Title Page

- **Data availability statement**

The Indian Pines data that support the findings of this study are openly available at http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes. The Pavia data that support the findings of this study are openly available at http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes#Pavia_Universit. The Salinas data that support the findings of this study are openly available at http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes. The Houston2013 data that support the findings of this study are openly available at http://www.grss-ieee.org/community/technical-committees/data-fusion/2013-ieee-grss-data-fusion-contest/.

- **Funding statement**

This work was supported in part by the National Natural Science Foundation of China (41701479 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 135509136.

- **Conflict of interest disclosure**

Not applicable

- **Ethics approval statement**

Not applicable

- **Patient consent statement**

Not applicable

- **Permission to reproduce material from other sources**

Not applicable

- **Clinical trial registration**

Not applicable

A Complementary Integrated Transformer Network for Hyperspectral Image Classification

Diling Liao[1], Cuiping Shi [1,*], Liguo Wang [2]

1 College of Communication and Electronic Engineering, Qiqihar University, Qiqihar 161000, China; 2020910228@qqhru.edu.cn; shicuiping@qqhru.edu.cn.
2 College of Information and Communication Engineering, Dalian Nationalities University, Dalian 116000, China; wangliguo@hrbeu.edu.cn

* Correspondence: shicuiping@qqhru.edu.cn

Abstract—In the past, convolutional neural network (CNN) has become one of the most popular deep learning frameworks, and has been widely used in Hyperspectral image (HSI) classification tasks. Convolution (Conv) in CNN uses filter weights to extract features in local receiving domain, and the weight parameters are shared globally, which more focus on the high-frequency information of the image. Different from Conv, Transformer can obtain the long-term dependence between long-distance features through modeling, and adaptively focus on different regions. In addition, Transformer is considered as a low-pass filter, which more focuses on the low-frequency information of the image. Considering the complementary characteristics of Conv and Transformer, the two modes can be integrated for full feature extraction. In addition, the most important image features correspond to the discrimination region, while the secondary image features represent important but easily ignored regions, which are also conducive to the classification of HSIs. In this paper, a complementary integrated Transformer network (CITNet) for hyperspectral image classification is proposed. Firstly, three-dimensional convolution (Conv3D) and two-dimensional convolution (Conv2D) are utilized to extract the shallow semantic information of the image. In order to enhance the secondary features, a channel Gaussian modulation attention module (CGMAM) is proposed, which is embedded between Conv3D and Conv2D. This module can not only enhance secondary features, but suppress the most important and least important features. Then, considering the different and complementary characteristics of Conv and Transformer, a complementary integrated Transformer module (CITM) is designed. Finally, through a large number of experiments, this paper evaluates the classification performance of CITNet and several state-of-the-art networks on five common datasets. The experimental results show that compared with these classification networks, CITNet can provide better classification performance.

Index Terms—Convolutional neural network, Gaussian modulation, Transformer, Complementary integrated Transformer module.

I. INTRODUCTION

Hyperspectral images (HSIs) were captured by hyperspectral sensors and contain hundreds of narrow-band spectral bands. At present, HSIs is widely used in many fields, such as geological exploration [1], object detection [2], atmospheric environment monitoring [3], [4], and precision agriculture [5], [6]. The task of HSI classification is to identify the land cover categories corresponding to the pixels in the image [7]-[9]. However, since the acquisition of HSIs by sensors

is often affected by the atmosphere, shooting angle, and shooting instruments [10], [11], it is difficult to identify the land cover category corresponding to pixels accurately.

In recent years, many works have made great achievements in the field of computer vision by using deep learning (DL), including image classification [12]-[14], target detection [15], [16], semantic segmentation [17], and have been widely used in the field of HSI classification [18]. Popular backbone networks in DL include auto-encoders (AEs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), capsule networks (CapsNet), and graph convolutional networks (GCNs). In [19], in order to extract high-level features of images, a new hybrid framework based on principal component analysis (PCA) [20], DL architecture, and logistic regression (LR) [21] was proposed. At present, GAN-based methods mainly focus on spectral-spatial GANs [22] and semi-supervised GANs [23]. In [24], Lin et al. proposed an improved GAN method, and better classification results are obtained. Due to the limited sample size of HSI data, a semi-supervised adaptive neighborhood strategy based generative adversarial network (AN-GAN) was proposed [25], which effectively improves the performance of the classifier. In addition, Li et al. [26] proposed a GAN to automatically extract the change regions between optical and SAR remote sensing images, proving that a better GAN model can improve the generated images. Paoletti et al. [27] proposed a new spectral-spatial capsule network (SSCN), and effectively reduced the computational complexity. In addition, Hong et al. [28] proposed a mini-batch GCN (miniGCN), which provides a feasible solution to the large graphs problem in GCNs. Considering the spectral sequence characteristics, RNN [29] can accumulate and learn image spectral features orderly due to its natural sequence data design attributes, but RNN model cannot be calculated in parallel.

Among the above popular backbone networks, CNN is the most popular learning framework because of its powerful image feature extraction ability [30]-[32]. HSI contains rich spectral and spatial information. There is no doubt that the full extraction of spectral and spatial information contained in HSIs can effectively improve the classification results. In the early research work based on CNN, many excellent networks appeared in HSI classification. In [33], Makantasis et al. constructed CNN network, which can automatically extract the spectral and spatial information of images. In order to make better use of spatial information, Cao et al. [34] learned the spatial information of images by constructing CNN and updated CNN parameters using random gradient. In [35], Paoletti et al. proposed a deep pyramid residual network (PyResNet) to extract more spatial information. For extract more nonlinear, discriminant, and invariant features, Chen et al. [36] constructed a multilayer 2D convolution network (2DCNN). However, these methods based on 2-D CNN mostly extract the spatial features of the image, and also have a lot of computational complexity. In addition, for extract image spectral-spatial features and alleviate the problem of computational parameter explosion, Lee and Kwon [37] constructed a new end-to-end CNN by using multiple convolution kernels of different sizes, and extracted rich spectral-spatial features. Similarly, considering that the HSI of 2-D image and 1-D spectral information is very different from that of 3-D target image, a multi-scale 3-D convolutional neural network (3DCNN) was proposed [38]. Although, 3-D CNN has been proved to be able to effectively extract the spectral and spatial features of HSI, and effectively improve the classification performance [39], [40]. However, with the deepening of network, the error gradient will greatly update the network parameters, resulting in network instability or gradient disappearance [41]. For solve these problems, Zhong et al. [42] introduced the residual structure into the spectral module and spatial

module, and proposed a spectral-spatial residual network (SSRN). Roy et al. [43] proposed a hybrid spectral convolution neural network (Hybrid-SN), which uses 3-D CNNs and 2-D CNNs to acquire the spectral-spatial and spatial information of HSI. Although the method based on CNN shows a strong ability to extract spatial information and local context information, it is undeniable that the method based on CNN still has some limitations. On the one hand, it is difficult for CNNs to capture sequence attributes well, especially medium-term and long-term dependencies. When some image land cover categories are complex, it inevitably encounters the performance bottleneck in the task of HSI classification [44]. On the other hand, CNN uses convolution filter weights for feature extraction in local receiving domain, and the weight parameters are shared globally, resulting in CNN paying too much attention to spatial content information and ignoring important spectral features.

In the past two years, Transformer-based methods show great potential in computer vision tasks [45-48]. Among them, the most classic model is vision Transformer (ViT) [49], which performs well in the field of image processing. In [50], a spectral-spatial Transformer (SST) was proposed. First, SST use Visual Geometry Group Network (VGGNet) [51] to extract spatial features and constructs a dense Transformer to obtain long-term dependencies. In order to solve the problem of spectral redundancy of HSI, a self-attention-based Transformer network (SATNet) [52] was proposed. In addition, Hong et al. [44] proposed a spectral Transformer (SF), reconsidered the Transformer from the perspective of spectral sequence, and learned the group adjacent spectral information by constructing a cross layer Transformer encoder module. Although these Transformer-based methods can effectively learn spectral information of HSIs, they ignore local semantic information, resulting in the lack of spatial information acquisition. To solve this problem, Le et al. [53] proposed a spectral-spatial feature tokenization Transformer (SSFTT) based on spectral-spatial feature. The network uses 3-D CNN and 2-D CNN extract shallow layer features, and designs a Gaussian weighted feature marker for feature transformation. Similarly, in [54], a spectral-spatial Transformer network (SSTN) was proposed, and used a factorized architecture search (FAS) framework to determine the hierarchical operation selection and block-level order of SSTN.

Conv uses filter weights to extract image features in local receiving domain, and the weight parameters are shared globally, which makes the extracted features more focus on the high-frequency information of the image [55]. On the contrary, Transformer is considered as a low-pass filter [56], which more focuses on the low-frequency information of the image. Considering the complementary characteristics of Conv and Transformer, integrating these two modes is beneficial to the full extraction of features. SSFTT [53] and SSTN [54] also fully verify this view. However, these methods only use the simple combination of Conv and Transformer, and the classification performance is not very satisfactory. In order to better integrate the two modes, this paper proposes a complementary integrated Transformer network (CITNet). Firstly, CITNet uses Conv3D and Conv2D to extract the spectral and spatial features of HSIs. Secondly, considering the importance of secondary features, a channel Gaussian modulation attention module (CGMAM) is designed, which is embedded between Conv3D and Conv2D to enhance the secondary features extracted by Conv3D. Following, for make full use of the advantages of Conv and Transformer, a complementary integrated Transformer module (CITM) is designed, which embeds Conv in Transformer. Finally, a linear classifier based on softmax is utilized for classification.

The main contributions of this paper are as follows:

1) A CITM module is designed in this paper, which fully considers the advantages of Conv and Transformer, embeds Conv in Transformer, and effectively fuses the obtained low-frequency information and high-frequency information.
2) Considering that the features extracted by Conv contain secondary features, it is also helpful to improve the classification performance. Therefore, a CGMAM module is proposed in this paper, which is used to enhance the secondary features extracted by Conv.
3) The proposed CITNet method in this paper systematically integrates CNN and Transformer. This method can extract high-frequency and low-frequency information of HSIs more effectively, and can significantly improve the classification performance. Experiments on five common datasets show the effectiveness of the CITNet.

The rest of this paper is organized as follows. Section II introduces all the modules of the proposed network in detail. In Section III, the model is analyzed, and the quantitative and visual results of the experiment are given. Finally, this paper is concluded in Section IV.

II. METHODOLOGY

The overall structure of the proposed CITNet is shown in Fig. 1, which is mainly composed of three parts, including feature extraction base on Conv, CGMAM, and CITM.

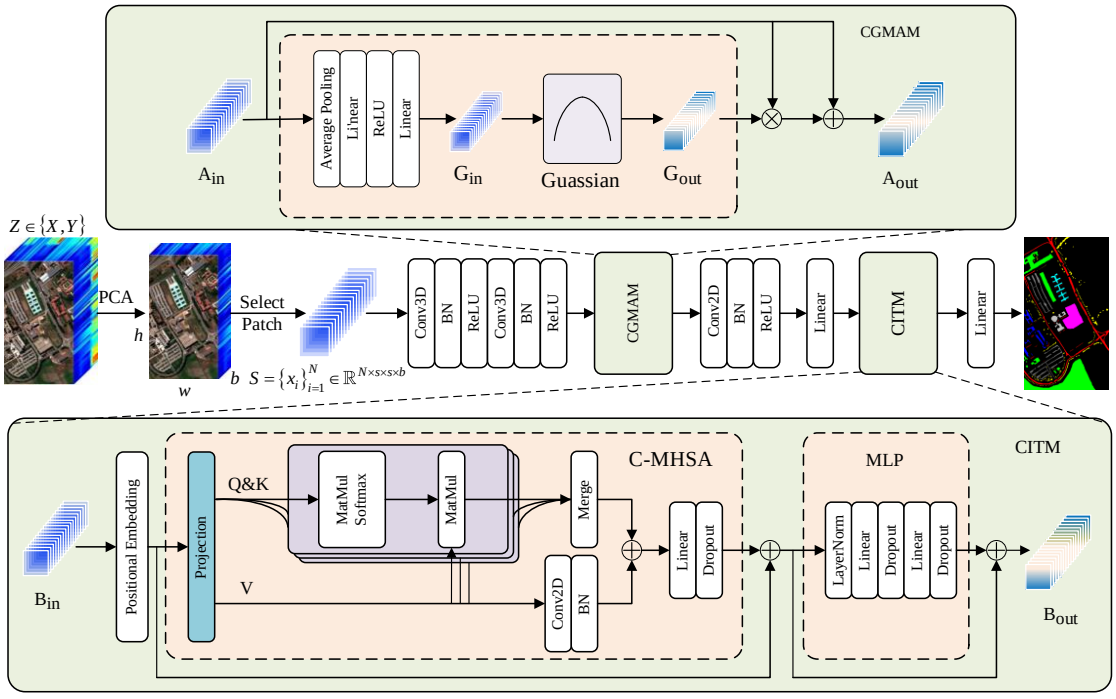

Fig. 1 The overall structure of the proposed CITNet.

A. Feature extraction based on Conv

In the field of HSI classification, CNNs have shown strong feature extraction ability. HSIs contain rich spectral and spatial information. Three-dimensional convolution (Conv3D) and two-dimensional convolution (Conv2D) are adopted for feature extraction of HSIs, which can not only acquire the spectral-spatial joint features of the image, but acquire the spatial features of the

image solely. Therefore, the CITNet first adopts Conv3D and Conv2D for feature extraction.

The original HSI is $Z \in \{X, Y\}$, $X \in \mathbb{R}^{h \times w \times l}$ is HSI data, and $Y = \{y_1, y_2, ..., y_C\}$ is HSI label. Where, $h \times w$ is the space size of the image, l is the number of bands of image, and C is the largest number of category labels. Although HSI carry a lot of useful spectral information, there are still many redundant spectral. Therefore, PCA method is used to preprocess the original HSI data for reduce the computational complexity. After PCA preprocessing, the number of bands changes from l to b, and the output is $X_{pca} \in \mathbb{R}^{h \times w \times b}$. Then, 3D cube extraction is performed on X_{pca}. N adjacent 3D cubes $x \in \mathbb{R}^{s \times s \times b}$ are generated by X_{pca}, and $s \times s$ is the space size. In particular, the central pixel of all x is (x_i, x_j), the label of each x is determined by the label of the central pixel, and all x has their corresponding labels. Where $0 \le i < h$, $0 \le j < w$. Then, in addition to the background data, the remaining data samples are divided into training dataset and test dataset.

Then, the spectral-spatial feature of $x \in \mathbb{R}^{s \times s \times b}$ is extracted through Conv3D, which can be expressed as

$$v_{i,j}^{x,y,z} = f\left(\sum_{d} \sum_{\alpha=0}^{H_i=1} \sum_{\beta=0}^{W_i=1} \sum_{\gamma=0}^{R_i=1} K_{i,j,d}^{\alpha,\beta,\gamma} \cdot v_{(i-1),d}^{(x+\alpha),(y+\beta),(z+\gamma)} + b_{i,j} \right) \tag{1}$$

Where, $f(\cdot)$ is the activation function; $v_{i,j}^{x,y,z}$ represents the neuron at position (x, y, z) of the j-th feature map of layer i-th. H_i, W_i and R_i represent the height, width and depth of the 3-D convolution kernel of layer i-th respectively. $K_{i,j,d}^{\alpha,\beta,\gamma}$ is the weight parameter of the d-th feature cube at position (α, β, γ). $b_{i,j}$ is the bias term. In the proposed CITNet, Conv3D module (including Conv3D layer, BN layer, and ReLU layer) are used for feature extraction. The output feature size is $\mathbb{R}^{s \times s \times c}$ and c is the number of channels.

In order to fully extract features, Conv2D is adopted to further extract the spatial features, which can be expressed as

$$v_{i,j}^{x,y} = f\left(\sum_{d} \sum_{\alpha=0}^{H_i^{'}-1} \sum_{\beta=0}^{W_i^{'}-1} K_{i,j,d}^{\alpha,\beta} v_{(i-1),d}^{(x+\alpha),(y+\beta)} + b_{i,j} \right) \tag{2}$$

Where $H_i^{'}$ and $W_i^{'}$ is the height and width of 2-D convolution kernel respectively. $K_{i,j,d}^{\alpha,\beta}$ represents the weight parameter of the d-th feature map at position $(H_i^{'}, W_i^{'})$.

The feature extraction of the data is carried out through a Conv2D module (including Conv2D layer, BN layer, and ReLU layer), and the output feature size is still $\mathbb{R}^{s \times s \times b}$.

B. Channel Gaussian modulation attention module

It was demonstrated in [57] that the most important feature corresponds to the discrimination region, while the secondary feature represents the important but easily ignored region. The most important features are essential to improve the discriminative ability, while the secondary features

are also conducive to better classification. Therefore, in order to enhance the secondary features, this paper proposed a channel Gaussian modulation attention module (CGMAM), which is utilized to enhance the secondary channel features. The input is $A_{in} \in \mathbb{R}^{s \times s \times c}$ (c represents the number of channels). Firstly, the input A_{in} is average pooling, linear, and activation layer to obtain the output feature G_{in} including channel dependence. Then, G_{in} redistributes the distribution of features through Gaussian modulation function, and enhance the secondary features of the channel to obtain the output feature map G_{out}. Finally, G_{out} performs channel weighting with the original input A_{in}. However, the output obtained at this time retains only secondary features. Therefore, the weighted output and the original input A_{in} are added pixel by pixel to obtain the output A_{out}, which contain the enhanced secondary features and the original important features. The above operations can be expressed as

$$A_{out} = \mathcal{G}\left(H\left(P_s\left(A_{in}\right)\right)\right) \otimes A_{in} + A_{in} \tag{3}$$

Where $P_s(\cdot)$ represents the average pooling function, $H(\cdot)$ represents the linear and activation function layer, $\mathcal{G}(\cdot)$ represents the Gaussian modulation function and $\otimes$ represents the channel by channel weighting.

In particular, in CGMAM, Gaussian modulation function is used to redistribute the distribution of features.

$$G_{out} = \mathcal{G}\left(G_{in}\right) \tag{4}$$

Input G_{in} can map all feature values to Gaussian distribution through Gaussian modulation function. The mean μ and variance σ of Gaussian distribution can be represented as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} G_{in}^{i} \tag{5}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(G_{in}^{i} - \mu\right)^2} \tag{6}$$

In order to better explain the Gaussian modulation function, we visualize the distribution of feature values before and after Gaussian modulation. As can be seen from Fig. 2, after Gaussian modulation, the large feature values and small feature values are suppressed, while the middle feature values are enhanced.

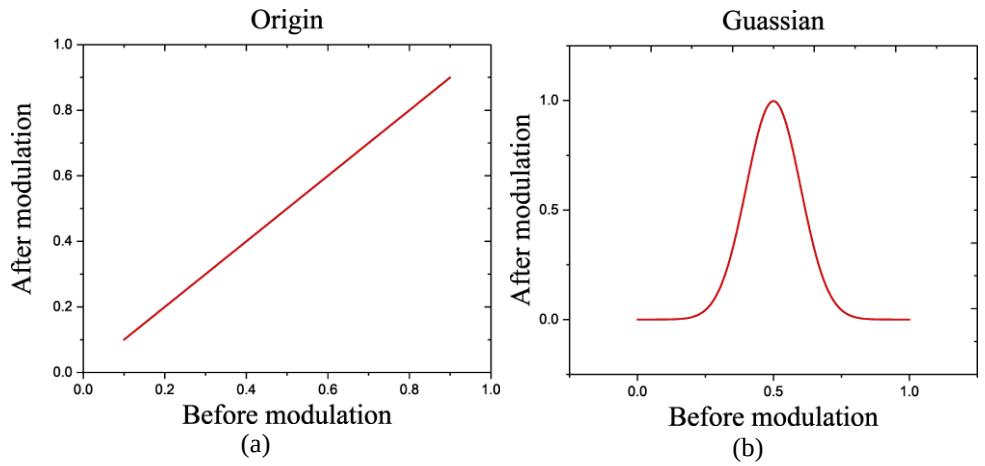

Fig. 2 Distribution before and after Gaussian modulation. (a) Original distribution of feature values, (b) The distribution after enhancing secondary features with Guassian modulation.

C. Complementary Integrated Transformer Module

In recent years, Transformer has been widely used in natural language processing (NLP). In [45], ViT is a Transformer-based classical network applied to image classification tasks and achieved satisfactory classification performance. Unlike CNN, Transformer can obtain the long-term dependence between remote features and deep semantic features through modeling. Considering the complementary characteristics of Conv and Transformer, the combination of these two modes is beneficial to the full extraction of features. Therefore, this paper designed a complementary integrated Transformer module (CITM). An illustration of CITM is shown in Fig.3. The structure of CITM is shown in Fig. 3 (a). CITM mainly consists of position embedding and complementary multi-head self-attention (C-MHSA) module. Among them, C-MHSA is the key part of CITM, which is shown in Fig. 3 (b).

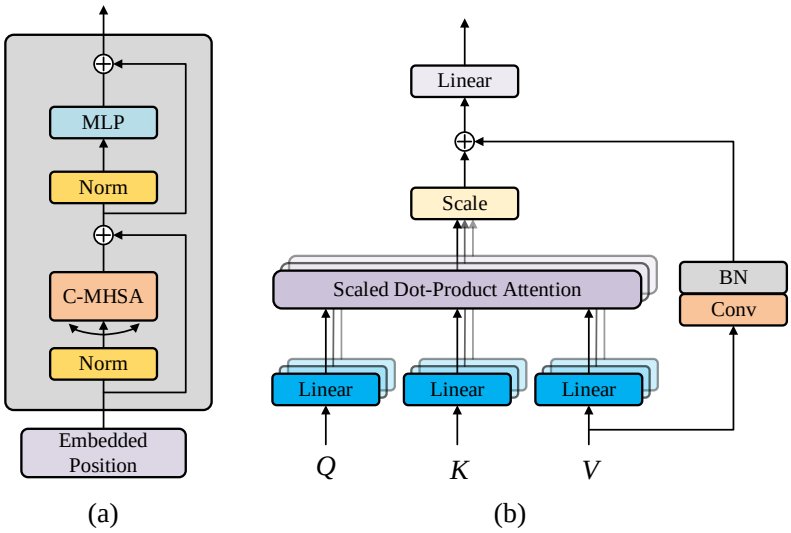

Fig. 3 The specific architecture description of CITM. (a) The architecture of CITM, (b) The architecture of C-MHSA (The key part of CITM).

Before implementing the CITM, in order to facilitate data processing, we reshape

$x \in \mathbb{R}^{c\times s\times s}$ into $x \in \mathbb{R}^{c\times ss}$, and obtain $B_{in} \in \mathbb{R}^{c\times z}$ through linear mapping.

As shown in Fig. 1, the output $B_{in} \in \mathbb{R}^{c\times z}$ after linear mapping is used as the input of position embedding. Then, the location information PE is encoded and added to the tokens represented by $[T_0^{cls}, T, ..., T_z]$. The resulting can be represented as

$$T_{in} = [T_0^{cls}, T_1, ..., T_z] + PE \tag{7}$$

Where T_0^{cls} represent classification token. Transformer can obtain deeper semantic features through modeling. It includes multi-head self-attention (MHSA), two LayerNorm (LN) and one MLP layer. Among them, Transformer can achieve excellent performance thanks to multi-head self-attention (MHSA). Generally, the input of MSHA includes Query (Q), Key (K), and Value (V). However, MHSA multiplication is considered to be a low-pass filter [52]. MHSA establishes different projection information in multiple different projection spaces, and more focuses on the low-frequency information of the image. On the contrary, Conv uses a filter to extract information in local receiving domain, and more focuses on the high-frequency information of the image. Considering the complementary characteristics of Conv and Transformer, a C-MHSA is proposed in this paper. Its structure is shown in Fig. 3 (b). The process of C-MHSA is

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

$$CMHSA = Concat(SA_1, SA_2, ..., SA_h)W + Conv(V) \tag{9}$$

Where d_K represents the dimension of K, h represents the number of headers, $W \in \mathbb{R}^{h\times d_K\times d_z}$ is the weight parameter, $Conv(\cdot)$ is the convolution function (including Conv and BN layers), and $Concat(\cdot)$ is the cascade function. Similar to MHSA, C-MHSA first linearly maps into three invariant matrices Q, K, and V, and uses the softmax function to calculate the score. Then, multiply the obtained result with V to obtain the self-attention (SA). Finally, the results of each head SA are connected together and fused with the V after convolution to obtain the C-MHSA output result. This integration method can effectively utilized the advantages of Transformer and Conv, and fully extract the high-frequency and low-frequency information of hyperspectral images. Finally, the output from C-MHSA is input to LN and MLP layers. The implementation process of CITM is summarized in Algorithm 1.

Algorithm 1: The implementation process of CITM

Input: Input $B_{in} \in \mathbb{R}^{c\times z}$

Output: The output of CITM is $B_{out} \in \mathbb{R}^{c\times z}$.

1: **for** i=1 to T **do**

2: Perform position embedding, the result denoted as T_{in}.

3: Perform the feature mapping layer and obtains three different output matrices Q, K and V.

4: V obtains the output $Conv(V)$ through convolution and BN layer.

4: Q and K perform inner product operation, and the softmax function is performed on the result $Soft\max\left(QK^T/\sqrt{d_K}\right)$.

5: The SA output $SA(Q,K,V)$ is generated by multiplying $Soft\max\left(QK^T/\sqrt{d_K}\right)$ and V.

6: The results of each head $SA(Q,K,V)$ are connected together and feature fusion is carried out to generate C-MHSA output.

7: Perform residual mapping layer.

8: Perform LN layer.

9: Perform MLP layer.

10: Perform residual mapping layer.

end for

D. Implementation

For better illustrate the proposed CITNet, this paper takes Indian Pines dataset as an example for detailed description. The dataset has a size of 145×145×200. Firstly, the output of the origin data after PCA preprocessing and 3D cube extraction is 13×13×30. In the first Conv3D, eight 7×7×7 convolution kernels are convoluted to obtain eight 13×13×30 features. Then the obtained features are passed through 64 1×1×30 convolution kernels are convoluted to acquire 64 feature with a size of 13×13×1 and reshaped it into 64 feature map with the size of 13×13. Then, 64 feature map with the size of 13×13 passes through CGMAM, and the output size is the same as the input. Finally, the output is passed through Conv2D, and 64 convolution kernel with the size of 7×7. Finally, the spatial dimension is flattened as a vector, and the output is $x \in \mathbb{R}^{64\times 169}$.

Next, in order to facilitate data processing, the $x \in \mathbb{R}^{64\times 169}$ obtained above is linearly mapped to obtain the feature $x \in \mathbb{R}^{64\times 64}$. Then, an all zero vector is connected to x as a learnable marker, and a learned position marker is embedded to obtain $T_{in} \in \mathbb{R}^{65\times 64}$. After the CITM module, the feature size remains unchanged. The process of HSI classification by the proposed CITNet is shown in Algorithm 2.

Algorithm 2: HSI classification based on the proposed CITNet

Input: HSI data $X \in \mathbb{R}^{h\times w\times l}$, real label $Y \in \mathbb{R}^{h\times w}$, PCA band number is 30, the size of cube space is 13×13, and the proportion of training samples is p%.

Output: The prediction label of the test dataset.

1: Set the batch size to 64, Adam optimizer (learning rate 0.005), and epoch T to 200.

2: The output of PCA is $X_{pca} \in \mathbb{R}^{h\times w\times b}$.

3: N adjacent 3D cubes $x \in \mathbb{R}^{s\times s\times b}$ are all generated by X_{pca}, and all x are proportionally p% randomly divided into training datasets, and the remaining data samples are test datasets.

4: **for** i=1 to T **do**

5: Perform Conv3D.

6: Reshape the output of Conv3D to $\mathbb{R}^{s\times s\times c}$ and perform CGMAM.

7: Perform Conv2D.

8: Reshape the output of Conv2D and after Linear mapping, the output is $B_{in} \in \mathbb{R}^{c\times z}$.

9: Perform CITM.

10: Use the softmax function to identify the label.

end for

11: Use test dataset with the trained model to get predicted labels

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Dataset description

Five HSI datasets have been considered in our experiments in this paper, including Indian Pines dataset, Pavia dataset, Salinas dataset, Houston 2013 dataset, and WHU-Hi-LongKou dataset. The category name and data sample division of all datasets are listed in Table I.

Indian Pines dataset: The dataset contains 145×145 pixels and 224 spectral bands. In addition to the water absorption band and low signal-to-noise ratio (SNR) band, there are still 200 bands to be used. In addition, the dataset contains 16 land cover categories, mainly including crops and plants.

Pavia dataset: The HSI was acquired by the sensor of reflective optics spectral image system (ROSIS-3). It has 115 spectral bands, and the space size of the image is 610×340, including 9 land cover categories. After removing 12 low SNR and noise bands, the remaining 103 spectral bands were used.

Salinas dataset: The HSI contains 512×217, and contains 224 spectral bands. After removing the discarded noise bands (the number of 108-112, 154-167, and 224), there are 204 spectral bands left. The Salinas dataset contains 16 land cover categories with a spatial resolution of 3.7m.

Houston 2013 dataset: The HSI was captured by the hyperspectral image analysis team and NCALM using sensors on the campus of the University of Houston and nearby urban areas. The space size is 349×1905, and contains 144 spectral bands, and its wavelength range is 380 ~ 1050 nm. Houston 2013 has 15 land cover categories.

WHU-Hi-LongKou dataset: The WHU-Hi-LongKou dataset was acquired on July 17, 2018, in Longkou Town, Hubei province, China, by an 8-mm focal length Headwall Nano-Hyperspec imaging sensor. The spatial size of the image is 550×400 and has 270 spectral bands. In addition, the image is about the agricultural scene, including 9 categories.

Table I Category name and data sample division quantity of all datasets.

Indian Pines				Pavia			
Class	Class name	Train	Test	Class	Class name	Train	Test
1	Alfalfa	4	42	1	Asphalt	331	6300
2	Corn-notill	142	1286	2	Meadows	932	17717
3	Corn-mintill	82	748	3	Gravel	104	1995
4	Corn	23	214	4	Trees	153	2911
5	Grass-pasture	48	435	5	Painted metal	67	1278
6	Grass-trees	72	658	6	Bare Soil	251	4778
7	Grass-pasture-mowed	3	25	7	Bitumen	66	1264
8	Hay-windrowed	47	431	8	Self-blocking	184	3498
9	Oats	3	17	9	Shadows	47	900
10	Soybean-notill	97	875	/	/	/	/
11	Soybean-mintill	245	2210	/	/	/	/
12	Soybean-clean	59	534	/	/	/	/
13	Wheat	20	185	/	/	/	/
14	Woods	126	1139	/	/	/	/
15	Bldg-Grass-Tree-Drivers	38	348	/	/	/	/
16	Stone-Steel-Towers	9	84	/	/	/	/
/	Total	1018	9231	/	Total	2135	40641
Salinas				Houston 2013			
Class	Class name	Train	Test	Class	Class name	Train	Test

1	Brocoil-green-weeds_1	100	1909	1	Healthy grass	125	1126
2	Brocoil-green-weeds_2	186	3540	2	Stressed grass	125	1129
3	Fallow	98	1878	3	Synthetic grass	69	628
4	Fallow-rough-plow	69	1325	4	Trees	124	1120
5	Fallow-smooth	133	2545	5	Soil	124	1118
6	Stubble	197	3762	6	Water	32	293
7	Celery	178	3401	7	Residential	126	1142
8	Grapes-untrained	563	10708	8	Commercial	124	1120
9	Soil-vinyard-develop	310	5893	9	Road	125	1127
10	Corn-senesced-green-weeds	163	3115	10	Highway	122	1105
11	Lettuce-romaine-4wk	53	1015	11	Railway	123	1112
12	Lettuce-romaine-5wk	96	1831	12	Parking Lot 1	123	1110
13	Lettuce-romaine-6wk	45	871	13	Parking Lot 2	46	423
14	Lettuce-romaine-7wk	53	1017	14	Tennis Court	42	386
15	Vinyard-untrained	363	6905	15	Running Track	65	595
16	Vinyard-vertical-trellis	90	1717	/	/	/	/
/	Total	2697	51432	/	Total	1495	13534
WHU-Hi-LongKou				/			
Class	Class name	Train	Test	/	/	/	/
1	Corn	172	34339	/	/	/	/
2	Cotton	41	8333	/	/	/	/
3	Sesame	15	3016	/	/	/	/
4	Broad-leaf soybean	316	62896	/	/	/	/
5	Narrow-leaf soybean	20	4131	/	/	/	/
6	Rice	59	11795	/	/	/	/
7	Water	335	66721	/	/	/	/
8	Roads and houses	35	7089	/	/	/	/
9	Mixed weed	26	5203	/	/	/	/
/	Total	1019	203523	/	/	/	/

B. Experimental setup

1) Evaluation Indicators

For HSI classification, there are three commonly used evaluation indicators: Overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) [61]. Let $H = \left(a_{i,j}\right)_{n\times n}$ be the confusion matrix between the real category information and the predicted category information. Where n is the number of categories, $a_{i,j}$ A is the quantity of category j classified as category i . Then, the OA value is

$$\mathrm{OA} = \frac{\sum_{i=1}^{n} a_{i,i}}{M} \times 100\% \tag{10}$$

Where, M is the total number of samples, and OA refers to the proportion of samples accurately classified in all samples. Another evaluation indicator AA refers to the classification accuracy of each category.

$$\mathrm{AA} = \frac{1}{n}\sum_{i=1}^{n} \frac{a_{i,i}}{\sum_{j=1}^{n} a_{i,j}} \times 100\% \tag{11}$$

Finally, the specific calculation of κ is as follows

$$\kappa = \frac{\sum_{i=1}^{n} a_{i,i} - \frac{\sum_{i=1}^{n}\left(a_{i,\_}\, a_{\_,i}\right)}{M}}{M - \frac{\sum_{i=1}^{n}\left(a_{i,\_}\, a_{\_,i}\right)}{M}} \tag{12}$$

In the above formula, $a_{i,\_}$ and $a_{\_,i}$ respectively represent all column elements corresponding to row i and all row elements corresponding to column i in the confusion matrix H.

2) Comparison methods

For comparison, some state-of-the-art classification networks are chosen, including 2DCNN [32], 3DCNN [34], PyResNet [31], Hybrid-SN [39], SSRN [38], ViT [45], SF [40], SSFTT[49], and SSTN[50].

2DCNN is composed of two convolution layers and two pool layers.

3DCNN is composed of two multi-scale 3-D convolution modules and a full connection layer. Each multi-scale convolution module contains four convolution kernels with a size of 1×1×1, 1×1×3, 1×1×5, and 1×1×11.

SSRN consists of spatial residual module and spectral residual module. Among them, the spatial residual module contains five convolution blocks, and each convolution block is composed of 3-D convolution layer and BN layer. The spectral residual module also contains five convolution blocks.

PyResNet consists of five different modules, namely C, P1, P2, P3, and output module. Among them, C contains a convolution layer and BN layer, while pyramid modules P1, P2 and P3 are composed of three pyramid residual units. Finally, the output module is classified by down sampling and a full connection layer.

Hybrid-SN is a hybrid CNN network. The spectral-spatial information of the image are extracted by 3-D CNN, and 3-D CNN contains three 3-D convolution layers. The spatial features of the image are extracted by 2-D CNN, and 2-D CNN contains a 2-D convolution layer.

ViT is a classical method based on Transformer. ViT includes a linear mapping component and Transformer encoder.

SF rethinks the classification of HSIs from the perspective of spectral sequence, and proposes a Transformer-based backbone network to replace the architecture based on CNN or RNN.

SSFTT is a spectral-spatial feature tokenization Transformer network.

SSTN is a spectral-spatial Transformer network, and a FAS framework is used to determine the hierarchical operation selection and block-level order of SSTN.

3) Implementation details

All experiments in this paper are implemented on the Pytorch software platform, and the hardware platform of the experiment is a desktop PC with Intel (R) Core (TM) i9-9900K CPU and a NVIDIA GeForce RTX 2080Ti GPU. It is worth noting that we use the batch size, learning rate and epoch to 64, 5e-3 and 200, respectively.

For fair comparison, all experiments in this paper were carried out in the above experimental environment, and all results were taken as the average of 20 experiments.

C. Model analysis

1) Ablation Experiments

Ablation research of the proposed CITM: Transformer can obtain deeper semantic features through modeling, and more focus on the low-frequency information of the image. On the contrary, Conv uses a filter to extract information in local receiving domain, which more focuses to the high-frequency information of the image. Considering the complementary characteristics of the two, a CITM module is proposed in this paper. The difference between this CITM module and the original Transformer module is that Conv is introduced into the multi-head self-attention part. In order to verify the effectiveness of the designed CITM module, this paper takes Indian Pines dataset as an example to study ablation. The results are shown in Table II, we can see that after the introduction of Conv, OA, AA, and κ have been greatly increased. In order to further verify the influence of convolution kernel size on the performance of CITM module, the convolution kernel with size of 3×3, 5×5, and 7×7 are adopted for experiments. It can be seen that the classification performance decreases with the increase of convolution kernel size. Therefore, the ablation research results fully prove that the introduction of Conv into Transformer can effectively improve the classification accuracy. With the gradual increase of the size of Conv kernel, the classification accuracy decreases gradually. This is because as the convolution kernel size increases, the fine features that can be obtained become less and less.

Table II Ablation research of the CITM module on Indian Pines dataset. (Optimal results are bolded)

With or without Conv		OA (%)	AA (%)	$\kappa\times100$
Without Conv		96.36	91.00	95.84
With Conv	3×3	**98.71**	**98.13**	**98.53**
	5×5	98.20	97.28	97.94
	7×7	97.93	96.67	97.64

Ablation research of the proposed CITNet: CITNet is mainly composed of three components, including Conv3d & Conv2d, CGMAM, and CITM. Conv3D & Conv2D are used to extract the spectral and spatial features of HSI. In order to enhance the secondary features, a CGMAM is proposed, which is embedded between Conv3D and Conv2D. In addition, this paper also designs a CITM module, which combines Transformer and Conv. To verify the effectiveness of these three components, some ablation experiments are performed on the Indian Pines dataset. The results of ablation experiment are shown in Table III. In the first case, the network only contains Conv3D & Conv2D, and the final classification accuracy is the worst. In the second case, the network includes Conv3D & Conv2D and CGMAM, and the classification accuracy is improved. In the third case, the network includes Conv3D & Conv2D and CITM, and the classification accuracy is better than the first two cases. Compared with the first case, the OA, AA, and κ of the third case increased by 6.89%, 9.08%, and 7.87%. In the fourth case, when the network contains these three components, the classification accuracy is the best. Therefore, the ablation research fully verified the effectiveness of the main components in CITNet.

Table III Ablation research of the CITNet on Indian Pines dataset. (Optimal results are bolded)

Cases	Components			Metric		
	Conv3D&Conv2D	CGMAM	CITM	OA (%)	AA (%)	$\kappa\times100$
1	√	×	×	91.34	84.70	90.12
2	√	√	×	93.88	88.46	93.02
3	√	×	√	98.23	93.78	97.99
4	√	√	√	**98.71**	**98.13**	**98.53**

2) Parameter sensitivity analysis

In the deep learning network, many parameters have an impact on the network performance. Among them, the learning rate and batch size directly affect the training process of the model. In other words, the learning rate directly affects the convergence state of the network, and the batch size affects the generalization performance of the network, and these two parameters will also affect each other. In order to explore the suitable learning rate and batch size of the proposed CITNet network, we conducted a combined experiment of different learning rate and batch size on five datasets. Among them, the selected learning rate set is $\{1e-4, 5e-4, 1e-3, 5e-3\}$, and the selected batch size set is $\{128, 64, 32, 16\}$. The experimental results are shown in Fig. 4.

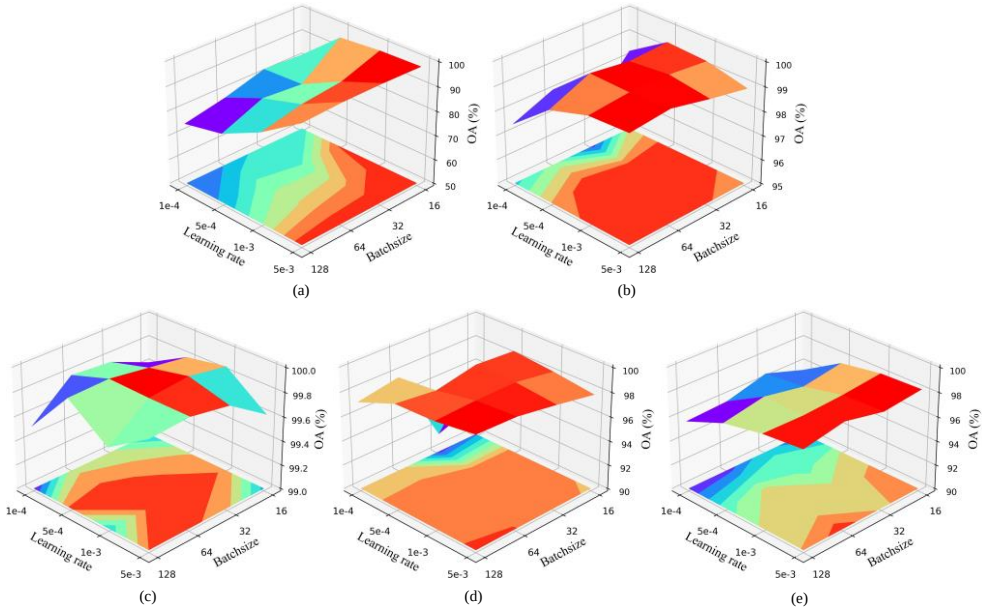

Fig. 4 Experimental results of different learning rates and batch sizes on all datasets. (a) Indian Pines, (b) Pavia, (c) Salinas, (d) Houston 2013, (e) WHU-Hi-LongKou.

In Fig. 4, red represents the maximum value area of contour and dark blue represents the minimum value area of contour. It can be seen from the Fig. 4 that in Indian Pines dataset, the OA value increases with the increase of learning rate, while different batch sizes have little impact on the OA value, as shown in Fig. 4 (a). In Pavia dataset, larger learning rate and batch size can often obtain larger OA value, as shown in Fig. 4 (b). In Salinas dataset, the optimal batch size is 64, and the corresponding different learning rates can obtain higher OA values, as shown in Fig. 4 (c). On the Houston 2013 dataset, it is obvious that the optimal learning rate is $5e-3$, and the corresponding optimal batch sizes are 128 and 64, as shown in Fig.4 (d). Similarly, on the WHU-Hi-LongKou dataset, the optimal learning rate and batch size are $5e-3$ and 64, respectively, as shown in Fig. 4 (e). To sum up, we select $5e-3$ and 64 as the learning rate and batch size of CITNet network.

3) Different input space sizes

For hyperspectral image classification, different input space sizes also have an impact on classification performance. In order to explore the optimal input space size of the five datasets on the proposed network, some experimental have been conducted. The size of the input space

selected in the experiment is 7×7，9×9，11×11，13×13，15×15，17×17，19×19, and 21×21. The results are shown in Fig. 5. As can be seen from Figure 5, the results show that the OA value of Indian Pines, Pavia, and WHU-Hi-LongKou datasets first increases and then decreases with the increase of input space. The OA values of Salinas and Houston 2013 datasets gradually increase with the increase of input space size, and tend to be flat after obtaining the highest OA value. Among them, when the space size of Indian Pines, Pavia, and WHU-Hi-LongKou is 13×13, the maximum OA value was obtained. When the input space size of Salinas and Houston 2013 is 14×14 and 19×19 respectively, the maximum values of OA are obtained, but they are not much different from OA value obtained by input space with the size of 13×13. Considering that the larger the input space will inevitably bring a large number of parameters, the size of the input space adopted by CITNet on the five datasets is 13×13.

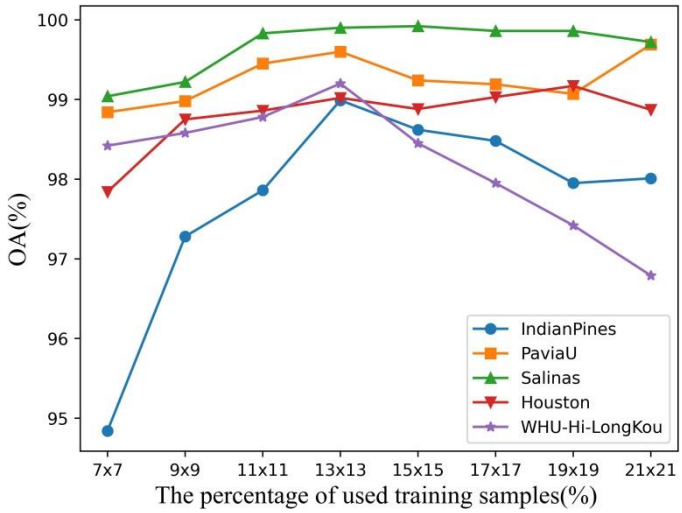

Fig. 5 Effect of different input space size on classification accuracy OA.

D. Result analysis

1) Quantitative analysis

Tables IV-VIII give the classification results of OA, AA, κ, and per-class on the five datasets. Through rough observation, it can be easily found that both CNN-based and Transformer-based methods have achieved satisfactory classification accuracy. In particular, compared with other methods, our method has the highest OA on all datasets. Specifically, among the CNN-based methods, the extracted features are insufficient due to the shallow network of 2DCNN and 3DCNN. It is not surprising that these two methods obtain the worst classification accuracy. Hybrid-SN combines 3-D CNN and 2-D CNN, which considers both spectral-spatial features and spatial features, and finally obtains high classification accuracy. Deep CNN network can obtain features with stronger discrimination ability, but it often brings the problem of gradient disappearance or gradient explosion. In order to solve these problems, SSRN introduces the residual structure into the spectral module and spatial module, and obtains the optimal classification accuracy based on CNN method. In addition, among the methods based on Transformer, ViT, as a classical model, shows the great potential of Transformer in HSI classification. Inspired by ViT, SF obtains gratifying classification results by fully considering the

spectral sequence. It is worth noting that the latest two Transformer-based SSFTT and SSTN have achieved better classification accuracy than most CNN-based networks on five datasets.

Finally, it can be found that among all CNNs and Transformers, our method obtains the highest classification accuracy. Compared with the SSRN with the highest classification accuracy in CNNs, the OA value of the proposed method is 0.17%, 0.84%, 0.05%, 0.04%, and 0.53% higher on the five datasets respectively. Compared with SSFTT with the highest classification accuracy in Transformers, the OA value of the proposed method is 1.28%, 0.48%, 0.49%, 0.67%, and 0.21% higher on the five datasets respectively. It is worth noting that for some categories that are difficult to classify, such as categories 15 and 16 of Indian Pines and category 13 of Houston 2013, the proposed method achieves 100% classification accuracy. This also fully shows that the proposed method effectively improves the ability of feature discrimination by integrated CNN and Transformer.

Table IV Classification accuracy of all methods on Indian Pines dataset. (Optimal results are bolded)

Methods	CNNs					Transformers				
	2DCNN[32]	3DCNN[34]	PyResNet[31]	Hybrid-SN[39]	SSRN[38]	ViT[45]	SF[40]	SSFTT[49]	SSTN[50]	Proposed
OA (%)	82.04	81.15	92.01	94.31	98.54	79.73	88.54	97.43	95.43	**98.71**
AA (%)	89.09	87.15	89.67	94.32	97.92	83.36	91.81	93.85	84.54	**98.13**
κ ×100	79.26	78.26	90.91	93.51	98.33	76.75	86.88	97.07	94.78	**98.53**
1	**100**	**100**	96.03	96.91	96.43	99.00	**100**	85.71	39.41	88.10
2	76.01	72.54	92.62	90.87	**99.38**	72.83	82.89	98.13	96.91	98.68
3	77.24	69.29	94.48	92.05	98.50	69.39	84.39	96.66	95.55	**99.20**
4	96.07	95.83	93.13	91.32	**100**	82.71	91.44	96.73	96.46	98.13
5	93.85	95.23	94.25	**98.73**	97.48	85.30	93.50	98.39	94.99	96.32
6	79.85	81.47	92.40	97.87	98.48	85.22	91.74	99.09	96.98	**99.70**
7	80.00	60.00	61.11	91.32	**100**	92.86	**100**	**100**	31.12	**100**
8	97.04	95.89	98.44	98.34	98.72	92.06	95.34	98.38	96.71	**99.54**
9	**100**	**100**	69.44	95.54	**100**	79.50	**100**	58.82	40.00	**100**
10	83.60	87.14	92.51	91.61	96.80	76.73	87.01	**99.31**	88.37	96.11
11	75.16	75.66	93.43	95.23	99.13	77.39	88.19	98.10	96.29	**99.86**
12	82.73	77.70	92.70	92.74	**96.43**	68.13	79.56	89.70	92.87	94.57
13	**100**	99.86	94.16	99.52	**100**	93.56	95.37	99.46	98.79	**100**
14	92.99	93.23	99.35	96.79	99.80	92.06	93.92	98.68	98.79	**99.91**
15	90.86	91.10	77.43	94.74	97.74	82.41	91.14	93.97	94.17	**100**
16	**100**	99	93.31	85.57	87.80	84.62	94.33	90.47	95.24	**100**

Table V Classification accuracy of all methods on Pavia dataset. (Optimal results are bolded)

Methods	CNNs					Transformers				
	2DCNN[32]	3DCNN[34]	PyResNet[31]	Hybrid-SN[39]	SSRN[38]	ViT[45]	SF[40]	SSFTT[49]	SSTN[50]	Proposed
OA (%)	94.55	93.69	94.70	97.99	98.79	94.35	95.89	99.15	97.20	**99.63**
AA (%)	93.55	93.38	93.96	97.49	97.81	92.15	93.64	98.62	96.75	**99.33**
κ ×100	92.74	91.56	92.94	97.33	98.40	92.48	94.55	98.87	96.27	**99.51**
1	91.07	88.64	89.74	97.19	**99.81**	90.74	93.23	99.67	95.89	99.79
2	97.18	96.27	97.99	99.22	**100**	97.57	98.96	99.99	98.83	99.97
3	83.47	83.55	88.97	97.66	81.52	82.37	82.55	**98.59**	92.36	97.04
4	99.31	98.99	99.41	99.08	**99.96**	99.14	99.83	93.71	97.21	99.00
5	99.84	99.44	99.13	98.18	99.76	96.87	99.73	99.84	96.13	**100**
6	95.81	95.83	96.09	98.68	99.98	94.43	97.79	99.54	99.95	**100**
7	89.88	91.04	87.77	97.36	**100**	79.37	80.51	99.53	99.08	**100**
8	86.76	87.87	88.93	92.16	**99.28**	89.59	90.31	98.00	92.89	98.91
9	98.66	98.76	97.61	97.87	**100**	99.30	99.88	98.67	98.41	99.22

Table VI Classification accuracy of all methods on Salinas dataset. (Optimal results are bolded)

Methods	CNNs					Transformers				
	2DCNN[32]	3DCNN[34]	PyResNet[31]	Hybrid-SN[39]	SSRN[38]	ViT[45]	SF[40]	SSFTT[49]	SSTN[50]	Proposed
OA (%)	96.01	96.62	98.22	98.99	99.85	97.87	97.72	99.41	94.03	**99.90**
AA (%)	98.02	98.22	98.97	99.29	**99.84**	99.43	98.85	99.37	98.08	99.83
κ ×100	95.55	96.24	98.02	98.88	99.83	97.55	97.46	99.34	93.40	**99.89**
1	99.81	99.72	99.89	99.86	**100**	97.87	99.40	99.95	99.37	**100**
2	99.62	99.31	**100**	99.98	**100**	99.43	99.97	99.92	99.86	**100**

3	99.39	98.51	99.98	99.82	**100**	97.55	98.75	99.89	98.52	**100**
4	99.51	99.42	97.31	99.63	99.56	98.98	99.79	**99.85**	99.03	99.47
5	99.10	99.02	99.40	99.45	99.49	98.68	99.13	98.66	98.85	**99.96**
6	99.95	99.96	99.94	99.87	99.99	99.92	99.93	99.79	99.91	**100**
7	99.62	99.76	**100**	99.81	99.96	99.84	99.96	99.97	99.92	**100**
8	90.84	92.18	97.40	98.15	99.88	92.45	94.86	98.55	94.63	**99.95**
9	99.73	99.93	99.99	99.88	**100**	99.39	99.93	99.88	99.45	99.86
10	95.84	96.64	99.39	99.29	99.78	98.16	98.25	99.13	99.64	**99.90**
11	97.98	98.69	98.52	99.05	99.63	94.14	99.11	**100**	98.80	99.61
12	99.58	99.59	99.55	99.87	99.95	99.08	99.75	**100**	99.91	**100**
13	98.95	99.06	99.51	99.07	99.93	98.55	**100**	99.20	99.42	99.31
14	99.17	99.18	99.82	98.02	99.79	97.35	99.69	95.67	99.60	**100**
15	89.75	90.77	92.94	97.15	99.55	89.02	93.05	99.93	83.00	**99.94**
16	99.54	99.85	99.90	99.73	**100**	99.50	99.96	99.48	**100**	99.30

Table VII Classification accuracy of all methods on Houston 2013 dataset. (Optimal results are bolded)

Methods	CNNs					Transformers				
	2DCNN[32]	3DCNN[34]	PyResNet[31]	Hybrid-SN[39]	SSRN[38]	ViT[45]	SF[40]	SSFTT[49]	SSTN[50]	Proposed
OA (%)	92.63	93.01	95.85	97.83	98.98	92.28	93.83	98.35	92.82	**99.02**
AA (%)	93.27	93.52	96.26	98.19	**99.01**	93.57	94.17	97.92	94.13	98.89
κ ×100	92.03	92.43	95.51	97.65	98.90	91.65	93.32	98.22	92.23	**98.95**
1	97.50	98.89	94.85	99.18	**99.59**	90.15	96.99	97.42	91.46	99.47
2	96.72	95.41	98.69	98.50	99.90	93.90	94.88	99.47	88.33	**99.91**
3	98.75	98.65	92.58	**100**	**100**	98.94	98.30	98.09	99.54	**100**
4	96.26	97.94	96.25	97.93	98.51	98.74	97.29	99.20	95.69	99.26
5	97.52	97.70	97.14	99.50	98.24	98.41	98.71	99.73	99.64	99.91
6	95.10	95.84	98.41	**99.74**	98.89	97.12	93.77	97.95	98.70	99.32
7	88.56	89.08	94.78	96.65	**100**	93.54	93.23	97.29	96.14	96.85
8	86.26	89.41	96.99	99.45	**99.90**	88.52	90.72	99.20	97.92	97.50
9	86.80	86.20	91.81	95.94	**99.48**	90.54	89.73	96.45	90.62	98.49
10	94.41	91.44	96.10	94.95	98.40	83.88	91.66	99.55	83.26	**99.91**
11	87.97	88.65	96.57	95.91	98.80	92.15	92.75	99.55	89.36	**100**
12	91.47	90.97	96.17	96.89	96.65	86.59	88.25	99.46	91.92	**99.82**
13	91.44	91.16	95.60	**98.59**	97.67	95.14	91.64	86.52	92.96	92.91
14	95.27	93.46	99.63	**100**	**100**	97.66	97.04	99.22	99.01	**100**
15	97.95	97.91	98.33	99.63	99.07	98.26	97.53	99.66	97.45	**100**

Table VIII Classification accuracy of all methods on WHU-Hi-LongKou dataset. (Optimal results are bolded)

Methods	CNNs					Transformers				
	2DCNN[32]	3DCNN[34]	PyResNet[31]	Hybrid-SN[39]	SSRN[38]	ViT[45]	SF[40]	SSFTT[49]	SSTN[50]	Proposed
OA (%)	89.95	95.12	95.24	98.60	98.67	86.48	92.43	98.99	97.20	**99.20**
AA (%)	80.61	89.72	91.25	96.73	**98.48**	73.71	82.60	97.81	94.87	97.88
κ ×100	86.56	93.57	93.91	98.16	98.25	82.07	90.05	98.68	96.31	**98.95**
1	96.03	99.75	99.09	99.20	99.97	83.68	96.11	98.24	**99.91**	99.75
2	59.83	66.72	81.42	96.53	99.70	39.75	72.39	**99.86**	82.17	99.12
3	96.67	98.92	73.08	94.00	**100**	59.38	83.48	99.07	**100**	96.50
4	85.20	95.10	99.07	98.63	97.43	87.93	93.71	**99.60**	97.24	99.23
5	56.31	76.30	89.13	95.13	99.80	35.03	67.89	97.57	**100**	97.30
6	91.99	97.81	98.71	98.74	**99.97**	95.80	94.80	99.45	99.93	99.48
7	98.86	98.80	99.58	99.69	99.54	99.25	97.6	99.82	99.77	**99.91**
8	67.55	82.63	85.58	**95.29**	91.65	92.09	80.60	90.93	77.33	94.57
9	73.03	91.40	95.62	93.36	98.23	70.50	57.07	95.79	**97.50**	95.09

2) Visual Evaluation

The classification maps of the comparison method and the proposed method on five datasets are shown in Fig. 6-Fig. 10. Through visual comparison, it can be known that the classification map of the proposed method CITNet is closest to the ground truth map on all datasets. It is obvious that due to the strong local context feature extraction ability of CNN, some CNN-based methods have obtained relatively smooth classification maps, including hybrid-SN and SSRN. This is also due to their use of 3-D CNN and 2-D CNN to extract the spectral and spatial information of HSI. The worst classification is 2DCNN which only considers spatial information.

Spectral feature is an important feature of HSI classification, and Transformer can get the long-term dependence between long-distance features through modeling, and adaptively pay attention to different regions and pay attention to the low-frequency information of more images. We can further find that the Transformer-based method cannot well classify some small-size isolated objects due to considering more low-frequency information, such as the red "Healthy grass" category and the bright green "Stressed grass" category in the Houston 2013 dataset. It is worth noting that although the classification results acquired by the Transformer-based methods ViT and SF still have many misclassified categories, it also fully shows the potential of the Transformer-based method. Therefore, our method integrates CNN and Transformer, which can not only fully extract the high-frequency features of local context, but also retain more low-frequency features of the image. Through the visual comparison, it is not difficult to verify the effectiveness of the proposed method.

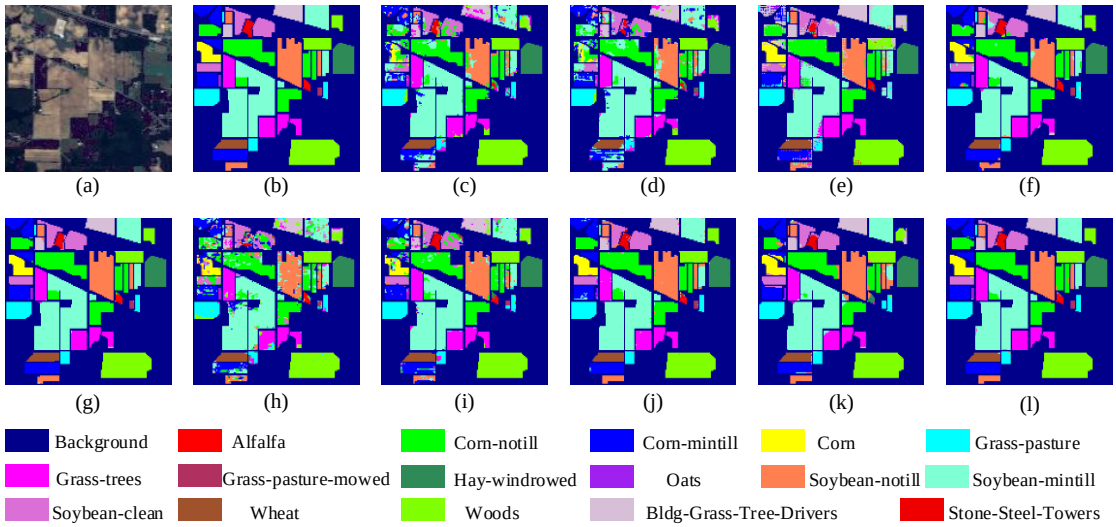

Fig. 6 Classification maps of all methods on Indian Pines dataset. (a) False color composite, (b) Ground truth map, (c)-(l) are classification maps of 2DCNN (82.04%), 3DCNN (81.15%), PyResNet (92.01%), Hybrid-SN (94.31%), SSRN (98.54%), ViT (79.73%), SF (88.54%), SSFTT (97.43%), SSTN (95.43%), and Proposed (98.71%).

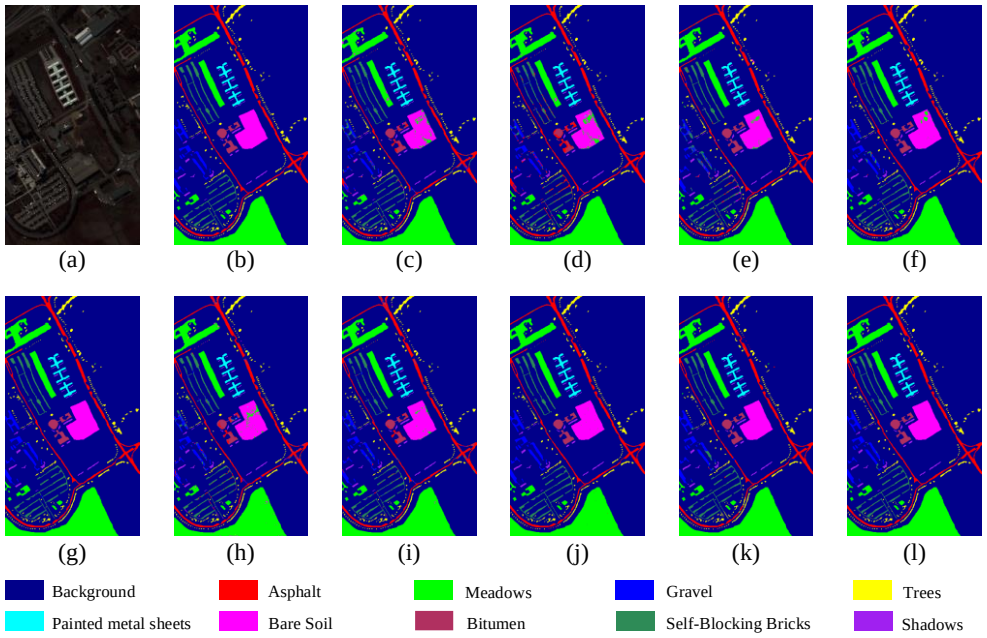

Fig. 7 Classification maps of all methods on Pavia dataset. (a) False color composite, (b) Ground truth map, (c)-(l) are classification maps of 2DCNN (94.55%), 3DCNN (93.69%), PyResNet (94.70%), Hybrid-SN (97.99%), SSRN (%), ViT (94.35%), SF (95.89%), SSFTT (99.15%), SSTN (97.20%), and Proposed (99.63%).

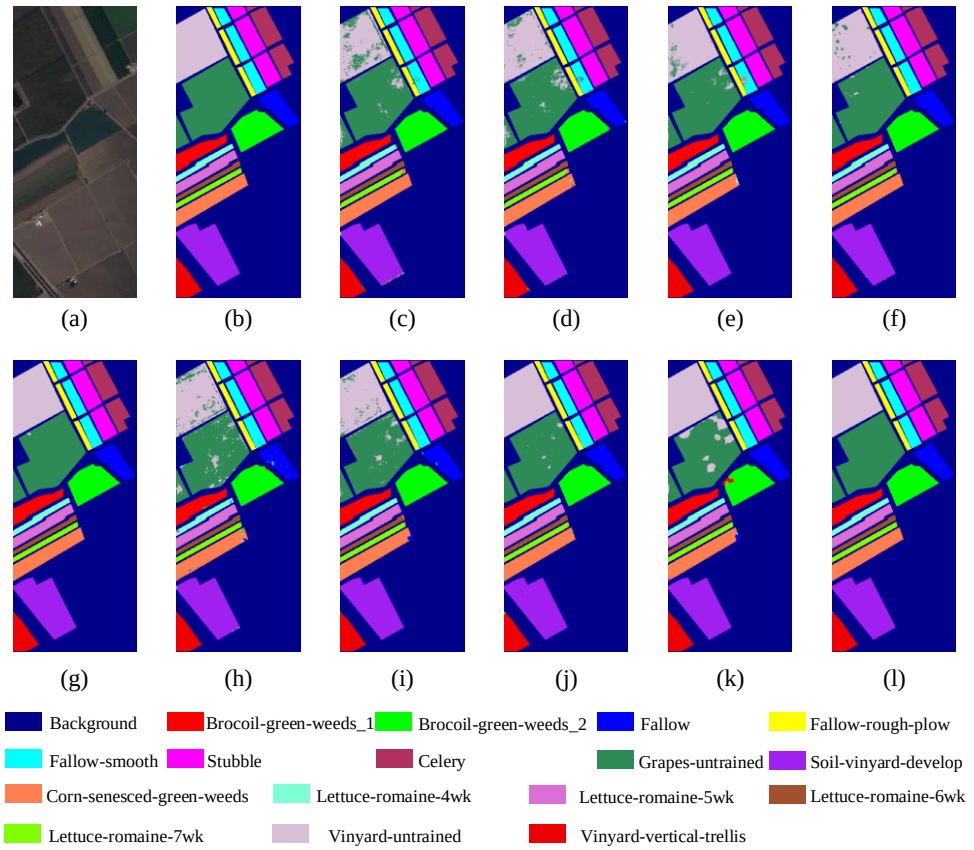

Fig. 8 Classification maps of all methods on Salinas dataset. (a) False color composite, (b) Ground truth map, (c)-(l) are classification maps of 2DCNN (96.01%), 3DCNN (96.62%), PyResNet (98.22%), Hybrid-SN (98.99%), SSRN (99.85%), ViT (97.87%), SF (97.72%), SSFTT (99.41%), SSTN (94.03%), and Proposed (99.90%).

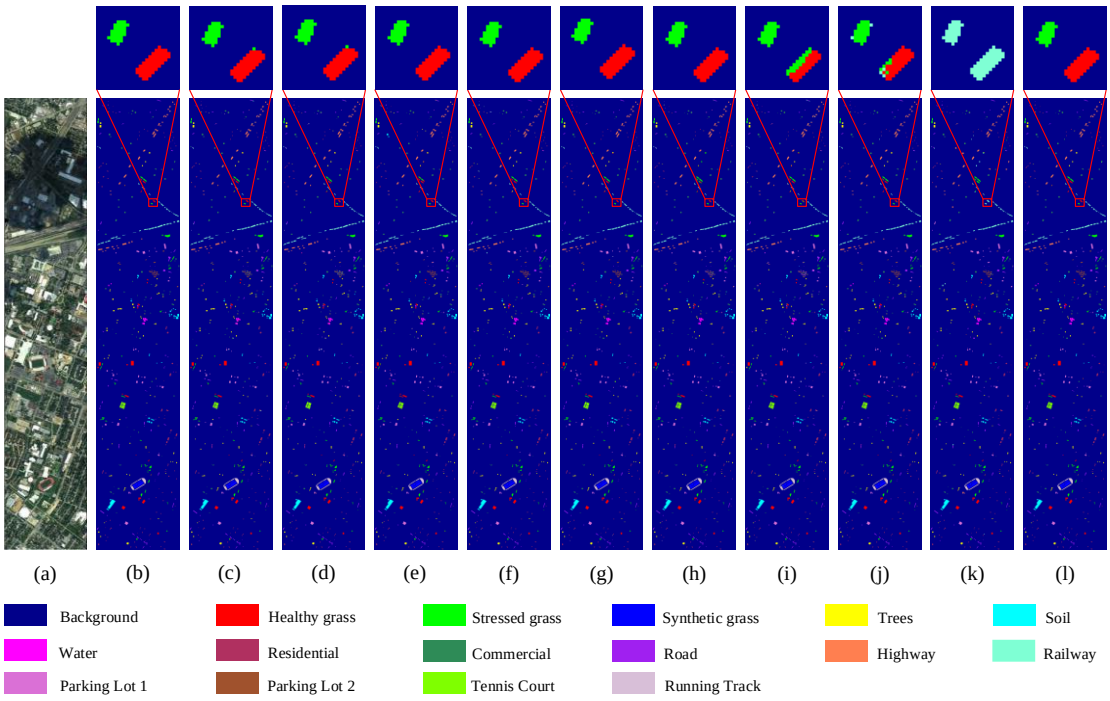

Fig. 9 Classification maps of all methods on Houston 2013 dataset. (a) False color composite, (b) Ground truth map, (c)-(l) are classification maps of 2DCNN (92.63%), 3DCNN (93.01%), PyResNet (95.85%), Hybrid-SN (97.83%), SSRN (98.98%), ViT (92.28%), SF (93.83%), SSFTT (98.35%), SSTN (92.82%), and Proposed (99.02%).

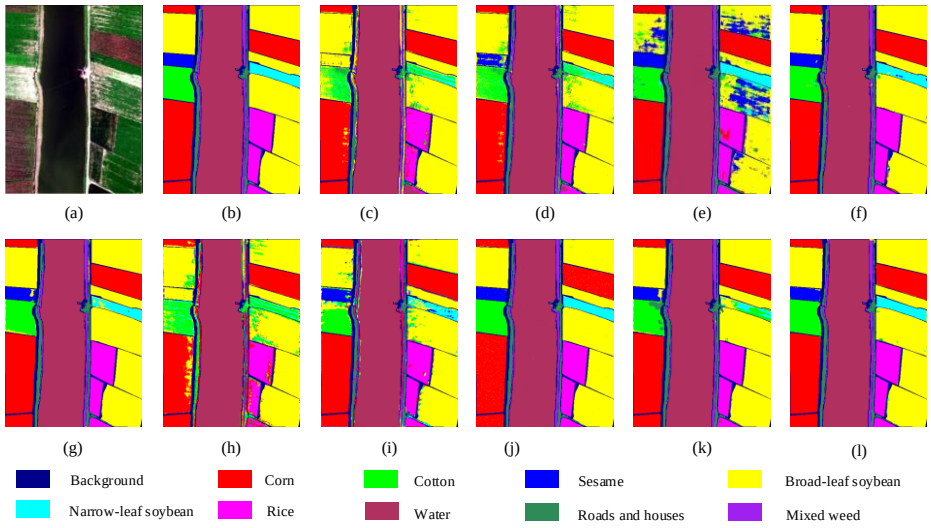

Fig. 10 Classification maps of all methods on WHU-Hi-LongKou dataset. (a) False color composite, (b) Ground truth map, (c)-(l) are classification maps of 2DCNN (89.95%), 3DCNN (95.12%), PyResNet (95.24%), Hybrid-SN (98.60%), SSRN (98.67%), ViT (86.48%), SF (92.43%), SSFTT (98.99%), SSTN (97.20%), and Proposed (99.20%).

Table IX Training time (min) and test time (s) of all methods on four datasets.

Methods		Indian Pines		Pavia		Salinas		Houston 2013	
		Train	Test	Train	Test	Train	Test	Train	Test
CNNs	2DCNN[32]	1.21	6.60	**1.48**	15.40	4.55	38.60	**1.40**	6.80
	3DCNN[34]	**0.89**	6.40	1.19	15.60	2.92	41.80	1.49	7.00
	PyResNet[31]	5.53	18.67	9.31	48.67	13.10	88.67	3.76	12.67

	Hybrid-SN[39]	2.57	6.74	18.28	38.40	34.79	77.40	15.00	9.20
	SSRN[38]	7.79	5.87	34.10	8.40	54.38	18.92	11.56	7.28
Transformers	ViT[45]	2.98	10.95	2.31	7.53	6.43	8.19	1.41	5.65
	SF[40]	1.24	4.89	2.19	13.64	7.29	22.22	1.70	3.26
	SSFTT[49]	1.14	1.06	2.02	3.22	**1.30**	**2.76**	1.52	1.31
	SSTN[50]	0.98	1.34	2.87	10.40	3.54	16.53	1.23	2.04
	Proposed	1.11	**0.71**	2.27	**3.21**	2.73	3.74	1.62	**1.04**

3) Time Cost Comparison

In order to further compare the proposed methods, Table IX gives the training time and test time required for all methods on the four datasets. By comparing the results in Table IX, it can be found that the network training time and test time required by 2DCNN and 3DCNN in CNN method are short, which is related to their shallow network. The training time and testing time of PyResNet, Hybrid-SN, and SSRN based on CNN are longer than those based on Transformer, which is also the advantage of Transformer method. Similarly, we can easily observe that the training time and testing time required by the Transformer-based method are similar, while the testing time required by the proposed method is the shortest on the Indian Pines, Pavia, and Houston 2013 datasets. Although the test time required by the proposed method in Salinas dataset is not the optimal result, it is the suboptimal result. Although deep CNN network can obtain better performance, its computational efficiency is poor. However, Transformer can obtain high-level semantic information without building a deep network, which has high computational efficiency. Therefore, the integration of CNN and Transformer can not only improve the computational efficiency, but also have good classification performance, which fully shows its great potential.

IV CONCLUSIONS

A CITNet network for hyperspectral image classification is proposed. First, CITNet uses Conv3D and Conv2D to extract the shallow layer features of the image. Then, a CGMAM embedded between Conv3D and Conv2D is designed to emphasize the secondary features extracted by Conv3D. Due to own limitations, Conv is not conducive to the establishment of long-term dependence, and is more inclined to the extraction of high-frequency information. On the contrary, Transformer modeling can get the long-term dependence between long-distance features and pay more attention to low-frequency information. Considering the complementary characteristics of Conv and Transformer, a CITM module is proposed, which integrates Conv and Tranformer. In order to verify the effectiveness of the designed network,, some quantitative experiments and visual evaluation have been conducted on five common datasets, and fully verified the effectiveness of CITNet. In the future work, we will aim at integrating the advantages of CNNs and Transformers, and introduce some advanced technologies (including migration learning and self-supervised learning) to further improve the Transformer framework.

FUNDING

This research was funded in part by the National Natural Science Foundation of China (41701479, 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 135509136.

REFERENCES

[1] N. Yokoya, J. C.-W. Chan, and K. Segl, "Potential of resolution enhanced hyperspectral data for mineral mapping using simulated EnMAP and Sentinel-2 images," *Remote Sens.*, vol. 8, no. 3, p. 172, 2016.

[2] Z. Wu, W. Zhu, J. Chanussot, Y. Xu and S. Osher, "Hyperspectral Anomaly Detection via Global and Local Joint Modeling of Background," in *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3858-3869, 15 July15, 2019.

[3] B. Zhao, L. Hua, X. Li, X. Lu, and Z. Wang, "Weather recognition via classification labels and weather-cue maps," *Pattern Recognit.*, vol. 95, pp. 272–284, Nov. 2019.

[4] Z. Wang, T. Yang, and H. Zhang, "Land contained sea area ship detection using spaceborne image," *Pattern Recognit. Lett.*, vol. 130, pp. 125–131, Feb. 2020.

[5] Y. Lanthier, A. Bannari, D. Haboudane, J. R. Miller and N. Tremblay, "Hyperspectral Data Segmentation and Classification in Precision Agriculture: A Multi-Scale Analysis," *IGARSS 2008 - 2008 IEEE International Geoscience and Remote Sensing Symposium*, 2008, pp. II-585-II-588.

[6] F. Xing, H. Yao, Y. Liu, X. Dai, R. L. Brown, and D. Bhatnagar, "Recent developments and applications of hyperspectral imaging for rapid detection of mycotoxins and mycotoxigenic fungi in food products," *Crit. Rev. Food Sci. Nutrition*, vol. 59, no. 1, pp. 173–180, Jan. 2019.

[7] T. V. Bandos, L. Bruzzone and G. Camps-Valls, "Classification of Hyperspectral Images With Regularized Linear Discriminant Analysis," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 3, pp. 862-873, March 2009.

[8] J. Zhang, X. Jin, J. Sun, J. Wang, and A. K. Sangaiah, "Spatial and semantic convolutional features for robust visual object tracking," *Multimed. Tools*. Appl., pp. 1-21, Aug. 2018.

[9] D. Zeng, Y. Dai, F. Li, R. S. Sherratt, and J. Wang, "Adversarial learning for distant supervised relation extraction," *Comput. Mater. Contin.*, vol. 55, no. 1, pp. 121-136, Jan. 2018.

[10] R. A. Borsoi, T. Imbiriba and J. C. M. Bermudez, "A Data Dependent Multiscale Model for Hyperspectral Unmixing With Spectral Variability," in *IEEE Transactions on Image Processing*, vol. 29, pp. 3638-3651, 2020.

[11] L. Drumetz, J. Chanussot, C. Jutten, W. -K. Ma and A. Iwasaki, "Spectral Variability Aware Blind Hyperspectral Image Unmixing Based on Convex Geometry," in *IEEE Transactions on Image Processing*, vol. 29, pp. 4568-4582, 2020.

[12] 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, doi: 10.1109/CVPR.2016.90.

[13] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," *International Conference on Neural Information Processing Systems. Curran Associates Inc*. 2017.

[14] S. Sabour, N. Frosst, and G. E Hinton, "Dynamic routing between capsules," 2017, arXiv:1710.09829. [Online]. Available: http://arxiv.org/abs/1710.09829.

[15] 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*

Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[16] 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.

[17] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 1 April 2017.

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

[19] Y. Chen, Z. Lin, X. Zhao, G. Wang and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094-2107, June 2014.

[20] T. J. Malthus and P. J. Mumby, "Remote sensing of the coastal zone: An overview and priorities for future research," *Int. J. Remote Sens.*, vol. 24, no. 13, pp. 2805–2815, Jan. 2003.

[21] G. M. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 6, pp. 1335-1343, June 2004.

[22] Y. Zhan, D. Hu, Y. Wang, and X. Yu, “Semi-supervised hyperspectral image classification based on generative adversarial networks,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 212–216, Feb. 2018.

[23] J. Feng, H. Yu, L. Wang, X. Cao, X. Zhang, and L. Jiao, “Classification of hyperspectral images based on multiclass spatial–spectral generative adversarial networks,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5329–5343, Aug. 2019.

[24] L. Zhu, Y. Chen, P. Ghamisi and J. A. Benediktsson, "Generative Adversarial Networks for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5046-5063, Sept. 2018.

[25] H. Liang, W. Bao, B. Lei, J. Zhang and K. Qu, "Adaptive Neighborhood Strategy Based Generative Adversarial Network for Hyperspectral Image Classification," *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 862-865.

[26] X. Li, Z. Du, Y. Huang, Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images", *ISPRS Journal of Photogrammetry and Remote Sensing*, Volume 179, 2021.

[27] 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.

[28] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza and J. Chanussot, "Graph Convolutional Networks for Hyperspectral Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966-5978, July 2021.

[29] 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.

[30] H. Zhang, Y. Li, Y. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 5, pp. 438–447, May 2017.

[31] Y. Xu, L. Zhang, B. Du and F. Zhang, "Spectral–Spatial Unified Networks for Hyperspectral

Image Classification," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5893-5909, Oct. 2018,.

[32] H. Zhai, H. Zhang, L. Zhang and P. Li, "Total Variation Regularized Collaborative Representation Clustering With a Locally Adaptive Dictionary for Hyperspectral Imagery," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 166-180, Jan. 2019.

[33] K. Makantasis, K. Karantzalos, A. Doulamis and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015, pp. 4959-4962.

[34] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu and J. Paisley, "Hyperspectral Image Classification With Markov Random Fields and a Convolutional Neural Network," in *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2354-2367, May 2018.

[35] 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.

[36] Y. Chen, H. Jiang, C. Li, X. Jia and P. Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232-6251, Oct. 2016.

[37] H. Lee and H. Kwon, "Going Deeper With Contextual CNN for Hyperspectral Image Classification," in *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4843-4855, Oct. 2017.

[38] M. He, B. Li and H. Chen, "Multi-scale 3D deep convolutional neural network for hyperspectral image classification," *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3904-3908.

[39] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 67, pp. 1–21, 2017.

[40] 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.

[41] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735 - 1780, 1997.

[42] 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.

[43] 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.

[44] 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.

[45] B. Heo, S. Yun, D. Han, S. Chun, J. Choe and S. J. Oh, "Rethinking Spatial Dimensions of Vision Transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11916-11925.

[46] B. Graham et al., "LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12239-12249.

[47] Zhou D, Kang B, Jin X, et al. "DeepViT: Towards Deeper Vision Transformer," *Computer Vision and Pattern Recognition*, 2021.

[48] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *Computer Vision and Pattern Recognition*, 2021.

[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *Computer Vision and Pattern Recognition*, 2020.

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

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[52] Y. Qing, W. Liu, L. Feng, and W. Gao, "Improved transformer net for hyperspectral image classification, " *Remote Sensing*, vol. 13, no. 11, p. 2216, 2021.

[53] 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.

[54] 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.

[55] A. Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Communications of the ACM*, 60:84 – 90, 2012.

[56] Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022.

[57] P. -T. Jiang, Q. Hou, Y. Cao, M. -M. Cheng, Y. Wei and H. Xiong, "Integral Object Mining via Online Attention Accumulation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2070-2079.