Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

Tripartite-structure Transformer for Hyperspectral Image Classification

Liuwei Wan[1] | **Meili Zhou**[1] | **Shengqin Jiang**[2] | **Zongwen Bai**[1] | **Haokui Zhang**[1]

[1]School of Physics and Electronic Information, Yanan Univeristy, Yanan, China

[2]School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, China

Correspondence
Meili Zhou, School of Computer Science, Yanan Univeristy, Yanan, China.
Email: zml@yau.edu.cn

Funding Information
This research was supported by the National Key R&D Program of China Grant No.2022YFE0138600; National Natural Science Foundation of China Grant No.62266045; Key projects of Shaanxi Yan'an Grant No.2021JB-04.

Abstract

Hyperspectral images contain rich spatial and spectral information, which provides a strong basis for distinguishing different land-cover objects. Therefore, hyperspectral image classification has been a hot research topic. With the advent of deep learning, convolutional neural networks (CNNs) have become a popular method for hyperspectral image classification. However, CNN has strong local feature extraction ability but cannot deal with long-distance dependence well. Vision Transformer (ViT) is a recent development that can address this limitation, but it is not effective in extracting local features and has low computational efficiency. To overcome these drawbacks, we propose a hybrid classification network that combines the strengths of both CNN and ViT, names Spatial-Spectral Former(SSF). The shallow layer employs 3D convolution to extract local features and reduce data dimensions. The deep layer employs a spectral-spatial transformer module for global feature extraction and information enhancement in spectral and spatial dimensions. Our proposed model achieves promising results on widely used public HSI datasets compared to other deep learning methods, including CNN, ViT, and hybrid models.

KEYWORDS
hyperspectral image classification, Vision Transformer (ViT), 3D-convolutional neural networks (3D-CNN)

1 | INTRODUCTION

Hyperspectral remote sensing uses sensors/imaging spectrometers to capture images of a target area in hundreds of wavelength bands, allowing for simultaneous acquisition of spatial and spectral information. Additionally, as spectral imaging technology continues to advance, improvements in spatial and spectral resolution are being made continuously. Compared to standard remote sensing images like multispectral remote sensing images, hyperspectral images provide a more comprehensive spectral band coverage, capturing a vast amount of information through numerous wavelength bands. The resulting 3D data blocks effectively capture and combine spatial and spectral information, making HSI classification techniques widely applied on medical [1], military [2], agricultural [3] and water resources management [4] scenarios.

However, HSI is a powerful technology that captures an extensive range of spectral bands for each spatial pixel, providing rich and detailed information about the Earth's surface. While this abundance of data presents valuable opportunities for ground object classification, it also poses challenges in extracting relevant and discriminative features effectively. Therefore, researchers have been actively exploring various methods to tackle this feature extraction problem in HSIs. In the early stages of HSI classification, traditional approaches mainly relied on shallow feature extraction techniques like Support Vector Machine (SVM) [5,6,7,8], Principal Component Analysis (PCA) [9], and Morphological Profile (MP) [10]. These methods served as initial steps in understanding the data and its characteristics, but they showed limitations when confronted with the complexity and intricacy of real-world HSI data. One of the main issues with these early models was their shallow structures, which prevented them from capturing the intricate patterns and relationships present in the high-dimensional HSI data. As a result, their classification performance often fell short of expectations, hindering their broader applicability in practical scenarios.

Since 2012, deep learning has gained attention in the field of computer vision, and numerous deep learning-based hyperspectral image (HSI) classification methods have been proposed. Deep learning-based HSI classification methods have achieved impressive performance by efficiently extracting data features. Convolutional neural networks (CNNs) are the most commonly used deep learning method for HSI classification due to their ability to capture the spectral and spatial information of HSI. CNN-based approaches are typically categorized into 1D-CNN, 2D-CNN, and 3D-CNN, depending on their convolution types. 1D-CNNs are used to extract deep spectral features [11,12], 2D-CNNs extract deep spatial features of pixels from spectrally compressed HSI blocks [13,14], and 3D-CNNs extract both spectral-spatial features [15,16]. HSI data is typically represented as 3D structures, so 3D-CNNs are commonly used to extract spectral-spatial features without the need for any pre-processing or post-processing [15]. Very recently, several variations of 3D-CNNs, such as migration learning [17,18] and neural structure search [19,20], have been proposed for HSI classification, providing additional evidence for the effectiveness and versatility of 3D-CNNs in this application. In summary, deep learning-based HSI classification methods have demonstrated superior performance, with 3D-CNN being the most commonly used model due to its ability to extract both spectral and spatial information without any preprocessing.

Although 3D-CNN-based methods for hyperspectral image (HSI) classification have shown promise performance, there is still room for improvement. While the local connectivity and shared weights of CNNs effectively capture local correlations and promote high feature extraction efficiency, they also result in a limited effective receptive field. This limitation can hinder the performance of CNN-based HSI classification methods. In 2020, Dosovitskiy et al. [21] proposed vision transformer(ViT), which is the first self-attention mechanism [22] based image classification model. ViT achieves higher classification accuracy on ImageNet-1k compared with previous various CNN models, thanks to the Transformer is able to process long-distance dependencies effectively. Inspired by this, some researchers attempt to introduce ViT into HSI classification. For example, Hong et al. [23] proposed using grouped spectrum embedding to obtain neighboring spectra features, and the Transformer multiple encoders with added jump connections to reduce the loss of valuable information during propagation. As CNN models can effectively model local relationships and ViT can extract information globally, some researchers have combined the two to create new structures. Methods like SSFTT [24] and SST [25] use CNN to extract shallow spatial features, treat each point in the final feature map as a sequence of words, and apply the Transformer on it. These methods extract spatial features using CNN, and spectral features using ViT, achieving higher classification accuracy than previous CNN-based algorithms.

Following the research line of combining the strengths of CNN and ViT, we attempt to design a more efficient hybrid model for HSI classification. Unlike previous approaches that use the vanilla ViT structure designed for RGB images, our proposed architecture is tailored for HSIs, resulting in better performance. In previous ViT-based HSI classification methods, ConvNets and ViTs are typically combined to overcome the long-distance dependence problem. However, these approaches inherit the disadvantages of raw ViTs, such as a large number of parameters, low computational efficiency, and difficulty in training. Additionally, the vanilla ViT is designed for RGB images, which have different characteristics compared to HSI. Although using the vanilla ViT for HSI classification is possible, it is not optimal.

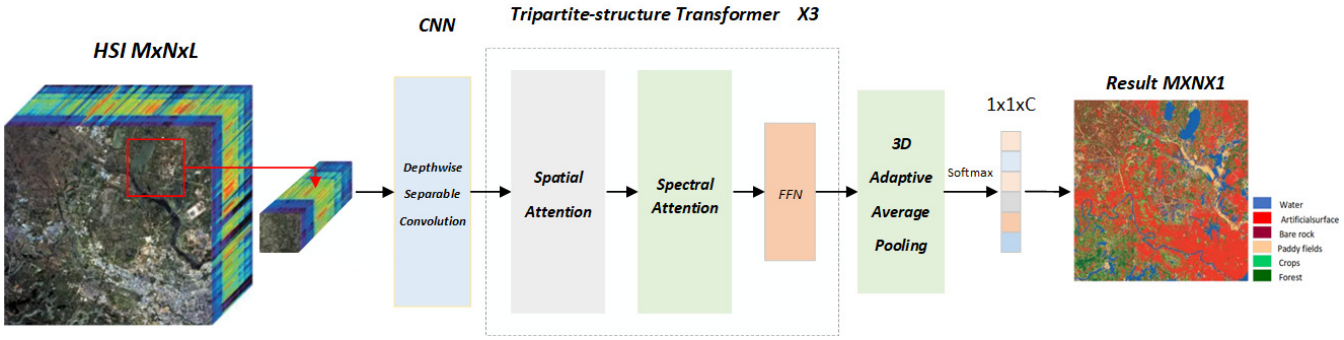


FIGURE 1 The overall workflow.

Our approach starts with the structural features of HSI and tailors a hybrid model specifically for HSI classification. We first use a 3D-CNN to extract local spectral-spatial features in shallow layers to compress the high dimension 3D data into a compacted 3D cube. This design fully utilizes the computational efficiency of CNN and reduces the computational load of ViT. We then design a tripartite-structure transformer block for HSI classification that includes multiple Spatial Attention, one Spectral Attention, and a channel mixer of FFN. The shallow features are input to Spatial Attention, and the deep Spatial features are extracted by multiple Spatial Attention. The deep Spectral features are extracted by inputting them into Spectral Attention. The channel mixer interacts fully with the information extracted from each channel. Our experimental results on three typical HSI classification datasets from Pavia Center, Pavia University, and Houston University validate the efficiency of our proposed network.

The remainder of this paper is organized as follows. Section II reviews related work. A detailed description of our method is given in Section III. Section IV gives the algorithm implementation details and compares the accuracy of our proposed SSF method with that of other methods. Finally, we conclude this work in Section V.

2 | RELATED WORK

2.1 | Hyperspectral Image Classification via CNNs

In recent years, the technology of classifying hyperspectral images based on CNN has achieved impressive performance, and CNNs have gone through three main stages for hyperspectral image classification.

From 2015 to early 2016, the focus of HSI classification was mainly on 1D-CNN and 2D-CNN. Classification methods based on 1D-CNN typically use 1D-CNN to convolve the spectral directions of HSIs and extract spectral features [11,26], which is known as the spectral feature-based classification method. On the other hand, 2D-CNN-based classification methods downscale HSIs on their spectral dimensions and extract neighborhood information while retaining its original spatial structure. The extracted information is then processed at a deeper level using 2D-CNN to extract deeper spatial features, which are then used to complete HSI classification [27,28]. These methods are known as spatial feature-based classification methods. However, methods that solely use 2D CNNs for classification may not retain structural information well, resulting in much smoother visual results compared to 1D CNN methods.

The second phase of HSI classification focused on combining 1D-CNN and 2D-CNN to improve accuracy. By taking advantage of the strengths of both models, a two-channel CNN structure was constructed. Yang et al. [29] used 2D-CNN to extract spatial features from downsampled hyperspectral images, 1D-CNN to extract spectral features from spectral vectors, and completed classification using fully connected layers and Softmax. Similarly, Zhang et al. [30] used a pyramid structure to fuse multilayer features and obtain spatial-spectral features.

The third stage of HSI classification is characterized by the increasing use of 3D-CNNs. Due to the 3D structure of HSIs, 3D-CNNs are well-suited for extracting spatial-spectral features [15]. The key advantage of 3D-CNNs lies in their ability to process the entire hyperspectral cube as a whole, treating it as a 3D volume. This enables them to capture the spatial-spectral relationships present within each pixel, thus offering a more holistic understanding of the hyperspectral data. As a result, a series of optimized 3D-CNN structures, including residual networks, lightweight models, and migration learning, have become more mainstream.

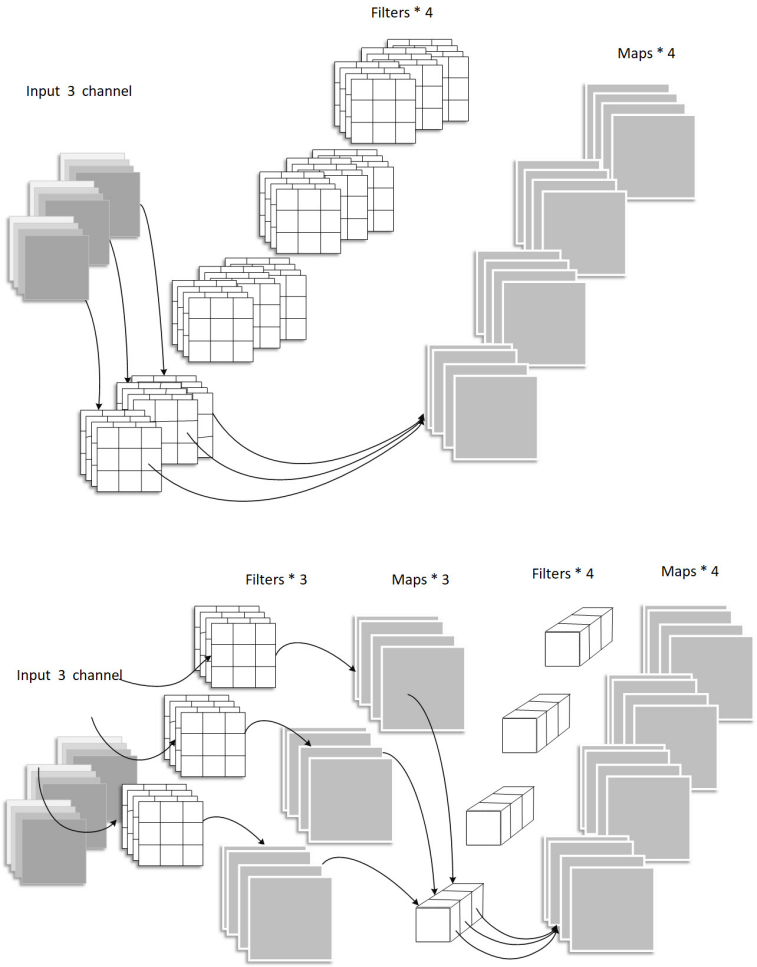


FIGURE 2 Difference between standard 3D-CNN and Depthwise Separable convolution.

For example, Meng et al.[31] proposed a lightweight spectral-space convolutional HSI classification module (LSSCM) to reduce network parameters and computational complexity, while Zhang et al.[17] added migration learning strategies to a lightweight model and proposed a model HSI classification module (LWNet) that achieved better classification accuracy using a deeper network, fewer parameters, lower computational costs, and fewer training samples. Compared with 1D-CNNs and 2D-CNNs, 3D-CNNs are more intuitive and efficient in classification results, as they are better suited to the data structure of HSIs. Inspired by these findings, in our proposed hybrid structure, we use 3D-CNNs in the shallow layer for spectral-spatial feature extraction.

2.2 | Vision transformers

Vaswani et al[22] introduced the Transformer structure for natural language processing (NLP) through a thorough examination of the attention mechanism. Compared to the RNN structure, which was widely used in NLP previously, the Transformer can process the sequence of elements in parallel, leading to significantly improved computational efficiency and better capturing the relationships between any pair of elements in the input sequence. As a result, the Transformer structure has gradually replaced the RNN structure and become dominant in the field of NLP.

In 2020, Dosovitskiy et al. introduced the Transformer structure to computer vision tasks with their proposed Vit model based on image features[21]. The Vit model segments each image into patches, which are then used as input sequences for feature extraction. A position embedding is introduced in Vit to ensure sensitivity to the position of the input patches, and an additional class token is set to perform the final classification. This novel approach allowed the Transformer to excel in image-based tasks, previously dominated by Convolutional Neural Networks (CNNs). However, despite its success, Vit had some limitations

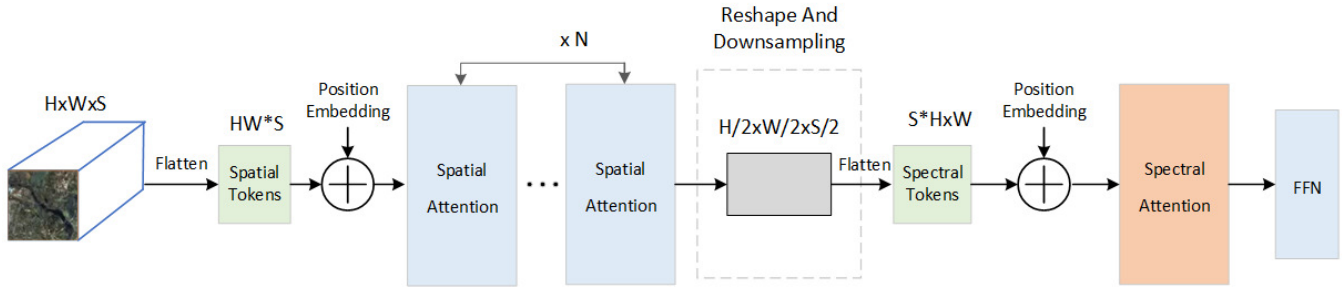


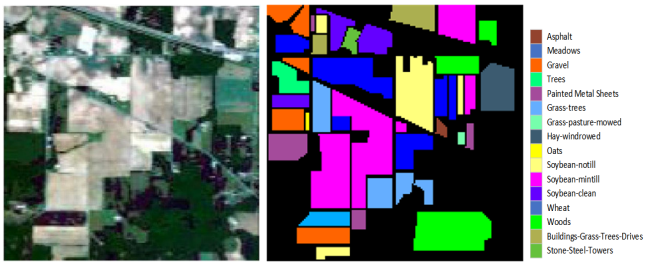
FIGURE 3 Tripartite-structure Transformer.

that hindered its practicality in certain applications. One of the major drawbacks was its high parameter count, resulting in substantial computational overhead. Additionally, training Vit was relatively challenging, requiring careful optimization to achieve satisfactory results. To address these issues, a series of Vit variants were proposed. Hong et al.[23] proposed the use of grouped spectral embeddings instead of all spectral bands for feature extraction, to model local detailed spectral differences. Cross-layer jump connections were added to multiple encoders of the Transformer to reduce information loss during layer-by-layer propagation. Some researchers have attempted to optimize Vit by incorporating CNNs. PVT[32] implements a pyramid structure by inserting convolutions at each stage of Vit to construct a hierarchical multilevel structure and reduce the number of tokens. Swin Transformer, proposed by Liu et al.[33], reduces the computational cost of attention by limiting it to pixels within a small window. It also introduces a shift-window-based MSA, which allows attention to span different windows. Swin Transformer achieves higher accuracy than previous CNN models on tasks such as dense prediction.

2.3 | Hybrid structures combining ConvNet and vision transformers

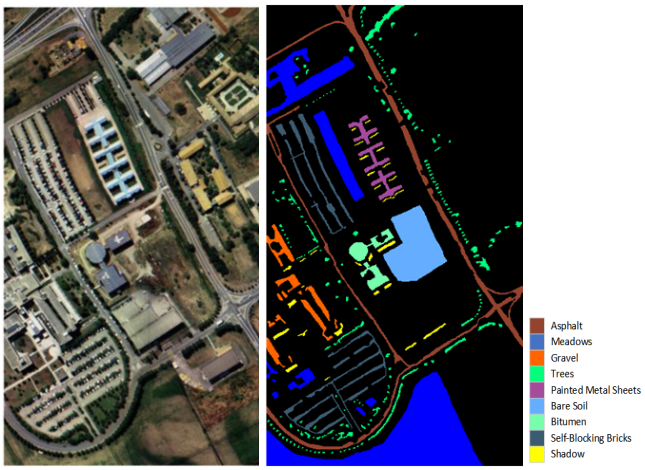
CNN effectively captures local correlations but cannot handle the long-range dependence of data well. Vit is good at extracting features from the global. However, its classification results are unsatisfactory because the original Vit has the drawbacks of a large number of parameters and low computational efficiency. Therefore, combining Vit and CNN to form new structures is becoming popular. In LeviT[34], Graham et al. combined ConvNets and transformers, resulting in a significant improvement in the speed/accuracy tradeoff compared to previous ConvNet and ViT models. BoTNet[35] utilizes multi-headed attention instead of standard convolution in the last few blocks of ResNet. Vit-C[36] applies convolution in the Patch Embedding step in the traditional Vit architecture. ConViT[37] employs a gated positional self-attentive mechanism with Soft convolution induction bias. The CMT[38] block consists of a local perceptual unit based on deep convolution and a lightweight Transformer module. CoatNet[39] introduces a novel transformer module that merges convolution and self-attention to effectively capture both local and global information. In MobileViT[40], the deeper stages of MobileNetv2[41] are replaced with their proposed MobileViT block. In ParCNet[42], Zheng et al proposed position aware circular convolution (ParC) to capture global features and combine the proposed ParC operation with traditional convolution to design hybrid structure.

Very recently, researchers have explored the use of hybrid structures for hyperspectral image (HSI) classification. One approach proposed by He et al.[25] involves fine-tuning a convolutional neural network (CNN) pre-trained by VGG to extract spatial features for each HSI band. The spatial features are then fed into a transformer to extract spectral features. To address the issue of limited HSI samples, the authors incorporate a migration learning strategy. Another approach, proposed by Wang et al.[43], divides the HSI spectrum into several equal-length subbands and applies 3D convolution with different kernels to fuse features. The fused features are connected to the residuals of the original image after linear embedding and classified using a Transformer. The authors use a centre mask technique to reconstruct the same centre pixel as much as possible by a decoder, which improves the learning of the relationship between the centre pixel and the domain pixel. However, this method only extracts shallow spatial features and relies on self-attention to extract deep spectral features. A third approach, proposed in the recent study by SSFTT[24], employs 3D-CNN to extract shallow spectral features and 2D-CNN to extract shallow spatial features. The extracted features are input into a Transformer module for feature learning after Gaussian weighted feature tagging. However, the authors only consider deep feature extraction for spatial features and neglect deep spectral features.



(a) (b)

FIGURE 4 Indian Pines dataset. (**a**) False-color map. (**b**) Ground-truth map.



(a) (b)

FIGURE 5 Pavia University dataset. (**a**) False-color map. (**b**) Ground-truth map.

3 | THE PROPOSED METHOD

In this section, we provide a detailed description of our proposed method(SSF). First, we present the overall framework of our classification architecture. Next, we delve into the two major modules that make up the framework: the depthwise 3D convolution module and the tripartite-structure transformer module(TST).

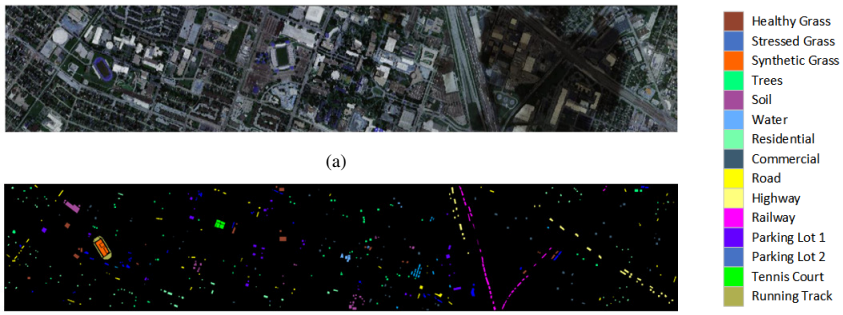
3.1 | The overall workflow

As illustrated in Figure 1, the overall workflow of our proposed classification framework consists of three major parts, including sample extraction, spectral-spatial feature extraction and 3D adaptive average pooling based classification.

During the sample extraction stage, we extract a cube of size $S \times S \times L$ as a sample, where S is the spatial size and L is the number of spectral bands. Each cube is extracted from a neighborhood window that is centered around a pixel. Note that S is typically an odd number to ensure that the center of the extracted cube corresponds to the pixel that needs to be classified. This cube serves as the original features of the center pixel.

For the spectral-spatial feature extraction, we designed a hybrid structure that incorporates depth-wise 3D convolution in the shallow layer and a tripartite-structure transformer module(TST) in the deep layer. Due to the high spectral resolution of hyperspectral imagery (HSI), applying a vanilla transformer module to HSI samples would result in high computation costs. On the other hand, 3D convolution is well-suited for extracting 3D features from HSI, and it has a development history of more than ten years. Moreover, current GPUs and deep learning tools are highly optimized for convolution operations, making 3D convolution highly computationally efficient. By applying 3D convolution in the shallow layer, we can leverage these advantages to 1) extract preliminary spectral-spatial features, and 2) reduce the feature dimension to decrease computation overload, which prepares the data for the transformer modules. After reviewing the spectral structure of HSI, we proposed a tripartite-structure transformer module and applied it in the deep layer.

For the final classification step, we use 3D adaptive average pooling to adjust the size of the features to a fixed value. Using adaptive average pooling is important to maintain a concise architecture. If we were to flatten the output of the last unit into a vector, as is done in other conventional structures, we would have to adjust the dimensionality of every fully connected layer for each HSI data set, as different HSI data sets have varying numbers of bands. By using adaptive average pooling, we can avoid this problem and keep the architecture consistent across different HSI data sets.



(a)

(b)

FIGURE 6 Houston 2013 dataset. (**a**) False-color map. (**b**) Ground-truth map.

3.2 | The depthwise 3D convolution

As mentioned above, we use 3D convolution in the shallow layer to extract primary spectral-spatial features as preparation for using the transformer. While convolution is computationally efficient, the traditional 3D convolution can still lead to high computation costs in this stage. To address this issue, we use depthwise 3D convolution instead. A comparison between the traditional 3D convolution and our adopted depthwise 3D convolution is illustrated in Figure 2.

Suppose the input feature cube F has the shape (s_i, w_i, h_i, c_i). The output feature cube, denoted as O, has the shape (s_o, w_o, h_o, c_o) after convolution. Here, s, w and h refer to the length, width, and height of the matrix, respectively, and c denotes the number of channels. When using conventional convolution, a convolution kernel K of size $k \times k \times k$ is applied, where k is the length of the filter. The output matrix of conventional convolution is calculated as follows:

$$O_{x,y,z,n} = \sum_{a,b,c,m} K_{a,b,c,m,n} F_{x+a-1,y+b-1,z+c-1,m}$$

where a, b, c and x, y, z denote the spatial locations of the elements for a conventational convolution kernel of size $k \times k \times k \times c_m \times c_n$.When using depthwise separable convolution, the input features are first passed through a depthwise convolution (one convolution kernel is responsible for one channel, and the number of channels of the generated feature map is the same as the number of input channels), The output matrix of depthwise convolution is calculated as follows:

$$\hat{O}_{x,y,z,m} = \sum_{a,b,c,m} \hat{K}_{a,b,c,m} F_{x+a-1,y+b-1,z+c-1,m}$$

where a, b, c and x, y, z denote the spatial locations of the elements for a depwise convolution kernel of size $k \times k \times k \times c$. And then passed through a pointwise convolution (a $1 \times 1 \times 1$ sized convolution kernel that combines the maps of the previous layer weighted in the direction of depth to generate a new feature map) for feature extraction.

In depthwise 3D convolution, each input channel is processed independently. As illustrated in Figure 2, if the input has 3 channels (denoted as m) and the output has 4 channels (denoted as n), the number of parameters for the conventional convolution operation is $k \times k \times k \times 3 \times 4$, and the computation cost is $k \times k \times k \times 3 \times H_o \times W_o \times S_o \times 4$. For the depthwise separable convolution, the number of parameters required for the operation is $k \times k \times k \times 3 \times 1 + 1 \times 1 \times 1 \times 3 \times 4$, and the computation cost is $k \times k \times k \times 1 \times H_o \times W_o \times S_o \times 3 + 1 \times 1 \times 1 \times 3 \times H_o \times W_o \times S_o \times 4$.

Compared to the standard 3D-CNN, the depthwise separable convolution reduces the number of parameters to approximately 1/4 and the computation cost to approximately $(k^3 + 4)/(4 \times k^3)$.

3.3 | Tripartite-structure Transformer

After extracting the shallow features of hyperspectral imaging (HSI) using depthwise 3D convolution, as described in Section 3.2, the proposed method further learns the relationship between high-level semantic features using the tripartite-structure transformer module(TST). This module comprises three main components, as shown in Figure 3.

TABLE 1 Training and Test sample

NO	Indian Pines			Pavia University			Houston2013		
	Class	Training	Test	Class	Training	Test	Class	Training	Test
1	Alfalfa	5	41	Asphalt	332	6299	Healthy grass	125	1126
2	Corn-notill	143	1285	Meadows	932	17717	Stressed grass	125	1129
3	Corn-mintill	83	747	Gravel	105	1994	Synthetic grass	70	627
4	Corn	24	213	Trees	153	2911	Trees	124	1120
5	Grass-pasture	48	435	Painted metal sheets	67	1278	Soil	124	1118
6	Grass-tree	73	657	Bare Soil	251	4778	Water	33	292
7	Grass-pasture-mowed	3	25	Bitumen	67	1263	Residential	127	1141
8	Hay-windrowed	48	430	Self-Blocking Bricks	184	3498	Commercial	124	1120
9	Oats	2	18	Shadows	47	900	Road	125	1127
10	Soybean-notill	97	875				Highway	123	1104
11	Soybean-mintill	245	2210				Railway	123	1112
12	Soybean-clean	59	534				Parking Lot 1	123	1110
13	Wheat	20	185				Parking Lot 2	47	422
14	Woods	126	1139				Tennis Court	43	385
15	Buildings-Grass-Trees-Drives	39	347				Running Track	66	594
16	Stone-Steel-Towers	9	84						
	Total	1024	9225	Total	2138	40638	Total	1502	13527

The first part of the proposed method extracts the relationship between high-level semantic features from the input feature map using spatial and spectral correlation extraction. After performing depthwise separable convolution, the HSI is transformed into a 3D feature map of size $H \times W \times S$. We first perform multiple spatial feature extractions followed by a single spectral feature extraction on the input features. To extract spatial feature relevance, we flatten the input 3D feature map into an S-dimensional 2D spatial vector of size HW and use position embedding to tag the position information of each semantic token. The Spatial attention module, which is based on multi-head attention, then extracts the spatial relevance of the input feature map using global information.

For spectral feature relevance extraction, we flatten the input 3D feature map into an HW-dimensional 2D spectral vector of size S and tag the position of each semantic token using position embedding. We then use the Spectral attention module, also based on multi-head attention, to extract the spectral feature relevance of the input feature map.

In both the spatial and spectral attention modules, three learnable weight matrices (W_Q, W_K, W_V) are predefined to learn multiple meanings. The tokens are linearly mapped into 3-D invariant matrices by these matrices, including the query Q, keys K, and values V. Attention scores are calculated using all Q and K, and score weights are calculated using the softmax function. In summary, attention is defined as follows:

$$Attention\,(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$d_k$is the dimension of Q or K.

The process of mapping query, key, and value to different learnable projections multiple times and then concatenating these results is known as multi-head attention. In this process, each result of the parallel computation of these attentions is called a head.

$$MutiHead\,(Q, K, V) = Concat\,(head_1 ... head_h)\,W^O$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

where h is the number of heads.W_O is the parameter matrix. $W_i^Q \in R^{d_m \times d_q}$, $W_i^K \in R^{d_m \times d_k}$, $W_i^V \in R^{d_m \times d_v}$, $W_i^O \in R^{d_m \times d_v \times h}$, where d_m is the number of tokens.

The second part of the proposed method is the Reshape and Downsampling module, which reshapes the output features from the previous layer into cubes of size $H \times W \times S$. We also incorporate a pyramid structure into the network as it deepens, inspired by the PVT architecture. The resolution of the feature map is reduced after reshaping, while the number of channels is doubled.

The third part is the Feed-Forward Network (FFN) module, which performs channel mixing to fully fuse information from different channels.

TABLE 2 Results of Indian Pines classification

	IndianPine				
Models	LSSCM	SpectralFormer	Vit	Hybrid Vit	SSF
1	0	77.74	58.38	97.56	1
2	78.67	90.18	74.11	81.48	97.5
3	79.65	92.85	75.00	76.57	98.68
4	91.08	89.26	83.89	66.67	1
5	74.02	93.4	76.33	87.13	97.9
6	91.63	96.36	97.72	97.26	99.66
7	0	86.27	57.63	1	1
8	1	71.84	49.79	93.49	1
9	0	77.84	40.25	66.67	1
10	71.09	1	95.06	86.63	98.3
11	91.67	93.89	83.04	94.66	95.66
12	69.29	76.97	43.33	64.61	98.42
13	27.03	97.78	1	95.68	1
14	96.75	69.23	87.18	97.63	1
15	66	90.9	72.73	91.07	98.21
16	0	1	60.00	79.76	1
OA	81.92	83.37	64.55	88.08	97.91
AA	58.56	87.72	72.15	86.05	99.02
k	79.24	81.14	60.16	86.36	97.57

4 | EXPERIMENTS

Experiments are conducted on a server with an Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz, 512 GB of memory,and Nvidia Tesla V100 32 GB graphics card. The proposed model is implemented by using the opensource framework Pytorch 1.7.1.

4.1 | Data Description

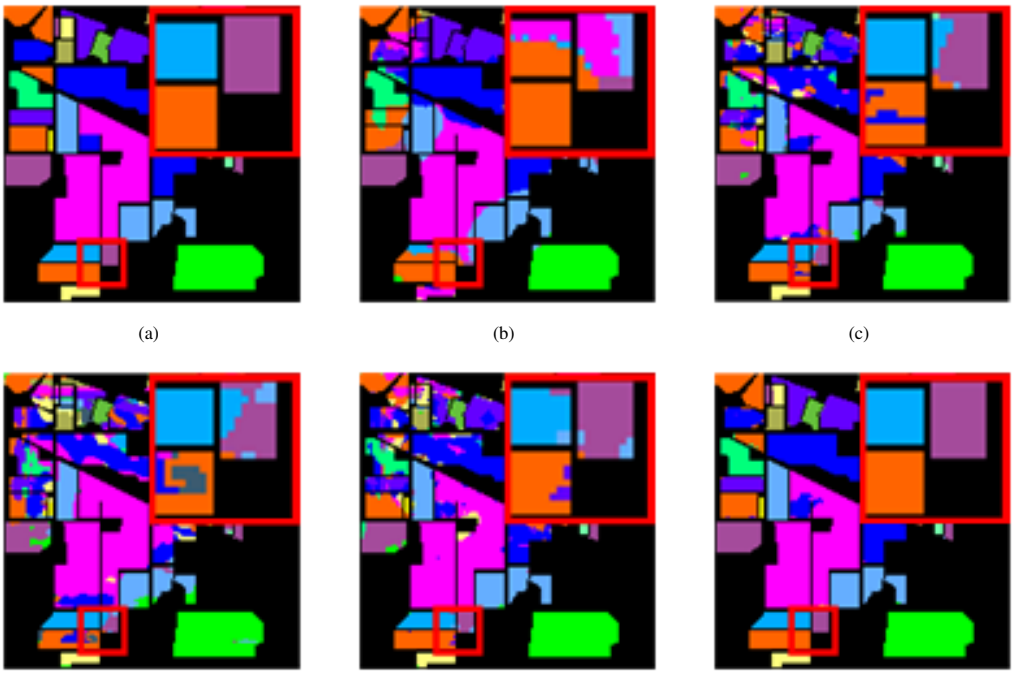
To verify the validity of our proposed hybrid structure, we use three public datasets: Indian Pines, Pavia University, and Houston University. The false-color composites and ground truth maps of these three HSIs are presented in Figure 4, Figure 5, Figure 6.

Indian Pines Data: In 1992, the airborne visible/infrared imaging spectrometer (AVIRIS) sensor was captured at the Indian Pines test site in Indiana, USA to generate the Indian Pines data set. It consists of 145 $\times$ 145 pixels, and the uncorrected data contains 224 spectral bands ranging from 400 nm to 2500 nm. After removing 24 bands covering the absorption region, the remaining 200 bands are used for 16 land-cover classifications. The false-color and ground-truth maps are shown in Figure 4(a) and Figure 4(b), respectively.

Pavia University Data: In 2001, the reflective optics system imaging spectrometer (ROSIS) sensor was captured at Pavia University in Italy to generate the Pavia University dataset. It consists of 610 $\times$ 340 pixels, and the uncorrected data contains 115 spectral bands ranging from 430 nm to 860 nm. After removing the 12 noisiest bands, the remaining 103 bands are used in 9 land-cover classifications.The false-color and ground-truth maps are shown in Figure 5(a) and Figure 5(b), respectively.

Houston2013 Data: The Houston dataset was acquired by the ITRES-CASI 1500 sensor at Houston University and its surrounding area, initially for the 2013 IEEE GRSS Data Fusion Competition. It consists of 3491905 pixels and contains 144 spectral bands ranging from 364 nm to 1046 nm, which were used in 15 land-cover classifications. The false-color and ground-truth maps are shown in Figure 6(a) and Figure 6(b), respectively.

Table 1 lists the land-cover category names, the number of training samples, and test samples regarding these three datasets. For the Indian Pines and Houston 2013 dataset, the table indicates that a random 10% of the total sample number was used as the training set, while the remaining 90% served as the test set. The dataset encompasses various land-cover categories, each with a specific number of samples for training and testing. Regarding the Pavia University dataset, a slightly different approach was adopted. Here, a random 5% of the total samples were selected as the training set, and the remaining 95% were designated as the test set.



(a) (b) (c)

(d) (e) (f)

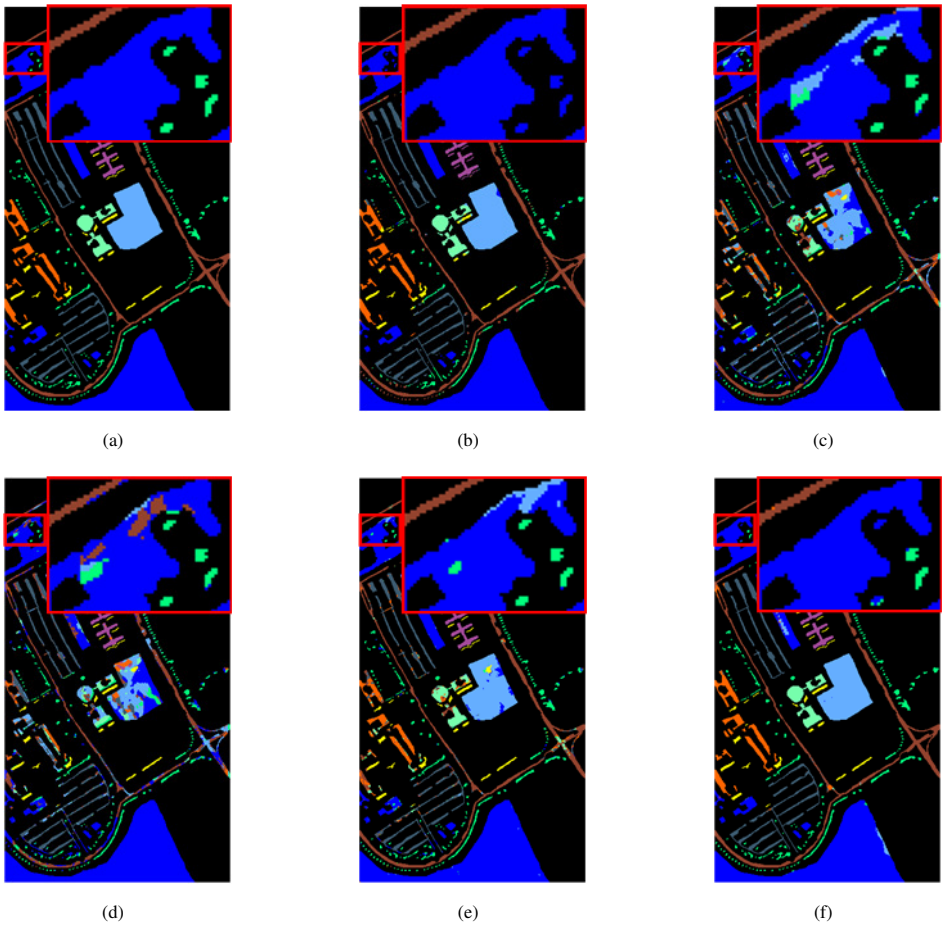
FIGURE 7 Classification maps of the Indian Pines dataset.(**a**) Ground-truth map. (**b**) LSSCM. (**c**) SpectralFormer. (**d**) Vit. (**e**) Hybird Vit. (**f**) SSF.

TABLE 3 Results of Pavia University classification

Pavia University					
Models	LSSCM	SpectralFormer	Vit	Hybrid Vit	SSF
1	98.87	82.00	81.03	90.98	97.57
2	99.96	71.40	65.30	98.26	98.14
3	94.23	64.68	64.08	69.21	99.28
4	72.11	94.88	87.74	91.34	99.36
5	99.92	99.82	99.01	99.84	1
6	98.95	94.12	94.55	91.73	99.92
7	98.41	85.93	85.11	71.5	1
8	98.31	97.56	84.99	86.76	99.32
9	55.56	92.83	99.12	96.78	1
OA	96.22	80.74	76.45	92.64	98.64
AA	90.7	87.02	84.55	88.49	99.29
k	94.98	75.56	70.26	90.22	98.18

4.2 | Implementation Details

We construct three networks with the same contour structure for the three different datasets, differing slightly in their parameters. Specifically, all three datasets use a cube with a spatial resolution of 27×27 as the input sample, and all use AdamW as the optimizer with a weight_decay of 1e-5. Since IndianPines and Houston are relatively challenging to train, their learning rate is set to 1e-5, and Pavia According to the cosine annealing strategy, the first 30% of epochs of the three datasets are in the warm-up stage. The learning rate gradually increases from 10% of the initial learning rate to the initial learning rate. The learning rate is gradually increased from ten per cent of the initial learning rate to the initial learning rate. Then seventy per cent of the epoch learning will gradually decrease, and the learning rate is minimized to one per cent of the input learning rate. In the training process, we set the epoch of Indian Pine to 1200 and the batch size to 64 because of its slow convergence, and the epoch of the other two datasets to 600 and the batch size to 128. The classification performance of each model is evaluated.



(a) (b) (c)

(d) (e) (f)

FIGURE 8 Classification maps of the Pavia University dataset.(**a**) Ground-truth map.(**b**) LSSCM. (**c**) SpectralFormer. (**d**) Vit. (**e**) Hybird Vit. (**f**) SSF.

4.3 | Classification Results and Analysis

In this section, we select several representative methods for comparison experiments to demonstrate the effectiveness of our proposed model. They are LSSCM[31], Vit[21], SpectralFormer[23], and our proposed SSF model. LSSCM is based on convolution, and Vit and SpectralFormal are based on Transformer; The authors of the corresponding papers provide all models. And we add simple 3D-CNN modules to the original model of Vit to form Hybrid Structures of the most basic CNN and Transformer, denoted as Hybrid-Vit. The original Vit representation is performed using the Vit model provided in SpectralFormer. Table 2, Table 3, Table 4 shows the results of comparing these algorithms on the three datasets.

Table 2 represents the classification results of these algorithms on the Indian Pines dataset. As seen from the table, the classification accuracy of our method is 97.91%, which is higher than the results of all the other CNN and Transformer variants. Analyzing the results, we observe that ViT solely focusing on extracting spectral information resulted in the lowest accuracy among all the methods. This outcome can be attributed to the limited capability of ViT to effectively capture spatial patterns, which are crucial for accurate HSI classification. SpectralFormer, while an improvement over ViT, still primarily processes spectral features and failed to deliver satisfactory results. Though LSSCM, an enhanced model based on 2D-CNN, demonstrated better performance than ViT and SpectralFormer, its accuracy remained unsatisfactory. On the other hand, the Hybrid Vit method, which combines elements of spatial and spectral information extraction, outperformed ViT, SpectralFormer, and LSSCM, reinforcing the importance of considering both spatial and spectral features for HSI classification tasks. The comparison between Transformer and CNN structures indicates that Transformer-based models generally exhibit better results than traditional CNN-based models. However, the initial ViT approach was not fully suitable for HSI classification. Therefore, we need to customize Vit according to the characteristics of HSI, such as spectralformer, but HybridVit with a hybrid structure has a better result. And

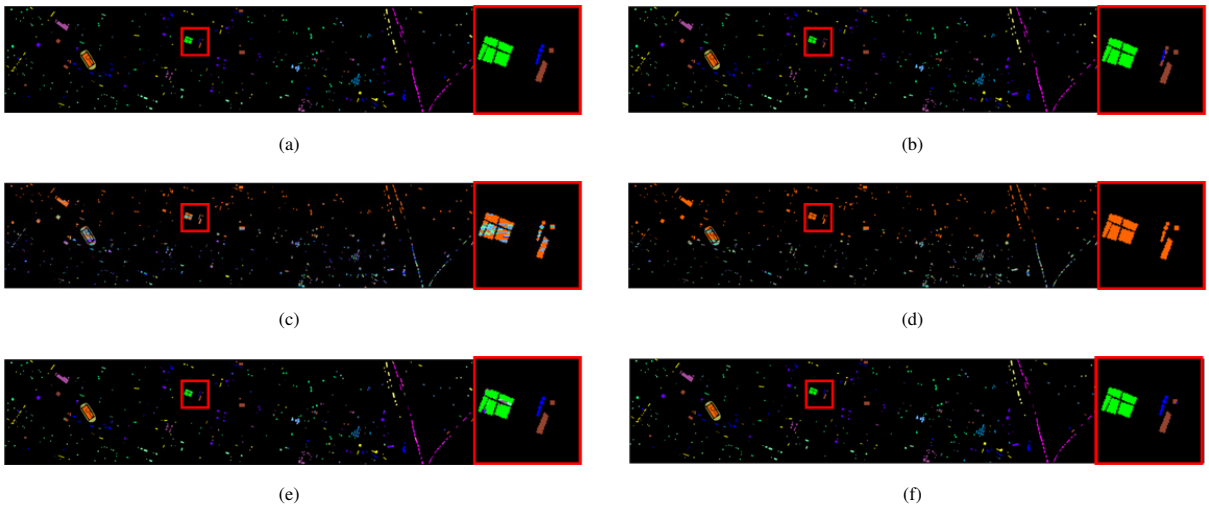


FIGURE 9 Classification maps of the Houston 2013 dataset. (**a**) Ground-truth map.(**b**) LSSCM. (**c**) SpectralFormer. (**d**) Vit. (**e**) Hybird Vit. (**f**) SSF.

TABLE 4 Results of Houston 2013 classification

Houston 2013					
Models	LSSCM	SpectralFormer	Vit	Hybrid Vit	SSF
1	93.32	84.71	82.24	93.70	99.50
2	91.46	98.68	94.92	93.15	99.80
3	93.58	96.63	99.60	1	1
4	85.63	99.24	99.24	96.41	97.59
5	99.24	99.43	97.06	97.82	1
6	83.27	94.41	99.30	78.18	1
7	95.65	84.33	86.29	98.51	98.13
8	95.46	77.68	50.99	78.92	99.30
9	84.59	86.40	70.63	92.95	98.70
10	99.81	43.24	75.58	98.27	1
11	97.9	79.13	83.49	86.00	99.49
12	97.9	81.84	46.78	88.65	99.39
13	88.44	69.12	58.59	61.81	99.73
14	1	1	92.71	88.15	1
15	85.92	98.94	99.37	94.30	98.11
OA	93.46	84.82	80.51	90.99	99.23
AA	92.81	86.25	82.45	89.39	99.32
k	92.92	83.54	78.87	90.25	99.17

we combined CNN and customize ternary-Vit to achieve the best performance. The classification results of these methods are shown in Figure 7.

Table 3 shows the classification statistics of the five methods on the Pavia University dataset, and our method beats the other four algorithms based on three evaluation metrics. However, the classification accuracy of our proposed method for the "Asphalt" and "Meadows" categories is not as good as that of LSSCM, probably due to the dominance of spatial information in these two categories, which is not sensitive to spectral information. Some of the information in "Meadows" was misclassified, and most was incorrectly assigned to "Bare Soil". Despite these challenges, it is essential to highlight that our proposed method still achieved an impressive overall classification accuracy of 98.64%. This outstanding accuracy demonstrates the strength and potential of our hybrid model, especially in handling the complexities inherent in HSI data. The classification results of these methods are shown in Figure 8. The superior performance of our proposed method is evident in the distinct and well-separated class boundaries, showcasing its ability to effectively capture both spectral and spatial information, thus leading to more accurate and reliable classification outcomes.

Table 4 represents the classification results of these algorithms on the Houston 2013 dataset. Due to the dataset's complexity, we recognized the need to enhance the training process by increasing the number of training samples, which was set to 15% of the total samples. This augmentation aimed to provide the algorithms with a more robust and diverse training set, potentially leading to improved classification accuracy. As expected, with the increase in training samples, the classification accuracy of many methods demonstrated improvement. However, it is noteworthy that our proposed method consistently outperformed several other algorithms, showcasing its robustness and effectiveness even in the face of increased complexity. Our method achieved an outstanding classification accuracy of 99.23%, which speaks to its strong ability to handle the intricate spectral characteristics and spatial patterns present in the Houston 2013 dataset. In some methods such as SpectralFormer and Vit, some of the information in "Tennis Court" was misclassified, and most was incorrectly assigned to "Synthetic Grass", and LSSCM misclassified some of the information in "Stressed Grass" assigned to "Healthy Grass", our proposed method achieved a good classification of those information, further underlining its ability to discern subtle spectral differences and spatial patterns crucial for accurate classification. Figure 9 provides a visual representation of the classification results obtained by the various algorithms. The superiority of our proposed method is evident in the clear and well-defined class boundaries, showcasing its ability to effectively leverage both spectral and spatial information to achieve highly accurate classifications.

5 | CONCLUSIONS

In this paper, we propose a method SSF to improve the performance of HSI classification, which inherits the excellent local feature extraction capability of ConvNet and the advantages of Vit in handling long-range dependencies and global context understanding. Unlike the traditional hybrid structure, our structure first extracts the shallow features of HSI by a variant CNN structure Depthwise separable Convolution, this initial step effectively captures essential local patterns and spatial information present in the data. And followed by 3 TSTs for deep feature extraction of spectral features and spatial features of the data, thereby enhancing the representation of rich spectral characteristics and spatial relationships in the data. The model performs better on all three tasks than other ConvNet, Vit, and hybrid models. In future work, we hope to explore a new Token mixer and channel mixer combined into a MetaFormer more suitable for HSI classification.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China(2022YFE0138600),the National Natural Science Foundation of China(62266045) and the Key projects of Shaanxi Yan'an(2021JB-04).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

REFERENCES

1. Lu G, Fei B. Medical hyperspectral imaging: a review. *Journal of biomedical optics.* 2014;19(1):010901–010901.
2. Briottet X, Boucher Y, Dimmeler A, et al. Military applications of hyperspectral imagery. In: . 6239. SPIE. 2006:82–89.
3. Lu B, Dao PD, Liu J, He Y, Shang J. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sensing.* 2020;12(16):2659.
4. Khan MJ, Khan HS, Yousaf A, Khurshid K, Abbas A. Modern trends in hyperspectral image analysis: A review. *Ieee Access.* 2018;6:14118–14129.
5. Melgani F, Bruzzone L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing.* 2004;42(8):1778–1790.
6. Ye Q, Huang P, Zhang Z, Zheng Y, Fu L, Yang W. Multiview learning with robust double-sided twin SVM. *IEEE Transactions on Cybernetics.* 2021;52(12):12745–12758.
7. Ye Q, Zhao H, Li Z, et al. L1-Norm distance minimization-based fast robust twin support vector *k*-plane clustering. *IEEE transactions on neural networks and learning systems.* 2017;29(9):4494–4503.
8. Chen YN, Thaipisutikul T, Han CC, Liu TJ, Fan KC. Feature line embedding based on support vector machine for hyperspectral image classification. *Remote Sensing.* 2021;13(1):130.
9. Prasad S, Bruce LM. Limitations of principal components analysis for hyperspectral target recognition. *IEEE Geoscience and Remote Sensing Letters.* 2008;5(4):625–629.
10. Fauvel M, Benediktsson JA, Chanussot J, Sveinsson JR. Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing.* 2008;46(11):3804–3814.
11. Zhang H, Li Y. Spectral-spatial classification of hyperspectral imagery based on deep convolutional network. In: IEEE. 2016:44–47.
12. Hu W, Huang Y, Wei L, Zhang F, Li H. Deep convolutional neural networks for hyperspectral image classification. *Journal of Sensors.* 2015;2015:1–12.
13. Tian C, Zhang Y, Zuo W, Lin CW, Zhang D, Yuan Y. A heterogeneous group CNN for image super-resolution. *IEEE transactions on neural networks and learning systems.* 2022.

14. Aptoula E, Ozdemir MC, Yanikoglu B. Deep learning with attribute profiles for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters.* 2016;13(12):1970–1974.
15. Li Y, Zhang H, Shen Q. Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sensing.* 2017;9(1):67.
16. Chen Y, Jiang H, Li C, Jia X, Ghamisi P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing.* 2016;54(10):6232–6251.
17. Zhang H, Li Y, Jiang Y, Wang P, Shen Q, Shen C. Hyperspectral classification based on lightweight 3-D-CNN with transfer learning. *IEEE Transactions on Geoscience and Remote Sensing.* 2019;57(8):5813–5828.
18. Xie F, Gao Q, Jin C, Zhao F. Hyperspectral image classification based on superpixel pooling convolutional neural network with transfer learning. *Remote sensing.* 2021;13(5):930.
19. Zhang H, Gong C, Bai Y, Bai Z, Li Y. 3-D-ANAS: 3-D asymmetric neural architecture search for fast hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing.* 2021;60:1–19.
20. Xue X, Zhang H, Bai Z, Li Y. 3D-ANAS v2: Grafting Transformer Module on Automatically Designed ConvNet for Hyperspectral Image Classification. *arXiv e-prints.* 2021:arXiv–2110.
21. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.* 2020.
22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in neural information processing systems.* 2017;30.
23. Hong D, Han Z, Yao J, et al. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing.* 2021;60:1–15.
24. Sun L, Zhao G, Zheng Y, Wu Z. Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing.* 2022;60:1–14.
25. He X, Chen Y, Lin Z. Spatial-spectral transformer for hyperspectral image classification. *Remote Sensing.* 2021;13(3):498.
26. Mei S, Ji J, Bi Q, Hou J, Du Q, Li W. Integrating spectral and spatial information into deep convolutional neural networks for hyperspectral classification. In: IEEE. 2016:5067–5070.
27. Zhang Q, Xiao J, Tian C, Chun-Wei Lin J, Zhang S. A robust deformed convolutional neural network (CNN) for image denoising. *CAAI Transactions on Intelligence Technology.* 2022.
28. Tian C, Xu Y, Zuo W, Zhang B, Fei L, Lin CW. Coarse-to-fine CNN for image super-resolution. *IEEE Transactions on Multimedia.* 2020;23:1489–1502.
29. Yang J, Zhao Y, Chan JCW, Yi C. Hyperspectral image classification using two-channel deep convolutional neural network. In: IEEE. 2016:5079–5082.
30. Zhang H, Li Y, Zhang Y, Shen Q. Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network. *Remote sensing letters.* 2017;8(5):438–447.
31. Meng Z, Jiao L, Liang M, Zhao F. A lightweight spectral-spatial convolution module for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters.* 2021;19:1–5.
32. Wang W, Xie E, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: 2021:568–578.
33. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021:10012–10022.
34. Graham B, El-Nouby A, Touvron H, et al. Levit: a vision transformer in convnet's clothing for faster inference. In: 2021:12259–12269.
35. Srinivas A, Lin TY, Parmar N, Shlens J, Abbeel P, Vaswani A. Bottleneck transformers for visual recognition. In: 2021:16519–16529.
36. Xiao T, Singh M, Mintun E, Darrell T, Dollár P, Girshick R. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems.* 2021;34:30392–30400.
37. d'Ascoli S, Touvron H, Leavitt ML, Morcos AS, Biroli G, Sagun L. Convit: Improving vision transformers with soft convolutional inductive biases. In: PMLR. 2021:2286–2296.
38. Guo J, Han K, Wu H, et al. Cmt: Convolutional neural networks meet vision transformers. In: 2022:12175–12185.
39. Dai Z, Liu H, Le QV, Tan M. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems.* 2021;34:3965–3977.
40. Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178.* 2021.
41. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted residuals and linear bottlenecks. In: 2018:4510–4520.
42. Zhang H, Hu W, Wang X. Parc-net: Position aware circular convolution with merits from convnets and transformer. In: Springer. 2022:613–630.
43. Wang Y, Jia S, Zhang Z. Multiscale Convolutional Transformer with Center Mask Pretraining for Hyperspectral Image Classificationtion. *arXiv preprint arXiv:2203.04771.* 2022.