98.95	99.06	99.07	99.84	97.44	99.99	98.55	**100**	99.20	99.42	99.03	**100**
14	96.45	99.17	99.18	98.02	98.13	98.65	98.90	97.35	99.69	95.67	99.60	97.82	**99.90**
15	99.71	89.75	90.77	97.15	97.79	89.80	90.40	89.02	93.05	**99.93**	83.00	85.42	99.86
16	**100**	99.54	99.85	99.73	99.88	99.32	99.19	99.50	99.96	99.48	**100**	99.10	99.18

Table VII

Classification accuracy of OA, AA, Kappa and various categories of all methods on WHU-Hi-LongKou dataset. The best classification results are bold.

Methods	Classifier	CNNs						Transformers					Proposed
	ISVM [45]	2DCNN [27]	3DCNN [28]	Hybrid-SN [29]	PyResNet [34]	LiteDepthwiseNet [41]	MCRSCA [30]	ViT [25]	SF [47]	SSFTT [48]	SSTN [49]	GAHT [54]	
OA (%)	98.97	89.95	95.12	98.60	97.73	98.87	98.35	86.48	92.43	98.99	97.88	96.89	**99.39**
AA (%)	96.03	80.61	89.72	96.73	96.56	98.52	94.46	73.71	82.60	97.81	96.13	92.55	**99.20**
κ ×100	98.65	86.56	93.57	98.16	97.01	98.51	97.83	82.07	90.05	98.68	97.21	95.91	**99.46**

1	99.14	96.03	99.75	99.20	98.83	98.88	99.67	83.68	96.11	98.24	99.86	99.58	**99.97**
2	**99.94**	59.83	66.72	96.53	95.35	98.37	94.67	39.75	72.39	99.86	90.14	87.33	99.25
3	81.16	96.67	98.92	94.00	99.68	**99.93**	90.91	59.38	83.48	99.07	99.41	92.14	99.57
4	99.30	85.20	95.10	98.63	99.53	98.25	98.85	87.93	93.71	**99.60**	97.18	96.12	99.51
5	89.64	56.31	76.30	95.13	96.01	99.11	82.43	35.03	67.89	97.57	97.81	79.07	**99.64**
6	99.64	91.99	97.81	98.74	99.03	99.73	99.20	95.80	94.80	99.45	**99.79**	99.51	99.66
7	99.95	98.86	98.80	99.69	96.64	99.90	99.96	99.25	97.6	99.82	99.76	99.89	**99.98**
8	**98.13**	67.55	82.63	95.29	91.76	93.47	92.20	92.09	80.60	90.93	86.26	86.13	90.45
9	97.32	73.03	91.40	93.36	92.19	**99.00**	92.23	70.50	57.07	95.79	94.94	93.24	98.13

2) Visual Evaluation

Fig. 7-10 show the classification results of all methods on four datasets. It can be clearly seen that the visual effect of the proposed method is closer to the real ground object map. On the Indian Pines dataset, the CNN based classification method has poor classification effect on edge categories, while the classification method combining CNN and Transformer has better classification results than CNN, which also benefits from more abundant features extracted, including global and local features. In addition, due to the small number of samples in category 16 and the complex coverage of ground features, the proposed AS^{2}FNet focuses more on long-distance dependencies, resulting in classification accuracy for category 16 is not very well. However, among other categories, most of the classification results obtained by our method are optimal. However, among other categories, most of the classification results obtained by our method are optimal. Pavia dataset contains fewer bands, and the distribution of buildings is more complex. The proposed S^{2}FTNet method has less noise in the classification result map, while most of the comparison methods have more classification errors in the category "Meadows". For Salinas dataset, two categories that are easy to observe, Vinyard untrained and Grapes untrained, our method has the best visual effect, followed by SSFTT. Among them, 2DCNN, 3DCNN, ViT and SF in the comparison method have serious misclassification. For the WHU-Hi-LongKou dataset, the images mainly include crops with similar spectra. Our method combines CNN and Transformer to build a spatial and spectral extraction module, which fused spectral information and spatial information well. The obtained classification result has better edge effect and less intra class noise.

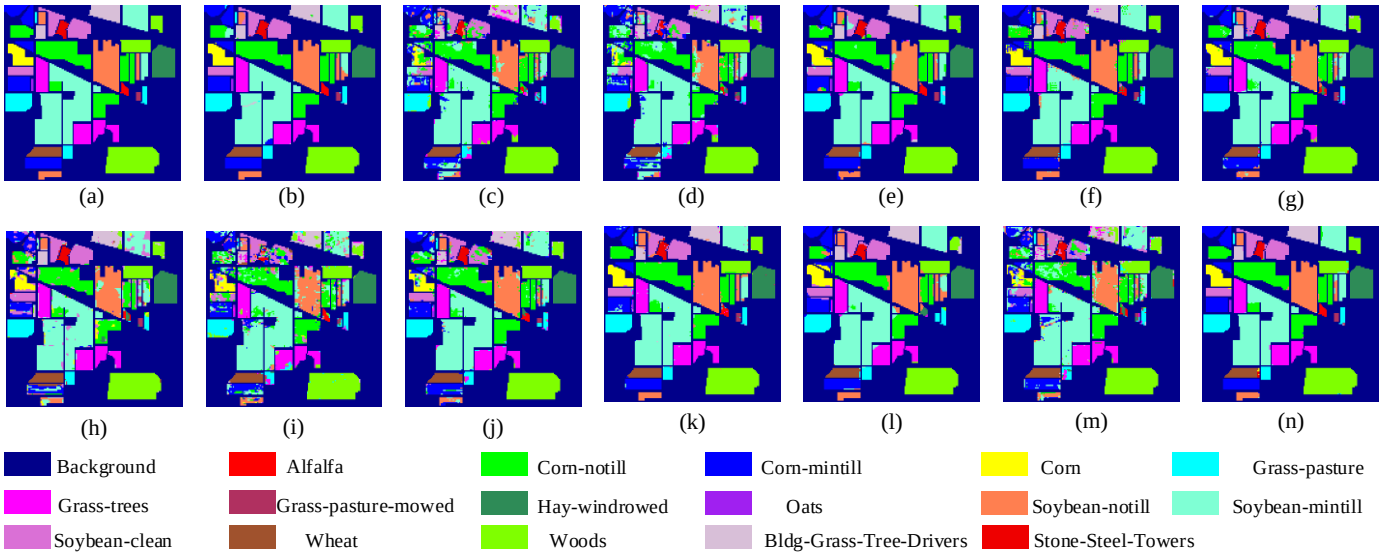


Fig. 7. Classification visualization maps of all methods on Indian Pines dataset. (a) Real ground feature map,(b)-(n) classification map of ISVM, 2DCNN, 3DCNN, Hybrid-SN, PyResNet, LiteDepthwiseNet, MCRSCA, ViT, SF, SSFTT, SSTN, GAHT and proposed respectively.

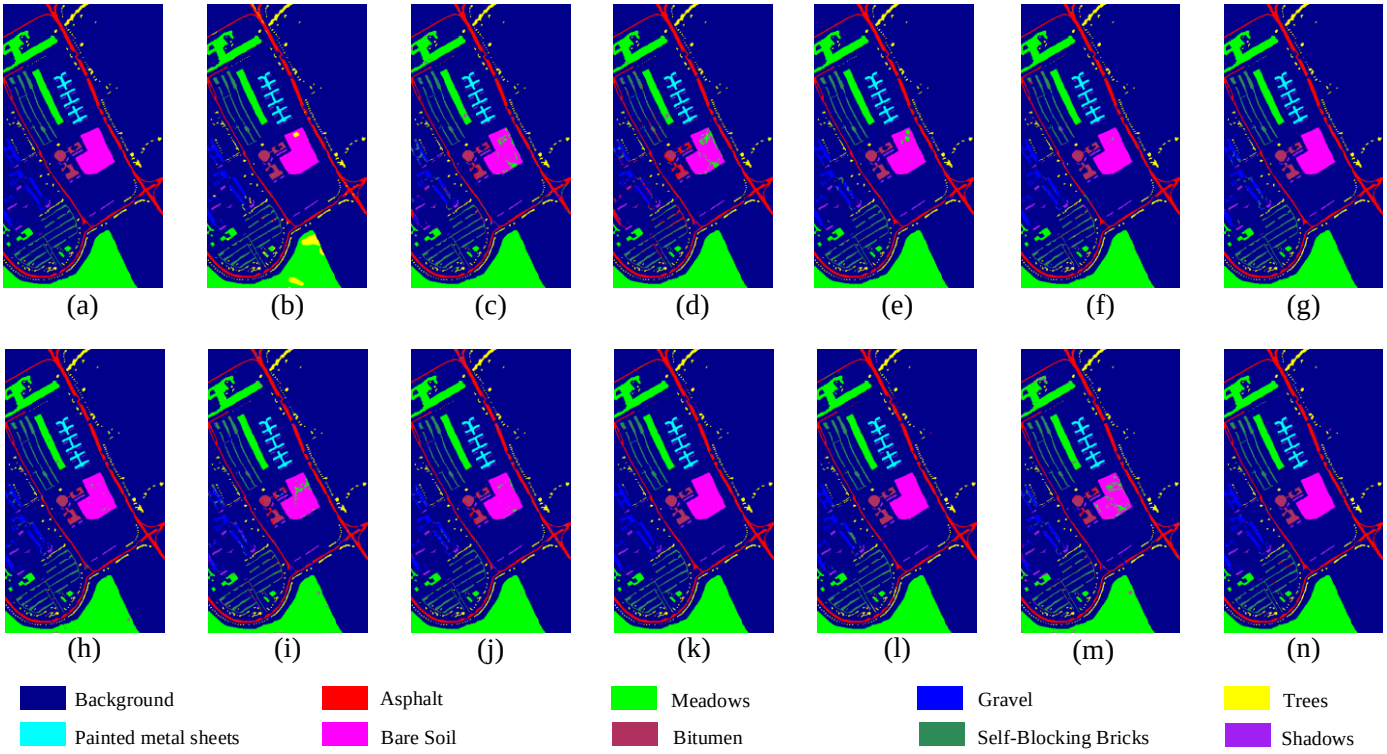


Fig. 8. Classification visualization maps of all methods on Pavia dataset. (a) Real ground feature map,(b)-(n) classification map of ISVM, 2DCNN, 3DCNN, Hybrid-SN, PyResNet, LiteDepthwiseNet, MCRSCA, ViT, SF, SSFTT, SSTN, GAHT and proposed respectively.

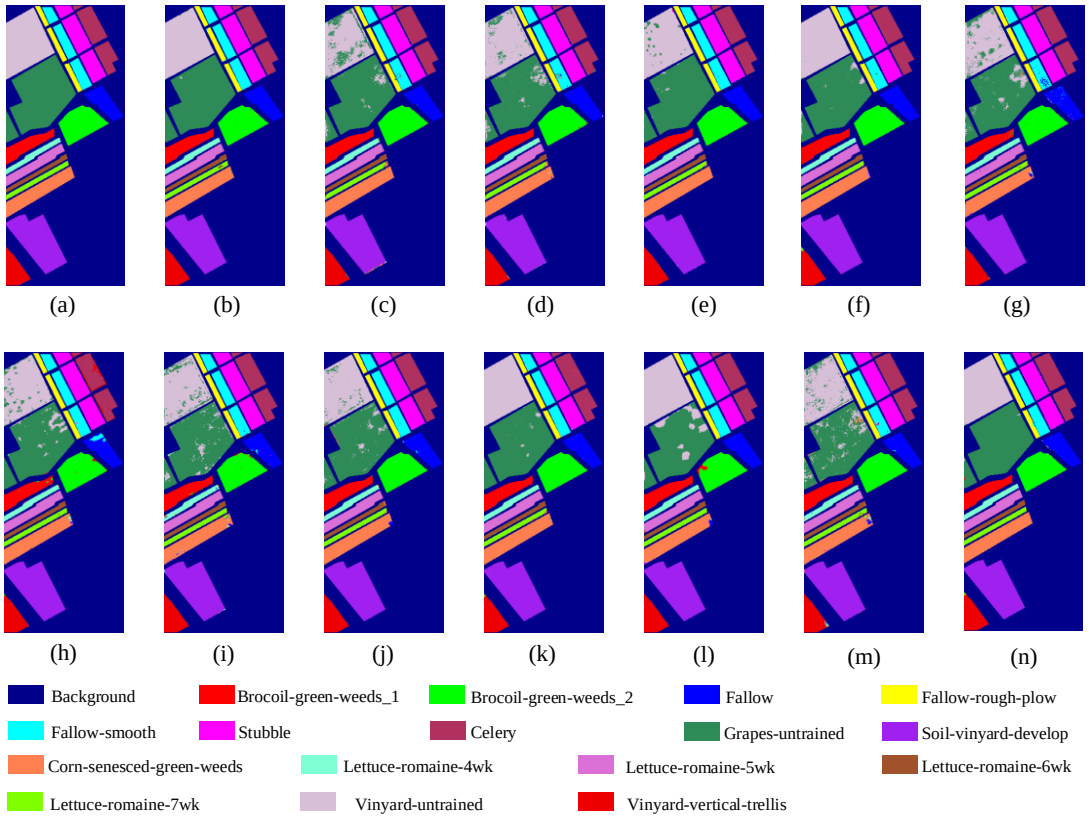


Fig. 9. Classification visualization maps of all methods on Salinas dataset. (a) Real ground feature map,(b)-(n) classification map of ISVM, 2DCNN, 3DCNN, Hybrid-SN, PyResNet, LiteDepthwiseNet, MCRSCA, ViT, SF, SSFTT, SSTN, GAHT and proposed respectively.

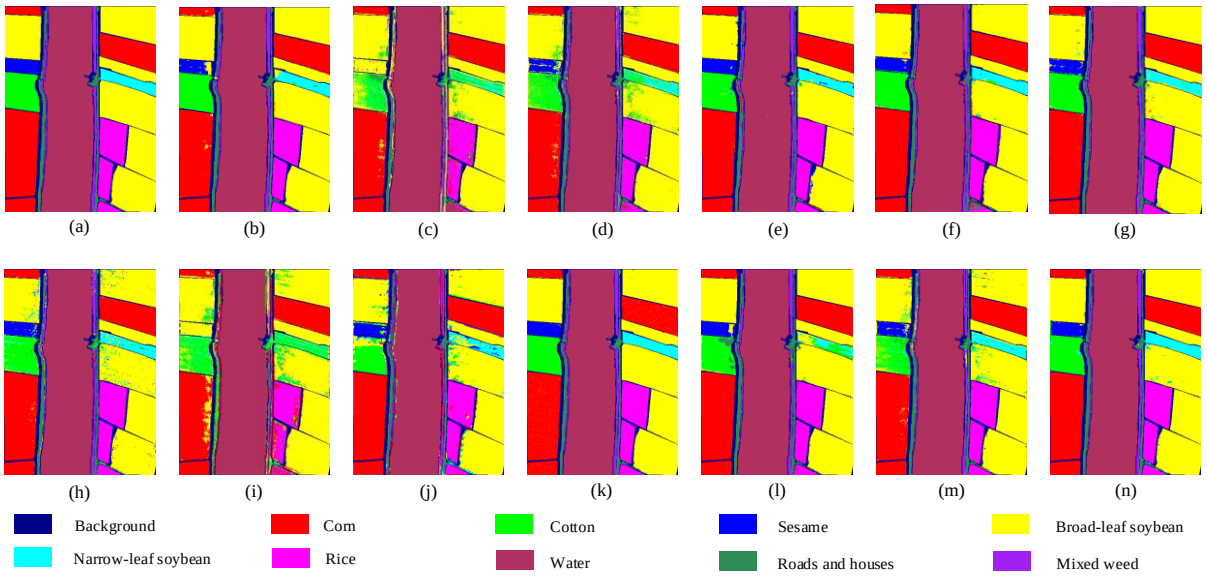


Fig. 10. Classification visualization maps of all methods on WHU-Hi-LongKou dataset. (a) Real ground feature map,(b)-(n) classification map of ISVM, 2DCNN, 3DCNN, Hybrid-SN, PyResNet, LiteDepthwiseNet, MCRSCA, ViT, SF, SSFTT, SSTN, GAHT and proposed respectively.

In order to more clearly illustrate the effectiveness of the proposed S^2FTNet method, we compared T-SNE visualization of features obtained by various methods (including 3DCNN, Hybrid-SN, and SSTN) on four datasets. The experimental results are shown in Fig. 11-14. Different colors represent labels of different categories. From left to right, they are the category distribution results of methods 3DCNN, Hybrid-SN, SSTN and proposed. More specifically, on the Indian Pines dataset, both 3DCNN and SSTN methods have serious label mixing. Although Hybrid-SN has obtained a better intra class distance than 3DCNN and SSTN, the inter class distance is still not satisfactory. However, our method has more obvious cluster, showing better intra class and inter class distance. For Pavia dataset, 3DCNN and SSTN methods performed poorly, and category 2 (yellow), category 4 (gray) and category 9 (yellow) were still seriously mixed. Compared with the Indian Pines dataset, Hybrid-SN performs better. However, our approach is still significantly better. For Salinas dataset, the category distribution of 3DCNN, SSTN and Hybrid-SN is mostly in a strip shape, with a large gap in the distance within the category. However, most of the categories of our methods are clustered and have large intra class distance. Due to the large number of sample categories in the WHU-Hi-LongKou dataset, its category distribution visualization effect is relatively full, but it is not difficult to see that there are some mixed categories in 3DCNN, SSTN and Hybrid-SN, and the category distribution is relatively scattered. On the contrary, our method obtains that the features of the same category are more clustered, and the distribution of different categories is more dispersed. In general, the proposed method S^2FTNet has better inter class distance and minimized intra class distance, and plays an important role in capturing the relationship between HSI classification samples.

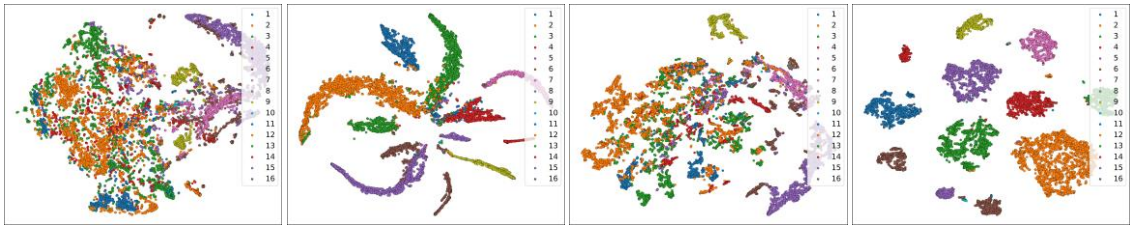


Fig 11. T-SNE visualization of different methods on Indian Pines dataset. (a) 3DCNN, (b) Hybrid-SN, (c) SSTN and (d) proposed.

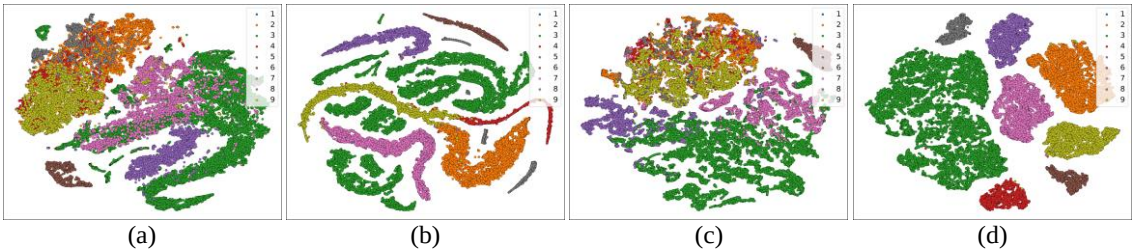
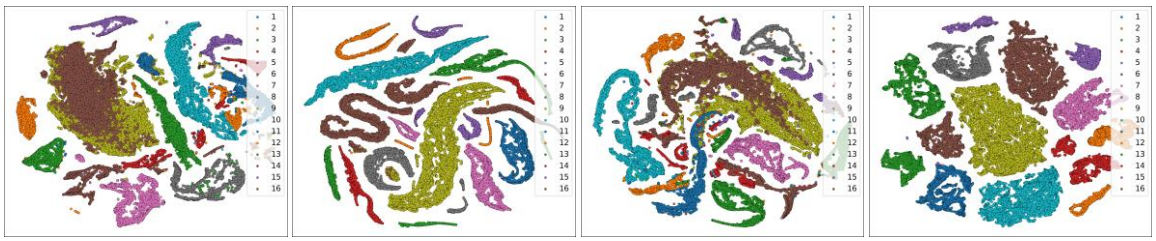
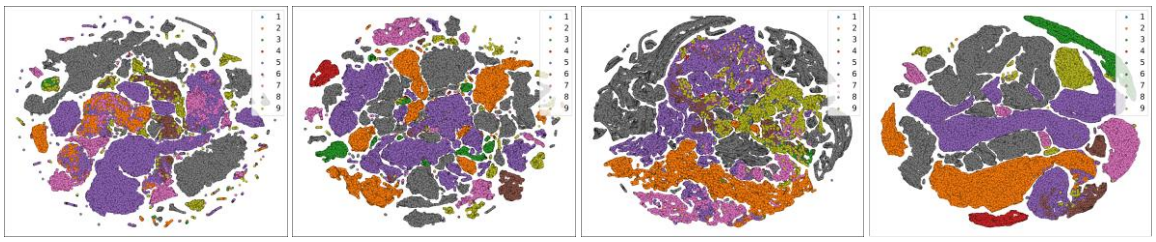


Fig. 12. T-SNE visualization of different methods on Pavia dataset. (a) 3DCNN, (b) Hybrid-SN, (c) SSTN and (d) proposed.



(a) (b) (c) (d)

Fig. 13. T-SNE visualization of different methods on Salinas dataset. (a) 3DCNN, (b) Hybrid-SN, (c) SSTN and (d) proposed.

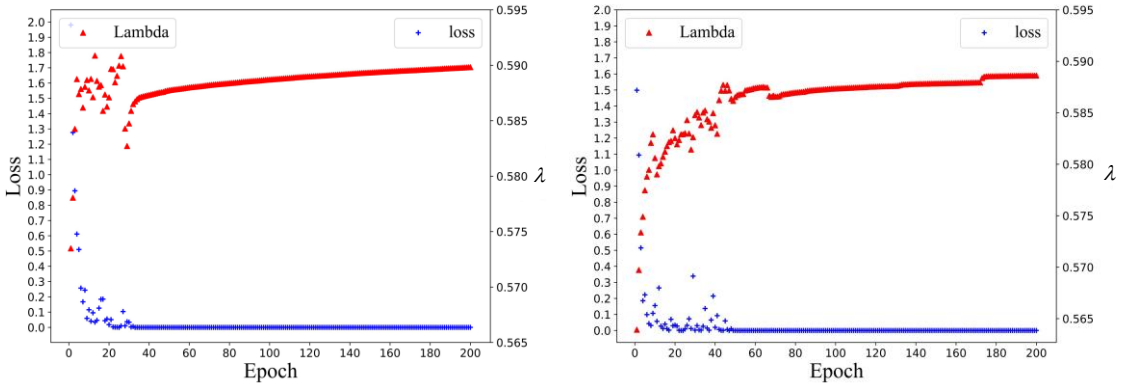


(a) (b) (c) (d)

Fig. 14. T-SNE visualization of different methods on WHU-Hi-LongKou dataset. (a) 3DCNN, (b) Hybrid-SN, (c) SSTN and (d) proposed.

3) Model Hyperparametric Analysis

In the designed network, considering the different importance of features extracted from spatial and spectral branches and their different contributions to the final classification results, we introduced a balance factor λ into the network and weighted the two branches by fractions. It will be updated gradually with the change of loss value during the training. In order to observe the changes of balance factor λ and loss value, we selected two datasets for the experiment, Indian Pines and WHU-Hi-LongKou datasets. The experimental results are shown in Fig. 15 (a) and (b). The red dot represents the balance factor λ value, and the blue dot represents the loss value. The abscissa represents the training Epoch. The left and right ordinates have different magnitudes. The left ordinate is the loss value, and the right ordinate is the balance factor value. It can be found that, on the one hand, the training Epoch of the two datasets is about 40, and the Loss value is close to 0, which shows that the combination of these two branch features can achieve faster convergence. On the other hand, the balance factor λ updated slowly and tends to be stable with the increase of Epoch, and the stable value is about 0.590. The above results show that the features extracted by the SpaFormer branch and the SpeFormer branch are different in importance, and the SpaFormer branch accounts for a larger proportion than the SpeFormer branch, and the spectral-spatial features obtained are more abundant. Finally, the classification performance can be effectively improved by adaptive fusion of these two features. For Indian Pines and WHU-Hi-LongKou datasets, the former has many categories, while the latter has large spatial resolution. The long-distance spectral and spatial features extracted by the SpaFormer branch contribute greatly to the classification results of the two datasets.



(a) (b)

Fig. 15. Changes of balance factor λ and loss value on different datasets. (a) Indian Pines dataset, (b) WHU-Hi-LongKou dataset.

4) Combining different classifiers

In this section, we select two spatial and spectral classifiers, EPF and IEPF, and combine our method with these two classifiers for experiments. The experimental results are shown in Table VIII. From the results, we can see that the method with the lowest classification accuracy OA value is EPF. IEPF has improved EPF and greatly improved classification accuracy. In addition, our method combines EPF and IEPF classifiers, and it

can be found that compared to EPF, the classification accuracy of S^{2}FTNet_EPF has been improved on all four datasets. Similarly, compared to IEPF, S^{2}FTNet_IEPF can also effectively improve classification performance. This shows that our proposed method can effectively extract spatial and spectral features.

Table VIII
Comparison of classification OA (%) results of S^{2}FTNet combined with different classifiers.

datasets	Indian Pines	Pavia	Salinas	WHU-Hi-LongKou
SVM_EPF	88.95	76.21	94.38	97.33
SVM_IEPF	97.72	99.56	99.63	98.97
S^{2}FTNet	98.50	99.38	99.80	99.39
S^{2}FTNet_EPF	97.34	78.39	99.07	98.86
S^{2}FTNet_IEPF	98.96	99.90	99.91	99.59

5) Model efficiency analysis

In order to evaluate the running efficiency of the proposed methods, this paper conducts running efficiency test experiments for all methods, and Table IX shows the results of the experiments. As can be seen from the table, compared with the method SSFTT, which requires the shortest training time and test time, the training time and test time required for the method S^{2}FTNet proposed in this paper are slightly longer. This is because the proposed method is a two branch Transformer structure. Compared with other Transformer based methods, S^{2}FTNet generally requires less running time. In addition, compared with the CNN based method, the Transformer based method requires much less training time and testing time. In general, the efficiency of Transformer based method is significantly higher than that of CNN based method. Compared with other methods, the running time of the proposed S^{2}FTNet is relatively close to that of the optimal method. The experiment fully shows that S^{2}FTNet not only has good classification accuracy, but also has satisfactory operation efficiency.

Table IX
Comparison of running time of all methods on four datasets.

datasets	time	ISVM	2DCNN	3DCNN	Hybrid-SN	PyResNet	LiteDepthwiseNet	MCRSCA	ViT	SF	SSFTT	SSTN	GAHT	Proposed
Indian Pines	Train.(min)	0.41	1.20	4.46	1.26	7.96	5.73	5.36	9.93	12.53	0.56	1.36	0.83	1.40
	Test.(s)	5.33	6.73	32	2.84	6.98	3.53	5.00	10.95	71.46	1.06	1.19	0.81	1.30
Pavia	Train. (min)	5.82	1.40	5.96	1.23	16.46	5.83	13.10	7.70	11.96	1.03	2.53	1.40	2.63
	Test.(s)	8.31	15.4	78	12.76	32.87	7.95	43.13	17.53	40.92	3.23	5.10	2.85	5.15
Salinas	Train. (min)	1.51	4.53	14.60	0.90	20.50	15.33	10.80	21.43	20.83	0.63	3.53	2.10	3.73
	Test.(s)	34.62	38.6	41.8	16.21	41.39	21.37	14.10	24.56	33.30	2.76	7.16	4.27	7.29
WHU-Hi -LongKou	Train. (min)	1.55	1.40	7.50	6.16	5.43	7.76	3.80	4.60	5.30	0.26	1.46	1.26	1.30
	Test.(s)	22.73	6.8	7.0	79.10	171.8	130.42	30.00	38.13	44.25	8.68	32.25	21.70	35.98

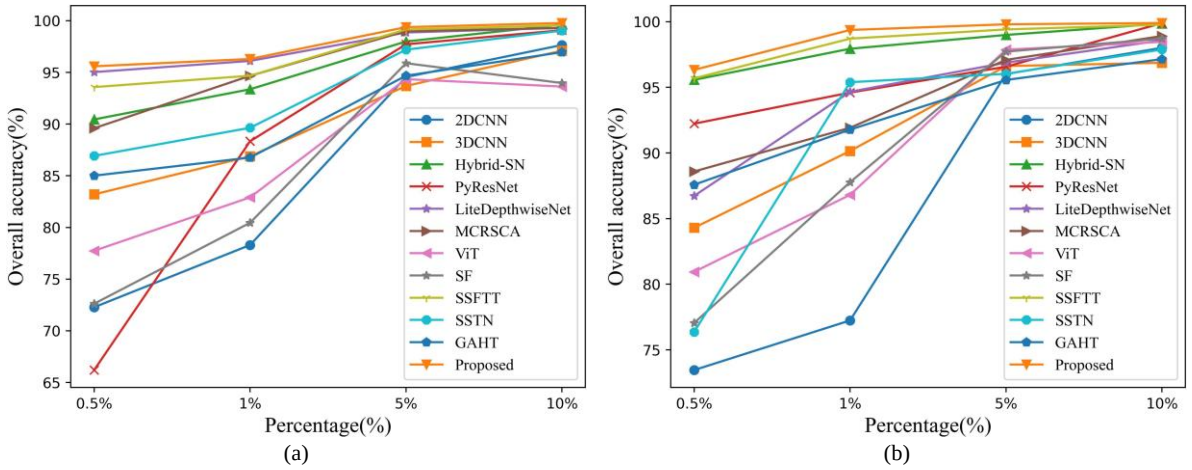


Fig. 16. Comparison of different training sample percentage. (a) Pavia dataset, (b) Salinas.

6) Comparison of different training sample percentage

The percentage of training samples plays a decisive role in hyperspectral image classification. However, the lack of labeled samples limits the training of the model. Therefore, it is necessary to verify the effectiveness of the method under small training samples. In this paper, we selected 0.5%, 1%, 5%, and 10% of Pavia and Salinas datasets for experiments. The experimental results are shown in Fig. 16. The abscissa represents the percentage of training samples, and the ordinate represents the OA value. It can be seen that our method has obtained the best results under different training sample percentage. In addition, the sub-optimal methods for Pavia and Salinas datasets are LiteDepthwiseNet and SSFTT, respectively. It is worth noting that our method has an OA value exceeding 95% on both datasets at a 0.5% sample percentage. Through small sample experiments, we verify that the proposed method can also achieve better classification accuracy under limited training samples.

IV. Conclusion

In this paper, we proposed a S^2FTNet method, which fully considers the spectral sequence and long-distance dependence of HSI data. Different from the traditional CNN based methods, the proposed method combines CNN and Transformer frameworks, making up the disadvantage that CNN is difficult to describe HSI long-distance correlation. Specifically, the proposed S^2FTNet includes two branches, SpaFormer branch and SpeFormer branch. Among them, the SpaFormer branch adopts CNN and the improved Transformer block to establish the long-distance dependence of spectral and spatial, which enriches the spectral-spatial features. The SpeFormer branch adopts the method of preserving spectral sequence, combined with the improved MHD-SA and Conv, to explore the long-distance dependence between different spectral bands. Due to the different importance of the extracted features, in order to balance the high-level semantic features extracted from the two branches, this paper also proposed an AS^2FM. Finally, in order to verify the advantages of the proposed method, three classical datasets and a new dataset are chosen and a series of experiments are carried out, which verified the effectiveness of the proposed method.

In the future, we will further explore the hyperspectral image classification method based on Transformer, and extract more representative semantic features through a small number of labeled samples to reduce the demand of the model on the number of training samples.

Acknowledgment

The authors would like to thank the Editor-in-Chief, the Associate Editor, and the reviewers for their insightful comments and suggestion. In addition, the authors would like to thank Professor Zhong's team from the State Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering, Wuhan University, China, for the WHU-Hi-LongKou dataset.

References

[1] Q. Huang, W. Li, B. Zhang, Q. Li, R. Tao and N. H. Lovell, "Blood Cell Classification Based on Hyperspectral Imaging With Modulated Gabor and CNN," in *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 160-170, Jan. 2020.

[2] G. A. Lampropoulos, T. Liu, S. -E. Qian and C. Fei, "Hyperspectral Classification Fusion for Classifying Different Military Targets," IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium, Boston, MA, USA, 2008.

[3] D. Hong et al., "Interpretable Hyperspectral Artificial Intelligence: When nonconvex modeling meets hyperspectral remote sensing," in *IEEE Geoscience and Remote Sensing Magazin*e, vol. 9, no. 2, pp. 52-87, June 2021.

[4] D. M. Tratt, K. N. Buckland, E. R. Keim and P. D. Johnson, "Urban-industrial emissions monitoring with airborne longwave-infrared hyperspectral imaging," 2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Los Angeles, CA, USA, 2016, pp. 1-5.

[5] L. Mou and X. X. Zhu, "Learning to Pay Attention on Spectral Domain: A Spectral Attention Module-Based Convolutional Network for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 110-122, Jan. 2020.

[6] 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," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 2485-2501, 2020.

[7] J. He, L. Zhao, H. Yang, M. Zhang and W. Li, "HSI-BERT: Hyperspectral Image Classification Using the Bidirectional Encoder Representation From Transformers," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 165-178, Jan. 2020.

[8] [8] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778-1790, Aug. 2004.

[9] M. D. Farrell and R. M. Mersereau, "On the impact of PCA dimension reduction for hyperspectral detection of difficult targets," in IEEE Geoscience and Remote Sensing Letters, vol. 2, no. 2, pp. 192-195, April 2005.

[10] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," in IEEE Geoscience and Remote Sensing Letters, vol. 3, no. 1, pp. 93-97, Jan. 2006.

[11] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[12] Chen, Yunpeng, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan and Jiashi Feng. "Dual Path Networks." *NIPS* (2017).

[13] Sabour, Sara, Nicholas Frosst and Geoffrey E. Hinton. "Dynamic Routing Between Capsules." ArXiv abs/1710.09829 (2017): n. pag.

[14] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligen*ce, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[15] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian and S. Kwong, "Nested Network With Two-Stream Pyramid for Salient Object Detection in Optical Remote Sensing Images," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9156-9166, Nov. 2019.

[16] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation, " *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[17] 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 Transactions on Geoscience and Remote Sensing*., vol. 54, no. 10, pp. 6232–6251, 2016.

[18] Z. Zhong, J. Li, D. A. Clausi and A. Wong, "Generative Adversarial Networks and Conditional Random Fields for Hyperspectral Image Classification," in *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3318-3329, July 2020.

[19] J. Wang, F. Gao, J. Dong and Q. Du, "Adaptive DropBlock-Enhanced Generative Adversarial Networks for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5040-5053, June 2021.

[20] R. Hang, Q. Liu, D. Hong and P. Ghamisi, "Cascaded Recurrent Neural Networks for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5384-5394, Aug. 2019.

[21] S. Hao, W. Wang and M. Salzmann, "Geometry-Aware Deep Recurrent Neural Networks for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 2448-2460, March 2021.

[22] H. Zhang, J. Zou and L. Zhang, "EMS-GCN: An End-to-End Mixhop Superpixel-Based Graph Convolutional Network for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-16, 2022, Art no. 5526116.

[23] Y. Ding, Y. Chong, S. Pan, Y. Wang and C. Nie, "Spatial-Spectral Unified Adaptive Probability Graph Convolutional Networks for

Hyperspectral Image Classification," in *IEEE Transactions on Neural Networks and Learning Systems*.

[24] M. E. Paoletti et al., "Capsule Networks for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 4, pp. 2145-2160, April 2019.

[25] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv Prepr. arXiv2010.11929, 2020.

[26] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, 2015.

[27] W. Zhao and S. Du, "Spectral–Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4544-4554, Aug. 2016.

[28] A. Ben Hamida, A. Benoit, P. Lambert and C. Ben Amar, "3-D Deep Learning Approach for Remote Sensing Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4420-4434, Aug. 2018.

[29] 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," in *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 277-281, Feb. 2020.

[30] R. Shang, H. Chang, W. Zhang, J. Feng, Y. Li and L. Jiao, "Hyperspectral Image Classification Based on Multiscale Cross-Branch Response and Second-Order Channel Attention," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-16, 2022.

[31] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification, " in *IEEE Geoscience and Remote Sensing*., vol. 59, no. 7, pp. 5966–5978, 2021.

[32] Z. Zhong, J. Li, Z. Luo and M. Chapman, "Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 847-858, Feb. 2018.

[33] He, K., Zhang, X., Ren, S., & Sun, J. (2016). " Deep residual learning for image recognition. " In Proceedings of the *IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[34] 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," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 740-754, Feb. 2019.

[35] Jie H , Li S , Gang S , et al. Squeeze-and-Excitation Networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, PP(99).

[36] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the *European conference on computer vision* (ECCV). 2018.

[37] Fu J , Liu J , Tian H , et al. Dual Attention Network for Scene Segmentation[C]// 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

[38] Wang L, Peng J, Sun W. Spatial–Spectral Squeeze-and-Excitation Residual Network for Hyperspectral Image Classification [J]. *Remote Sensing*, 2019, 11(7):884-.

[39] He, Ke, et al. "A Dual Global–Local Attention Network for Hyperspectral Band Selection." in *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-13.

[40] Mei, Shaohui, et al. "Hyperspectral image classification using attention-based bidirectional long short-term memory network." in *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021): 1-12.

[41] B. Cui, X. -M. Dong, Q. Zhan, J. Peng and W. Sun, "LiteDepthwiseNet: A Lightweight Network for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-15, 2022.

[42] Z. Meng, L. Jiao, M. Liang and F. Zhao, "A Lightweight Spectral-Spatial Convolution Module for Hyperspectral Image Classification," in *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022.

[43] X. Kang, S. Li and J. A. Benediktsson, "Spectral–Spatial Hyperspectral Image Classification With Edge-Preserving Filtering," in IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 5, pp. 2666-2677, May 2014.

[44] S. Zhong, C. -l. Chang and Y. Zhang, "Iterative Edge Preserving Filtering Approach to Hyperspectral Image Classification," in IEEE Geoscience and Remote Sensing Letters, vol. 16, no. 1, pp. 90-94, Jan. 2019.

[45] S. Zhong, S. Chen, C. -I. Chang and Y. Zhang, "Fusion of Spectral–Spatial Classifiers for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 6, pp. 5008-5027, June 2021.

[46] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification, " *Remote Sensing*, vol. 13, no. 3, p. 498, 2021.

[47] D. Hong et al., "SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-15, 2022.

[48] L. Sun, G. Zhao, Y. Zheng and Z. Wu, "Spectral–Spatial Feature Tokenization Transformer for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-14, 2022.

[49] Z. Zhong, Y. Li, L. Ma, J. Li and W. -S. Zheng, "Spectral–Spatial Transformer Network for Hyperspectral Image Classification: A Factorized Architecture Search Framework," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-15, 2022.

[50] X. Huang, M. Dong, J. Li and X. Guo, "A 3-D-Swin Transformer-Based Hierarchical Contrastive Learning Method for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-15, 2022, Art no. 5411415.

[51] J. Bai et al., "Hyperspectral Image Classification Based on Multibranch Attention Transformer Networks," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-17, 2022, Art no. 5535317.

[52] J. Zou, W. He and H. Zhang, "LESSFormer: Local-Enhanced Spectral-Spatial Transformer for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-16, 2022.

[53] R. Song, Y. Feng, W. Cheng, Z. Mu and X. Wang, "BS2T: Bottleneck Spatial–Spectral Transformer for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-17, 2022, Art no. 5532117.

[54] Mei, Shaohui and Song, Chao and Ma, Mingyang and Xu, Fulin, " Hyperspectral image classification using group-aware hierarchical transformer," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-14, 2022.

[55] Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.