MATNet: A Combining Multi-Attention and Transformer Network for Hyperspectral Image Classification

Bo Zhang, Yaxiong Chen, Yi Rong, Shengwu Xiong, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*— Hyperspectral image (HSI) has rich spatial–spectral information, high spectral correlation, and large redundancy between information. Due to the sparse background distribution of HSI, the existing methods generally perform poorly for the classification of class pixels located in the boundary areas of land cover categories. This is largely because the network is vulnerable to surrounding redundant information during the training stage, leading to inaccurate feature extraction and thus poor generalization ability of the model. Based on previous work, we propose an HSI classification network called MATNet which combines multi-attention and transformer. The network first uses spatial attention (SA) and channel attention (CA) to pay more attention to the more significant information parts, then uses the tokenizer module to make a semantic-level representation of different categories of ground objects, and then performs deep semantic feature extraction using the transformer encoder module. Finally, we design a loss function called Lpoly, which adds a polynomial to the label smoothing loss to tune the original first polynomial to accommodate different datasets and tasks. We perform experiments in several well-known HSI datasets as well as for visualization. The results show that our proposed MATNet performs well in extracting spatial–spectral features of HSIs and understanding semantic degrees of semantic degrees.**

***Index Terms*— Deep learning (DL), feature fusion, hyperspectral image (HSI) classification, natural networks, transformer.**

Manuscript received 24 October 2022; revised 17 January 2023; accepted 17 February 2023. Date of publication 9 March 2023; date of current version 24 March 2023. This work was supported in part by the NSFC under Grant 62101393, in part by the Project of Sanya Yazhou Bay Science and Technology City under SCKJ-JYRC-2022-17 and Grant Grant SCKJ-JYRC-2022-76, in part by the National Key Research and Development Program of China under Grant 2022ZD0160604, in part by the NSFC under Grant 62176194, in part by the Major Project of IoV under Grant 2020AAA001, in part by the Postdoctoral Project of Hainan Yazhou Bay Seed Laboratory under Grant B22E18102, in part by the Project of Sanya Science and Education Innovation Park of Wuhan University of Technology under Grant 2022KF0032, in part by the Youth Fund Project of Hainan Natural Science Foundation under Grant 6220N344, in part by the CAAI-Huawei MindSpore Open Fund under Grant CAAIXSJLJJ-2022-001A, in part by the Knowledge Innovation Program of Wuhan-Basic Research, in part by Sanya Science and Education Innovation Park of Wuhan University of Technology under Grant 2021KF0031 and Grant 2022KF0020, in part by the Fundamental Research Funds for the Central Universities under Grant WUT:223110001, in part by the Natural Science Foundation of Chongqing under Grant cstc2021jcyj-msxmX1148, in part by the Open Project of Wuhan University of Technology Chongqing Research Institute under Grant ZL2021-6, and in part by the Hainan Special PhD Scientific Research Foundation of Sanya Yazhou Bay Science and Technology City under Grant HSPHDSRF-2022-03-017. We thank MindSpore for the partial support of this work, which is a new deep learning computing framework(https://www.mindspore.cn/). *(Corresponding author: Shengwu Xiong.)*

Bo Zhang is with the Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China, also with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Yaxiong Chen, Yi Rong, and Shengwu Xiong are with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China, also with the Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China, also with the Hainan Yazhou Bay Seed Laboratory, Sanya 572025, China, also with the Wuhan University of Technology Chongqing Research Institute, Chongqing 401122, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China (e-mail: xiongsw@whut.edu.cn).

Xiaoqiang Lu is with the Qiyuan Laboratory, Beijing 100095, China.

Digital Object Identifier 10.1109/TGRS.2023.3254523

I. Introduction

With the increasing development of spectral imaging technology, the hyperspectral image (HSI) has higher dimensions and resolution, and the HSI classification task is more challenging than the ordinary image classification task. Faced with such high-dimensional image data, many researchers focus on the preprocessing of HSI data. The traditional dimensionality reduction methods include *principal component analysis* (PCA) [1], *independent component analysis* (ICA) [2], *linear identification analysis* (LDA) [3], and *locally linear embedding* (LLE) [4]. In addition, *Laplacian eigenmaps* (LEP) [5] is similar to LLE, and it constructs the relationship between the data from the local perspective. These methods can achieve the purpose of image dimension reduction, but some frequency bands which have some impact on the classification results will be lost. To minimize the loss of spatial–spectral information, Xia et al. [6] proposed a novel ensemble approach, namely, *rotation random forest via kernel PCA* (RoRF-KPCA) to increase the diversity that characterizes the ensemble architecture. And Zhang et al. [7] proposed a superpixelwise PCA method for unsupervised feature extraction of HSI.

Deep learning (DL) methods have been widely used in various computer vision (CV) tasks, due to their powerful feature representation and learning ability [8], [9], [10], [11]. Early deep network models represented stacked autoencoders (SAEs) and deep blief networks (DBNs) [12], and then the emergence of AlexNet [13] brought the field of DL into a new era. The convolutional neural network (CNN) has developed rapidly in a very short time. And all kinds of neural network models have been applied to the field of CV, including image classification [14], object detection [15], tracking [16], and so on. Since the spectral information represented by each pixel can be regarded as a 1-D vector in HSI, Li and He [17] proposed a modified CNN based on 1-D-CNN and 2-D-CNN. Using a network

which contains 2-D–3-D CNN and multibranch feature fusion, Ge et al. [18] proposed an HSI classification method to retain the connection between the spatial context and the spectra. To address the huge computational overhead caused by 3-D-CNN, Jia et al. [19] proposed a lightweight CNN, which greatly saves computational overhead while ensuring classification accuracy. In the face of rich spatial spectral information, deep models always perform better than shallow models when the deep models have enough training samples to show the model representation ability of the characteristics. Zhang et al. [20] used transfer learning (TL) methods which contain the cross-sensor strategy and the cross-modal strategy to alleviate the problems caused by insufficient training samples. To alleviate the problem of an insufficient training sample, Yu et al. [21] proposed an unsupervised domain adaptation architecture with dense-based compaction (UDAD) for HSI classification (HSIC). From the perspective of spectral sequences in HSI, considering that the RNN has powerful sequence data processing power, Hang et al. [22] used RNN to effectively analyze hyperspectral pixels into sequence data. And an automated RNN method for HSI classification was proposed by Feng et al. [23]. Also, other network models based on *graph attention networks* (GANs) [24] and *graph convolutional networks* (GCNs) [25] are also applied to HSI classification and show good performance.

In general, the HSI is divided into several cubes before being fed into the network. Thus, adjacent space between pixels will be very similar, and for the local feature difference small fine-grained image, traditional convolutional network feature recognition and extraction ability are very limited. Inspired by the human visual attention mechanisms, the attention mechanisms are applied to all kinds of CV tasks. It obtains more detailed information by scanning the global image to find important area, and in the process of subsequent feature learning and representation to the area allocated more attention resources. At the same time, it can reduce the attention of irrelevant pixels. The use of attention mechanisms helps the tasks screen out important information. To explore bidirectional spectral correlation of HSI, Mei et al. [26] proposed an attention-based *bidirectional long short-term memory* (Bi-LSTM)-based network. The emergence of transformer [27] has dramatically transformed the NLP field, and this brand-new architecture uses the self-attention mechanism to process data sequences. The vision transformer (ViT) [28] was widely used in various tasks in the field of CV. Also, transformer has a strong ability to identify and extract deep semantic features, and Chen et al. [29] applied ViT in HSI classification, further showing the good performance of transformer in visual tasks.

Compared with the rapid development of hyperspectral imaging technology, the HSI data processing technology lags behind, which brings many challenges and difficulties to its application. The specific limitations can be roughly summarized as follows.

1) The HSI has typical high data volume characteristics due to large data volume. Within the same band, the HSI also has rich spatial context information. Due to the particularity of HSI acquisition from the remote sensing perspective, the training samples in HSI are highly scattered and sparse. But there are close connections between the different pixels.
2) In addition, the bands of the HSI have a strong correlation, and the spectral correlation coefficient of the image is large, which is easy to cause hyperspectral redundant information stacking. Moreover, the redundancy increases with the increasing number of imaging bands and imaging resolution, and it has a typical high redundancy characteristic.
3) Today, the public HSI datasets are rich and cover a wide range of scenarios. With the rapid development of spectral imaging technology, the images collected by different sensors are different, and the existing HSI image classification models generally have poor generalization ability. Therefore, how to design more robust models and loss functions is also challenging.

To solve these problems, we design a network called MATNet to achieve HSI classification. First, the *channel–spatial attention* (CSA) module is designed to select bands and spatial areas. And it helps pay more attention to somewhere important and reduce redundant information. For initially extracted features through Gaussian weighted feature Tokenizer, the feature map is defined as semantic tokens. These tokens represent high-level semantic concepts of different feature categories. Our previous work [30], [31] has proved the guiding role of depth features and category-level semantic features for feature learning. Thus, we design a *multilayer dense adaptive fusion* (MDAF) encoder structure to fully understand the transformed tokens. Finally, faced with the problem of large gaps in the number of classified samples, we propose a modified loss function called Lpoly loss. And the Lpoly loss adjusts the first polynomial coefficient of the cross-entropy (CE) based on the label smooth CE to further reduce the risk of overfitting. We perform experiments on three famous HSI classification datasets. And the results show that our proposed MATNet performs better than most previous networks for HSI classification.

The specific contributions can be roughly summarized as follows.

1) We propose an HSI classification network called MATNet that integrates multi-attention and transformer to understand the semantic information of HSI.
2) In the spatial–spectral feature extraction stage, we adopt the CSA module to pay more attention to the key areas that have greater influence on the classification results. The use of CSA ignores the redundancy between similar information as possible.
3) We add a polynomial to the label smoothing cross-entropy (LSCE) loss and call the modified loss as the Lpoly loss. It adjusts the degree of learning of different prediction labels to perform robust on different datasets.
4) To ease the imbalance of feature representation between low-level and high-level encoders, we propose the MDAF structure. The structure combines with different degrees of long connection and short connection. Moreover, it helps enrich the mixed information. The experimental results and visualization validate that the

proposed method achieves more robust performance than most previous networks.

II. Related Works

Here, we briefly review the related works about attention mechanisms. In dealing with the HSI classification task, the use of attention mechanisms in HSI greatly helps the classification task because HSI contains rich and diverse spatial, channel, and spectral information. Thus, it is urgent to screen its huge amount of information.

A. Spatial Attention

In HSI, the local critical areas are often concentrated in a small area. When some cases of mixed image elements or image elements are encountered on the classification boundary, the learning model can only classify if the correct image element features are found. But these features are usually in a small area in the image element neighborhood block. Therefore, if the model performs feature extraction on the full eigenvectors, suboptimal results may result due to the presence of certain unrelated regions in the eigenvectors. Using spatial attention (SA), we attempt to focus more on the spatial regions associated with the correct features. Chen et al. [32] proposed the double attention block so that 2-D global features could be fully extracted by subsequent convolution layers. In addition, the SA in [33] focuses on "where" is an informative part, which is complementary to the channel attention (CA).

However, for SA, the features in each channel of the image perform the same SA operation, and no attention to the information interaction between channel and channel.

B. Channel Attention

As we all know, one of the characteristics of HSI is the ultrahigh number of channels. The 3-D convolutional layer can only focus on the relationship between local adjacent frequency bands, and it is difficult to build the connection between distant frequency bands. The use of CA can enable the network to learn the weights of each channel and distinguish the importance between different channels. It also can strengthen the attention of channels with great influence on classification accuracy, while reducing the attention of weaker channels. It is also effective to establish the correlation between different channels. In [34], the importance of each feature channel was automatically learned. And then the useful features were promoted. In contrast, the unimportant features were suppressed. Wang et al. [35] proposed an *efficient CA* (ECA) module which ensures the efficient feature extraction and reduces the computational complexity.

For CA, each channel performs a global feature extraction. But it does not focus on the local and local connections, ignoring the information interaction within the space of each channel.

C. Spectral Attention

Unlike ordinary images, the high-dimensional frequency bands of HSI hid a lot of time series information, and the combination of the same pixel position can be regarded as the 1-D spectral data. The attention mechanism on the hyperspectral band selection is helpful to fully learn the interdependence and nonlinear relationship between bands to large information. It is significant for the bands to give more weight and realize bands' selection. A general architecture, the *spectral attention autoregressive model* (SAAM), was proposed in [36]. This architecture used two spectral attention models to determine relevant global patterns and remove local context's noise while performing the forecasting. Spectral attention is still in the development stage, and it is believed that there will be more effective ways to focus on the choice of spectral bands.

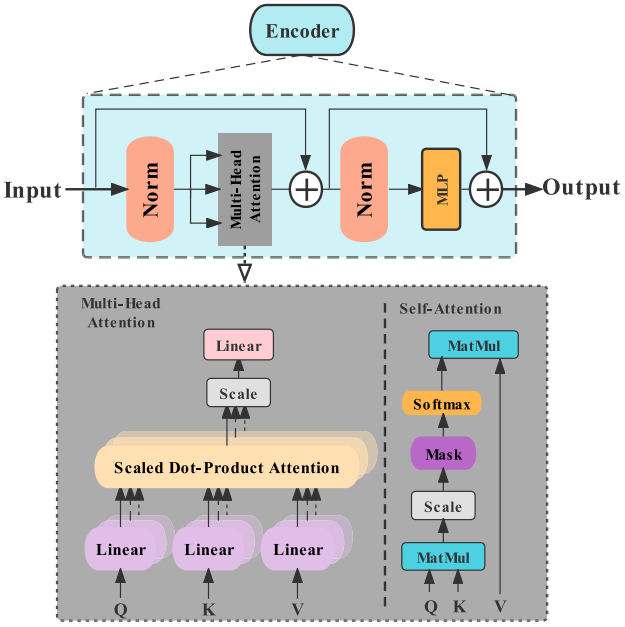

Fig. 1. Transformer encoder structure with the multi-head attention mechanism.

D. Self-Attention

To reduce the dependence on external information, the feature information inherent inside the image needs to be fully learned. It requires us to use the observed within the same sample to predict other regions. The introduction of the self-attention mechanism can capture the dependencies between the input sequence terms, especially those long-range dependencies that are easily lost in the sequence models. The NonLocalBlock designed by Wang et al. [37] effectively characterized the dependencies between more distant pixels or regions on the image. It fully used the spatial self-attention mechanism, but does not focus on the channel.

A large part of transformer's [27] great success in NLP field is reflected in the use of the self-attention mechanism. Unlike the dependence between RNN attention sequences, the transformer encoder introduces a brand new multi-head self-attention mechanism to obtain correlations between any position in space and makes more full use of internal information. As shown in Fig. 1, applying transformer to the image classification task, ViT's [28] performance is impressive.

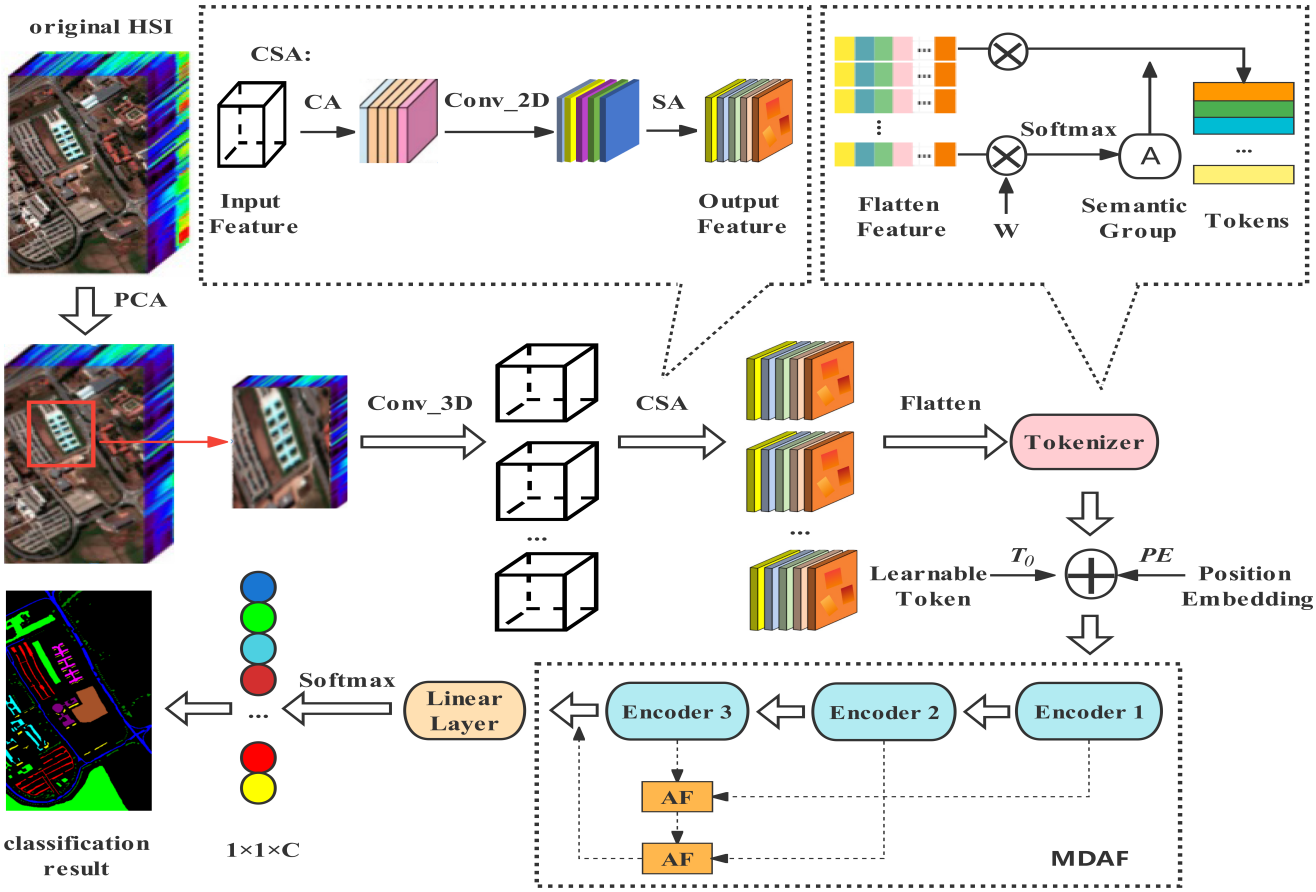

Fig. 2. Overview of the proposed MATNet framework for HSI classification.

He et al. [38] proposed HSI-Bert which is a pure transformer-based method that can capture the global dependence between the pixels, regardless of their spatial distance. Zhao et al. proposed convolutional transformer network (CTN) which combines the CNN and the transformer. It uses central location coding to combine pixel locations with spectral features to generate spatial location features and further obtain local–global features. In addition, Jia and Wang [39] proposed a multiscale convolution embedding module combining transformer to efficiently use unlabeled data, thus enabling the efficient extraction of spatial–spectral information. To pay more attention to effective information, Bai et al. [40] proposed a new multibranch transformer structure called SST-M, which collects SA and extracts spectral features.

Recent work has dedicated transformer's self-attention to build connections within the spectra. Hong et al. [41] proposed the SpectralFormer which is a novel way for extracting locally spectral representations.

III. Proposed Method

In this section, we will detail the specific details of the structure of the network. The proposed MATNet is shown in Fig. 2. It mainly includes a 3-D convolution layer, a CSA module combining CA, a 2-D convolution layer, and SA. In addition, a Gaussian weighted feature tokenizer module, an MDAF module, and softmax function are included in the MATNet.

A. Channel–Spatial Attention

We adopt a combination of CA and SA. The use of CA and SA helps network put more resources on the key bands and areas of the image and ignores interference information as far as possible. What is more, we add a 2-D convolutional layer between the CA and the SA. According to the research in [42], we can find that the spatial information always makes more important influence than spectral information in HSI classification. By adding a 2-D convolution layer, the network can learn more 2-D spatial features before performing the SA, which also allows the subsequent SA to pay arbitrary attention to the more detailed areas at different angles. Yin et al. [43] proposed an HSI classification network called the *multibranch 3D-dense attention (MBDA) network* which only contained the SA.

1) Channel Attention: In general, we will reduce the dimensions of HSI before extracting the features. Although some classical dimensional reduction methods can initially screen out the frequency bands with high information redundancy, the retained frequency bands have different importance for the classification results [44]. Thus, the use of mixed attention can fully focus on areas that have a greater impact on the classification results. Regarding the design of CA, the specific process is shown in Fig. 3. For input features, the average pooling layer and the maximum pooling layer are aggregated. And then the resulting two different spatial context descriptors $\boldsymbol{F}^c_{\text{avg}}$ and $\boldsymbol{F}^c_{\text{max}}$ are input into the same shared *multilayer*

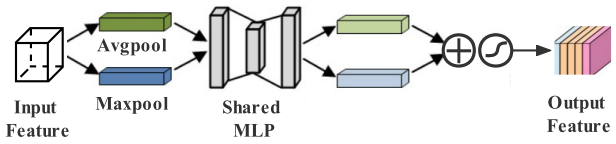

Fig. 3. Channel attention.

Fig. 4. Spatial attention.

perceptron (MLP), where the hidden activation size of MLP is set to r, and then sum element by element to obtain the final output features. The process is expressed by a mathematical formula as follows:

$$\begin{aligned}\mathbf{CA}(\mathbf{F}) &= \mathbf{Sig}(\mathrm{MLP}(\mathrm{Avg\,Pool}(\mathbf{F})) + \mathrm{MLP}(\mathrm{Max}\ P\ \mathrm{ool}(\mathbf{F}))) \\ &= \mathbf{Sig}\left(\mathbf{W_1}\left(\mathbf{W_0}\left(\mathbf{F^c_{avg}}\right)\right) + \mathbf{W_1}\left(\mathbf{W_0}\left(\mathbf{F^c_{max}}\right)\right)\right)\end{aligned} \tag{1}$$

where $\boldsymbol{F}$ represents the input feature, **Sig** denotes the sigmoid function, $\mathbf{W_0} \in R^{C/r\times C}$, and $\mathbf{W_1} \in R^{C\times C/r}$. Moreover, r is the reduction ratio and C is the number of channels. $\boldsymbol{W_0}$ and $\boldsymbol{W_1}$ are the MLP weights, and they are shared for inputs.

2) Spatial Attention: Due to the limited receptive fields of convolution and pooling operations, more correlations between neighboring pixels are often considered. The SA mechanism module considers the importance of all the positions from the global perspective, which fully reflects the superiority of SA. The specific process is shown in Fig. 4. This part first takes the input features through average pooling and maximum pooling operations separately, connects the resulting features, and finally generates SA feature maps using a convolutional layer. The process is expressed by a mathematical formula as follows:

$$\begin{aligned}\mathbf{SA}(\mathbf{F}) &= \mathbf{Sig}\left(f^{3\times 3}\left(\left[\mathrm{Avg\,Pool}(\mathbf{F}); \mathrm{Max\,Pool}(\mathbf{F})\right]\right)\right) \\ &= \mathbf{Sig}\left(f^{3\times 3}\left(\left[\mathbf{F^s_{avg}}; \mathbf{F^s_{max}}\right]\right)\right)\end{aligned} \tag{2}$$

where **Sig** denotes the sigmoid function, F is the input feature, and $f^{3\times 3}$ represents a 3×3 convolution operation.

B. Gaussian Weighted Feature Tokenizer

The flattened feature map of the input is defined as $\boldsymbol{X} \in \boldsymbol{R}^{hw\times c}$, where h is the height, w is the width, and c is the number of channels. The input features X and W are the 1×1 pointwise product. Then, the size of result after being transposed is $\boldsymbol{R}^{t\times hw}$. Next, the semantic information is concerned by the softmax function. $\boldsymbol{A}$ is represented by $\mathbf{softmax}(\boldsymbol{X} * \boldsymbol{W})^T$. Finally, A multiplies with X to make final semantic tokens $\boldsymbol{T}$ and $\boldsymbol{T} \in \boldsymbol{R}^{t\times c}$, where t indicates the number of tokens. This module makes the samples more separable with the consistence between semantic features and distribution features expressed by the tokens.

The semantic tokens T can be expressed as follows:

$$\boldsymbol{T} = \mathbf{softmax}(\boldsymbol{X} * \boldsymbol{W})^T \boldsymbol{X} \tag{3}$$

where $\boldsymbol{W}$ represents a weight matrix initialized with a Gaussian distribution, and it is defined as $\boldsymbol{W} \in \boldsymbol{R}^{c\times t}$. In addition, $*$ represents the 1×1 pointwise product.

C. Transformer Encoder

The traditional CNN-based HSI classification networks are limited by the size of the convolutional kernel. He and Chen [45] proposed a CNN-based spatial transformer network to optimize the original input. However, due to the limitations of the convolution operations, the obtained optimization results are often local rather than global. In the task of HSI classification, we often consider whether pixels with similar features belong to the same category. It is urgent to require that the network architecture should take into account the connections between the global pixels. The transformer encoder has a strong ability to focus on the correlation between pixels using the multi-head attention mechanism. The internal correlation obtained by integrating multiple self-attention layers reduces the dependence on the external information. The multi-head self-attention mechanism is shown in Fig. 1.

1) Multi-head Self-attention: For self-attention used in the transformer encoder, three learnable weight matrices are defined, including queries Q, keys K, and values V. Q and K are used to calculate the attention score, and the softmax function is used to calculate the weight of the score. In summary, self-attention is formulated as follows:

$$\mathrm{Self\_Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\boldsymbol{T}}}{\sqrt{d_K}}\right)\boldsymbol{V} \tag{4}$$

where d_K represents the dimension of K.

For multi-head attention, it involves the weight matrices of multiple groups in Q, K, and V. Then, the calculated values of each head attention are joined together.

We can obtain representations of attention. At the same time, they are linked into a larger feature matrix. Finally, the linear transformation matrix is used to ensure that the feature dimension is the same as the input data. Because self-attention cannot use the sequence information. Thus, we encode the location information into the feature embedding after transformed semantic tokens.

D. MDAF Module

The representation of low-level features has limited capability. And the high-level encoder may lose shallow features of ground categories, such as shape and color. Different from the proposed *DSS-TRM* [46] method which simply stacks the transformer layer, we use the cross-layer connection to fuse information between different layers. In this way, we can get the characteristics of the multilevel information which are more suitable for fine-grained classification. The module consists of adaptive fusion (AF) and multilayer dense connection.

1) Adaptive Fusion: For AF used in MDAF, it is shown in Fig. 5. The adaptation is mainly reflected in: for any two $h \times 1$ size feature vectors, first it is connected along the dimension of width, and then the $h\times 2$ size vector is convoluted in 2-D with size 1×2 along the h dimension. Finally, the feature vector of $h \times 1$ size is still obtained. In our

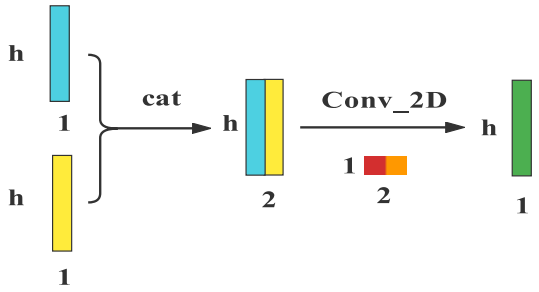

Fig. 5. Adaptive fusion.

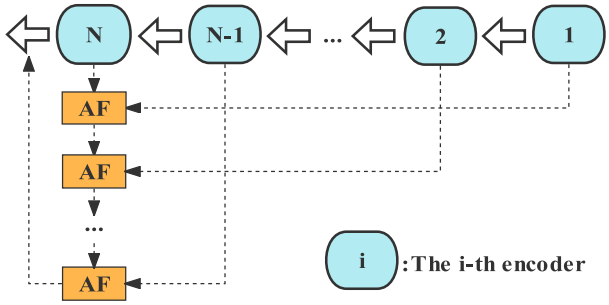

Fig. 6. Multilayer dense connection.

network, h was set to 201. The AF operation can effectively eliminate the redundant information between different layers and can also retain the specific information from different levels. Moreover, we can embed the AF into the network structure according to the needs of the task, without changing the size of the input and output tensors during the process. The process is expressed by a mathematical formula as follows:

$$L_k = \textbf{Conv\_2D}\big[\textbf{cat}\big(L_i, L_j\big)\big] \tag{5}$$

where **cat** represents the splicing operation, and the kernel size of **Conv_2D** is 1×2.

2) Multilayer Dense Connection: We can use the tokenizer module to fully obtain the characteristics describing the ground object. To extract deep semantic features, we choose to use the multilayer encoder structure. There are two main reasons. For one reason, features in the deep layers are more abstract than shallow ones. At the same time, there is a risk of information loss due to the defects of the convolution operation. To this end, we choose to adopt a cross-layer connection method. First, shallow and deep information perform an AF, and then the fusion information and middle information perform an AF. Our proposed connection method is more flexible than the connection method used in the SpectralFormer [41] method. Not only that, the fusion features through *multilayer dense connection* contain more information.

The structural diagram of this connection mode is shown in Fig. 6. We can choose the suitable number of layers according to the different needs of the task. We adjust the number of encoders in the module and conclude that the three-layer encoder structure can extract deep enough semantic features, but also ensure the small computational overhead. And we can express the process with the following formula:

$$\textbf{Output} = \textbf{AF}\{[\textbf{AF}(E_1, E_3)], E_2\} \tag{6}$$

where **AF** represents the process of AF. E_i represents the output of the ith encoder.

E. Label Smoothing Poly (Lpoly) Loss Function

1) Label Smoothing Cross-Entropy: The CE loss is widely used in classification tasks. But only the loss of the correct label position is considered. In this way, the loss of other label positions is ignored. The formula of CE is expressed as (7). The HSI classification with fewer samples carries the risk of overfitting. To alleviate this problem, the LSCE loss is proposed with adopting the regularization method. As a result, it helps increase the final loss and improve the learning ability of the network. The LSCE is expressed by a mathematical formula as (8)

$$L_{\mathrm{CE}} = \frac{1}{N}\sum_i L_i = -\frac{1}{N}\sum_i \sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{7}$$

where M is the number of categories, and y_{ic} stands for the sign function (0 or 1) if the true class of sample i is equal to c equals 1, otherwise 0. The p_{ic} represents the predicted probability that the observed sample i belongs to class c

$$\mathrm{LSCE} = (1-\lambda)\mathrm{CE}(i) + \lambda\sum \frac{\mathrm{CE}(j)}{N} \tag{8}$$

where $\mathrm{CE}(i)$ represents the standard CE loss of i, λ is a small positive number, i is the correct class, j is the incorrect class, and N is the number of classes.

2) Label Smoothing Poly (Lpoly): Label smoothing is a regularization strategy that ultimately suppresses the overfitting mainly by reducing the weight of the class of real sample labels when calculating the loss function. When the parameters in the formula are determined, the correspondence between the true label and the error label probability is also determined. For the CE loss, it can be decomposed into a series of weighted polynomial bases. Equation (9) shows that each polynomial basis is weighted by the corresponding polynomial coefficients, but the coefficient assignment may not be optimal for different tasks and datasets. For HSI, the distribution of samples is unbalanced between different ground categories. Moreover, due to the large gap in the HSI spatial–spectral information collected by different sensors, we also hope to dynamically change the smoothness of different prediction labels to make the classification results more stable. It also helps alleviate the problem of the limited number of HSI samples.

The Taylor expansion of the CE loss in the bases of $(1-P_t)^i$ is given as follows:

$$\begin{aligned} L_{\mathrm{CE}} &= -\log(P_t) \\ &= \sum_{m=1}^{\infty} 1/m(1-P_t)^m \\ &= (1-P_t) + 1/2(1-P_t)^2, \ldots \end{aligned} \tag{9}$$

where P_t is the model's prediction probability of the target ground-truth class.

Optimizing the CE loss using gradient descent requires taking the gradient of P_t

$$-\frac{\mathrm{d}L_{\mathrm{CE}}}{\mathrm{d}P_t} = \sum_{n=1}^{\infty}(1-P_t)^{n-1} = 1+(1-P_t)+(1-P_t)^2,\ldots \quad (10)$$

According to (9) and (10), we adjust the first polynomial coefficient of the CE based on the LSCE. In this way, we can dynamically adjust the relationship between the predicted category probabilities according to the huge variability of the dataset. The formula is expressed as follows:

$$L_{\mathrm{Lpoly}} = (1-\lambda)\mathrm{CE}(i) + \lambda\sum\frac{\mathrm{CE}(j)}{N} + \epsilon(1-P_t) \quad (11)$$

where $\mathrm{CE}(i)$ represents the standard CE loss of i, i is the correct class, j is the incorrect class, λ is a small positive number, ϵ is a number which is greater than -1 to ensure monotonicity, and N is the number of classes.

F. Implementation

To fully show the specific details of the MATNet network structure, we present the process in the form of Algorithm 1.

Algorithm 1 Multi-Attention-Based Joint Transformer Network

Input: An HSI cube data

Output: Predicted labels of the test samples.

Initialization:

Set batch size to 64; PCA bands' number to 40; patch size to 13; training sample rate μ; learning rate to $1e-3$; tokens' number to 200; loss function parameters λ and ϵ.

Repeat:

1: Perform a 3D convolution layer to obtain 3D feature maps;
2: Perform the CSA module that incorporates channel attention, a 2D convolution layer, and spatial attention to obtain 2D feature maps;
3: Flatten 2D feature maps into 1-D feature vectors;
4: Generate semantic tokens via tokenizer module;
5: Concatenate the learnable classification tokens $\boldsymbol{T}_0$ to get semantic tokens;
6: Embed position information on the semantic tokens;
7: Perform the MDAF module;
8: Input the first classification token to the last linear layer;
9: Use the softmax function to identify the labels;

end for: Predict test samples using the trained model.

Next, we take the Indian Pines dataset as an example to illustrate our designed MATNet. The process is shown in Table I.

IV. Experiments

In this section, we list three datasets for HSI classification used in the experiments. Then we verify the effective performance of the network, and the ablation experiments are performed on these modules and the modified loss function.

TABLE I
EXAMPLE OF INDIAN PINES DATASET TO ILLUSTRATE THE PROPOSED MATNET

Step	Step Name	Size
1	Original HSI	$145\times145\times200$
2	PCA dimensional reduction	$145\times145\times40$
3	Patch extraction	$13\times13\times40$
4	$3D$ convolution	$16\ 11\times11\times38$
5	Reshape	$11\times11\times608$
6	$2D$ convolution	$64\ 9\times9$
7	Flatten	$64\ 1\times81$
8	Reshape	64×81
9	Tokenizer	$\boldsymbol{T}\in\boldsymbol{R}^{200\times64}$
10	Add learnable token $\boldsymbol{T}_0$	$\boldsymbol{T}_{input}\in\boldsymbol{R}^{201\times64}$
11	Position Embedding	$\boldsymbol{T}_{input}\in\boldsymbol{R}^{201\times64}$
12	MDAF	$\boldsymbol{T}_{output}\in\boldsymbol{R}^{201\times64}$
13	Take out Classification token	$\boldsymbol{T}_0^{out}\in\boldsymbol{R}^{1\times64}$
14	Linear layer and softmax	$\boldsymbol{P}_i\in\boldsymbol{R}^{1\times16}$

TABLE II
DETAILED CATEGORIES AND NUMBER OF SPECIFIC TRAINING AND TEST SET SAMPLES FOR INDIAN PINES

Class No.	Class Name	Training	Testing
1	Alfalfa	5	41
2	Corn Notill	143	1285
3	Corn Mintill	84	746
4	Corn	24	213
5	Grass Pasture	48	435
6	Grass Trees	74	656
7	Grass Pasture Mowed	3	25
8	Hay Windrowed	48	430
9	Oats	2	18
10	Soybean Notill	97	875
11	Soybean Mintill	246	2209
12	Soybean Clean	59	534
13	Wheat	21	184
14	Woods	129	1136
15	Buildings Grass Trees	40	346
16	Stone Steel Towers	10	83
	Total	1033	9216

A. HSI Datasets

To demonstrate the performance of the MATNet, we used three public HSI datasets. They are Indian Pines, Pavia University, and Houston2013 datasets. The proportion of the training set and other details are shown in Table II–IV.

B. Experimental Setup

1) Evaluation Metrics: We evaluate the classification performance on three compelling indices. They are *overall accuracy* (OA), *average accuracy* (AA), and *kappa coefficient* (κ). The larger the values of the three metrics, the better the classification result [30], [31], [47], [48].

2) Configuration: Our proposed MATNet was implemented under the Mindspore and PyTorch framework. The workstation used in the proposed MATNet is Intel Xeon(R) Silver 4210R

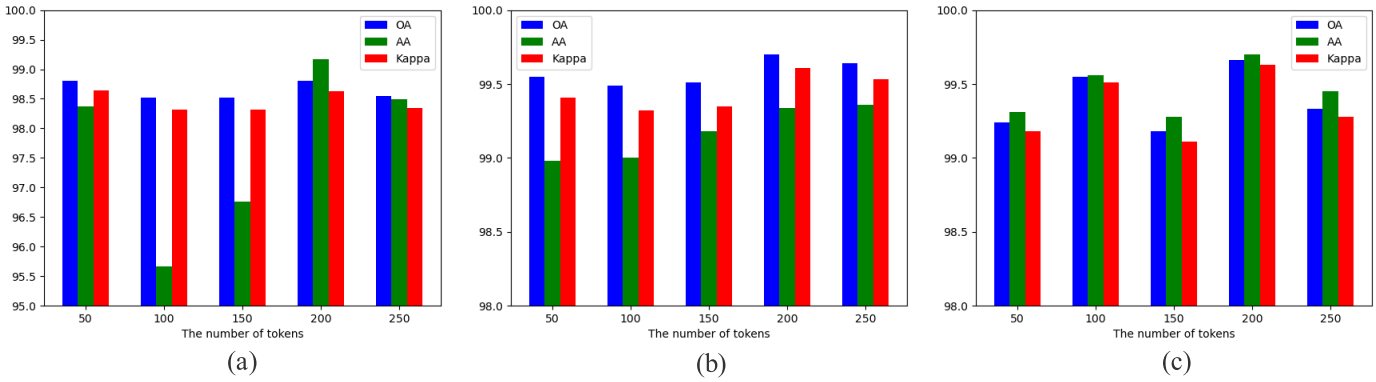

Fig. 7. Impact of different tokens for OA, AA, and κ on three datasets. (a) Indian Pines. (b) Pavia University. (c) Houston2013.

TABLE III
DETAILED CATEGORIES AND NUMBER OF SPECIFIC TRAINING AND TEST SET SAMPLES FOR PAVIA UNIVERSITY

Class No.	Class Name	Training	Testing
1	Asphalt	340	6291
2	Meadows	960	17689
3	Gravel	113	1986
4	Trees	157	2907
5	Metal Sheets	72	1273
6	Bare Soil	256	4773
7	Bitumen	67	1263
8	Bricks	187	3495
9	Shadows	48	899
	Total	2200	40576

CPU, 64-GB RAM, and an NVIDIA Quadro RTX 5000 16-GB GPU. The optimizer [49] used in our network was Adam. The learning rate was initialized and set to $1e-3$. We found in practice that the network has different feature extraction capabilities for different datasets. During the experiment, we concluded that the features of three datasets could be fully extracted by the network before 300 epochs, so we set the number of epochs to 300.

3) Comparison With State-of-the-Art Backbone Networks: To prove the performance of the proposed MATNet, we selected some popular methods for comparison: Support vector machine (SVM), extended morphological attribute profiles (EMAP) [50], 1-D-CNN [51], 2-D-CNN [52], 3-D-CNN [53], SSRN [54], Cubic-CNN [55], HybridSN [56], spectral-spatial feature tokenization transformer (SSFTT) [57], MBDA [43], and our proposed MATNet.

1) For SVM, EMAP, 1-D-CNN, 2-D-CNN, 3-D-CNN, and SSFTT, all the network settings are the same as the details in SSFTT [57].
2) For other networks, the parameters are set the same as in SSRN [54], Cubic-CNN [55], HybridSN [56], and MBDA [43].
3) For the proposed MATNet, the bands' number is reduced to 40 by performing PCA. The 3-D convolutional layer consists of 16 $3 \times 3 \times 3$ cores. The 2-D convolutional layer consists of 64 3×3 cores. The kernel size of the SA module is 3. The extracted patch is set to 13×13. The number of transformer encoder heads is 200, and the number of encoder layers in the MDAF module was set to 3.

For a fair comparison, the number of samples used to train and test is the same proportion of the total sample number for all the methods.

4) Parameter Analysis:

1) For the patchsize, we empirically set the size of patch to [9], [11], [13], [15], [17], [19] separately and perform contrast experiments on the Indian Pines dataset, finally set to 13. The specific results are shown in Fig. 6.
2) For the CSA module, our convolutional kernel sizes in this module are all 3, and choosing a smaller kernel size can greatly reduce the computation complexity. What is more, replacing large convolutions with multiple small convolutions can also increase the nonlinear expression power of the network.
3) For the tokenizer module, which is a high-level semantic concept used to represent and process HSI feature categories. Due to the large differences in different datasets, we set the number of last generated semantic tokens on three datasets. The experimental results are shown in Fig. 7. Based on the experimental results, we find that the number of tokens set to 200 is more suitable for the proposed framework.
4) For the MDAF module, we experiment on the number of layers of the encoder based on the number of tokens in the final output of the tokenizer module, and Table VII shows that using the three-layer encoder structure can learn and represent the features best. In addition, we also compare the training time and the test time of different encoder structures. Table VIII shows that the more the encoders used, the greater the cost of time. In summary, we selected a three-layer encoder structure in MATNet.
5) For the loss function, considering different spatial spectral information contained by the different datasets, we adopt the method of grid search and set different parameters to accommodate the different datasets. The classification results of different parameters are shown in Fig. 8. For the first parameter λ, we synthesize the overall performance and uniformly set the parameter

TABLE IV
DETAILS OF THREE HSI DATASETS AND THE PROPORTION OF THE TRAINING AND TEST SAMPLES

Dataset Name	Year	Equipment	Bands Number	Wavelength Range	Pixel Size	Training Proportion
Indian Pines	1992	AVIRIS sensor	200	400nm-2500nm	145×145	10%
Pavia University	2001	ROSIS sensor	103	0.43-0.86μm	610×340	5%
Houston2013	2013	ITRES CASI-1500 sensor	144	364nm-1046nm	349×1905	10%

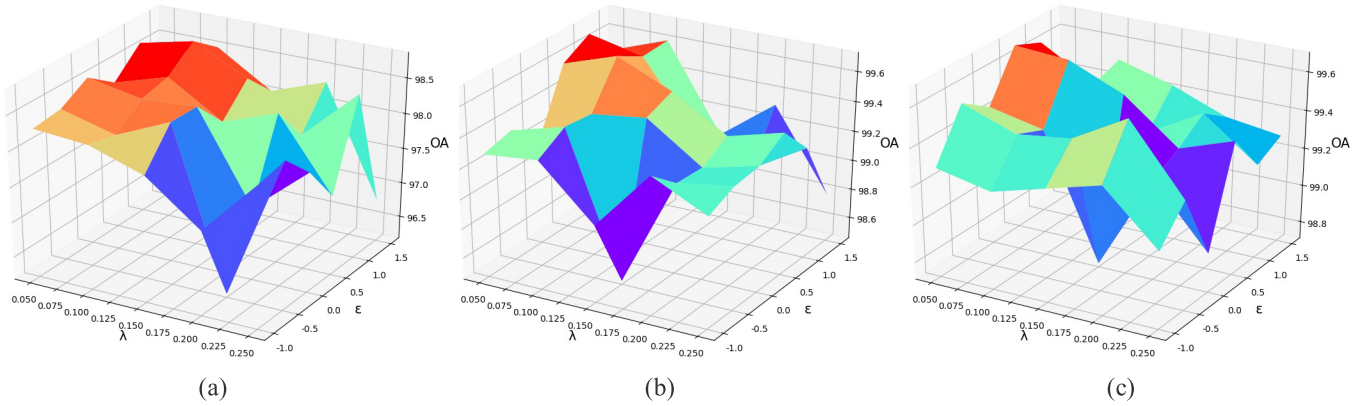

Fig. 8. Impact of different parameters from the loss function for OA on three datasets. (a) Indian Pines. (b) Pavia University. (c) Houston2013.

TABLE V
DETAILED CATEGORIES AND NUMBER OF SPECIFIC TRAINING AND TEST SET SAMPLES FOR HOUSTON2013

Class No.	Class Name	Training	Testing
1	Healthy Grass	130	1121
2	Stressed Grass	127	1127
3	Synthetic Grass	73	624
4	Tree	125	1119
5	Soil	126	1116
6	Water	33	292
7	Residential	127	1141
8	Commercial	124	1120
9	Road	126	1126
10	Highway	126	1101
11	Railway	123	1112
12	Parking Lot1	127	1106
13	Parking Lot2	48	421
14	Tennis Court	43	385
15	Running Track	67	593
	Total	1525	13504

TABLE VI
IMPACT AND FLOPS OF DIFFERENT PATCHSIZES ON THE INDIAN PINES DATASET. THE BOLD ONE IS THE OPTIMAL RESULT

Metric	The size of patch					
	9	11	13	15	17	19
OA (%)	96.94	98.38	**98.80**	98.47	98.36	98.08
AA (%)	95.11	97.99	**99.17**	97.85	93.67	95.92
κ*100	96.51	98.16	**98.63**	98.25	98.13	97.81
FLOPs(G)	1.982	2.560	3.328	4.285	5.431	6.767

to 0.1. As for the second parameter ϵ, it worked best when ϵ was set to 1 on Indian Pines, and the Pavia University and Houston2013 datasets were 0.5.

TABLE VII
IMPACT OF DIFFERENT ENCODER LAYERS ON THE HOUSTON2013 DATASET. THE BOLD ONE IS THE OPTIMAL RESULT

Metric	The number of encoder layers					
	1	2	3	4	5	6
OA (%)	99.22	99.26	**99.66**	99.58	99.44	99.30
AA (%)	99.32	99.34	**99.70**	99.57	99.53	99.33
κ*100	99.15	99.20	**99.63**	99.54	99.39	99.24

TABLE VIII
COST OF TIME BETWEEN DIFFERENT ENCODER LAYERS ON THE HOUSTON2013 DATASET

time	The number of encoder layers					
	1	2	3	4	5	6
training(s)	213.99	254.76	344.52	443.32	450.89	507.38
test(s)	1.51	2.01	2.68	3.22	3.70	4.24
FLOPs(G)	2.205	2.435	3.328	4.220	5.112	6.005
params(M)	0.417	0.435	0.615	0.714	0.813	0.912

C. Ablation Study

To estimate the effectiveness of the modules in the proposed MATNet, we perform the following ablation experiments.

1) CSA Module: For the CSA module, to validate the ability of the CA and SA to focus on important information, we design four different cases on Honston2013 to test separately, as shown in Table X. The results show that the classification result of using both CA and SA is the best, and the result of using neither CA nor SA is the worst. In addition, only CA is better than only SA.

At the same time, to verify that the CSA module we used has a competitive effect, we also compare with other attention mechanisms on Indian Pines, which are Non_Local [37], SELayer [58], scSE [59], and ECALayer [35]. We replaced the CSA using other attention modules and kept the other model structures consistent. The experimental results are shown in Table XI.

TABLE IX

CLASSIFICATION RESULTS OF SPECIFIC CATEGORIES OF DIFFERENT NETWORKS ON THE INDIAN PINES. THE BOLD ONE IS THE OPTIMAL RESULT

Class No.	SVM	EMAP	1-D CNN	2-D CNN	3-D CNN	SSRN	Cubic-CNN	HybridSN	SSFTT	MBDA	MATNet
1	65.63	62.50	43.75	48.78	41.46	83.15	87.86	87.80	95.12	92.43	**100.00**
2	63.44	81.57	77.93	78.13	90.51	95.31	96.35	94.39	97.67	**98.43**	95.64
3	60.25	83.19	56.72	83.51	79.36	94.23	93.65	96.52	98.87	97.72	**99.33**
4	41.11	85.89	45.18	47.42	46.01	90.68	82.54	83.89	91.55	99.20	**100.00**
5	87.05	78.61	87.57	75.12	95.17	97.79	96.69	98.16	96.32	97.24	**99.31**
6	97.21	79.08	98.63	92.99	**99.70**	98.67	96.69	99.54	99.54	98.88	99.54
7	89.47	52.63	65.11	60.00	88.00	97.92	90.16	92.97	**100.00**	**100.00**	**100.00**
8	96.66	91.19	97.36	98.37	**100.00**	99.26	98.46	**100.00**	**100.00**	**100.00**	**100.00**
9	32.26	50.12	37.14	66.67	48.89	89.49	89.93	86.27	88.89	82.77	**100.00**
10	73.84	81.32	66.03	87.77	86.06	97.48	93.94	97.94	97.71	**99.68**	98.97
11	84.36	86.91	82.49	89.09	97.51	98.16	97.45	**99.50**	98.69	99.04	99.09
12	42.89	78.43	73.49	63.67	74.91	93.07	93.18	94.57	**98.13**	96.04	97.75
13	98.58	96.35	99.30	**100.00**	99.46	98.59	99.12	94.59	97.28	99.89	99.46
14	94.02	93.91	93.78	95.33	99.74	99.72	99.39	99.29	99.91	99.70	**100.00**
15	42.65	77.36	55.39	66.76	84.10	93.31	84.26	92.35	98.84	99.79	**100.00**
16	92.19	84.38	81.54	91.57	93.98	93.79	89.69	**97.98**	85.54	93.09	97.59
OA (%)	76.39	83.69	79.37	84.47	91.03	94.78	94.90	96.62	97.47	98.68	**98.80**
AA (%)	72.18	76.53	70.87	77.83	82.18	94.87	93.85	95.66	96.57	97.71	**99.17**
κ*100	72.85	81.40	76.28	82.24	89.68	94.08	94.17	96.29	97.11	98.49	**98.63**

TABLE X

ABLATION EXPERIMENTS OF THE COMBINATION OF CSA ON HOUSTON2013. THE BOLD ONE IS THE OPTIMAL RESULT

Module		Metric		
CA	SA	OA (%)	AA (%)	κ*100
no	no	98.99	99.08	98.90
yes	no	99.33	99.36	99.27
no	yes	99.15	99.21	99.08
yes	yes	**99.66**	**99.70**	**99.63**

TABLE XI

IMPACT OF DIFFERENT ATTENTION ON THE INDIAN PINES DATASET. THE BOLD ONE IS THE OPTIMAL RESULT

Metric	Non_local	SELayer	scSE	ECALayer	CSA
OA (%)	98.07	98.20	98.42	98.50	**98.80**
AA (%)	93.12	96.15	98.11	96.41	**99.17**
κ*100	97.80	97.95	98.20	98.29	**98.63**

TABLE XII

ABLATION ANALYSIS OF THE PROPOSED STRUCTURE ON THREE DATASETS. THE BOLD ONE IS THE OPTIMAL RESULT

Metric	Indian Pines		Pavia University		Honston2013	
	MDAF	WITHOUT	MDAF	WITHOUT	MDAF	WITHOUT
OA (%)	**98.80**	98.27	**99.70**	99.58	**99.66**	99.32
AA (%)	**99.17**	93.54	**99.34**	99.19	**99.70**	99.42
κ*100	**98.63**	98.03	**99.61**	99.54	**99.63**	99.26

TABLE XIII

IMPACT OF DIFFERENT LOSS FUNCTIONS ON THREE DATASETS. THE BOLD ONE IS THE OPTIMAL RESULT

Metric	Indian Pines			Pavia University			Honston2013		
	LPOLY	LSCE	CE	LPOLY	LSCE	CE	LPOLY	LSCE	CE
OA (%)	**98.80**	97.82	97.09	**99.70**	99.61	99.46	**99.66**	99.39	99.15
AA (%)	**99.17**	96.88	94.29	**99.34**	99.22	98.80	**99.70**	99.48	99.28
κ*100	**98.63**	97.51	96.68	**99.61**	99.48	99.28	**99.63**	99.34	99.08

2) MDAF Module: For the MDAF module, to evaluate the effectiveness of our designed connection structure and feature fusion, we test the strategy without using any feature fusion (WITHOUT) on three datasets. And the results in Table XII show that our proposed MDAF module effectively incorporates important features between different levels to prevent the loss of shallow features.

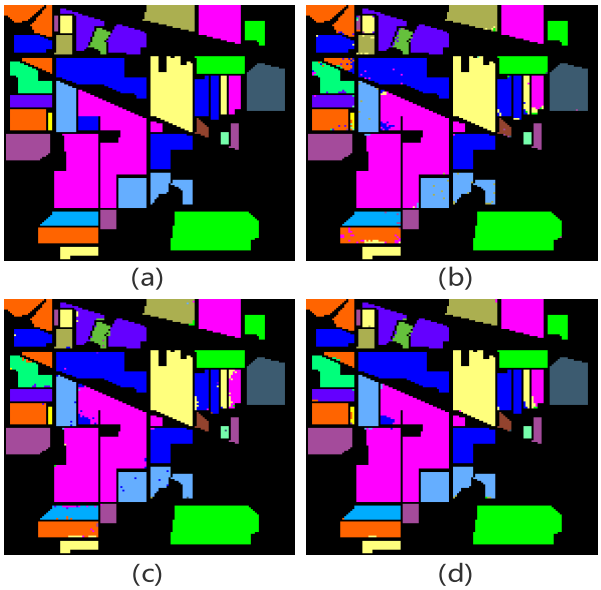

Fig. 9. Results' visualization of the Indian Pines dataset. (a) Ground truth. (b) CE. (c) LSCE. (d) Lpoly.

3) Loss Function: For the loss function, we tested different loss functions to verify that the Lpoly loss function makes the network more robust. Other compared loss functions are the traditional CE loss function and LSCE loss function. The results are shown in Table XIII, and we can find that the Lpoly loss function performs the best.

D. Classification Results and Analysis

1) Classification Results: Detailed classification results are shown in Tables IX, XIV, and XV for three HSI datasets. The results of the comparison mainly involve the above three indicators.

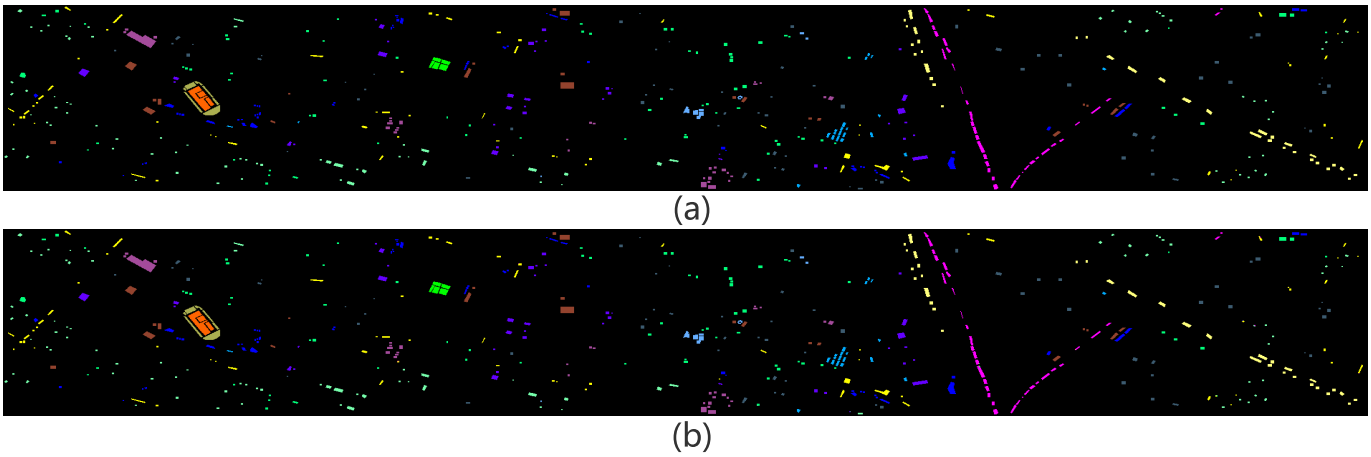

Fig. 10. Results' visualization of the Houston2013 dataset. (a) Ground truth. (b) MATNet.

TABLE XIV
CLASSIFICATION RESULTS OF SPECIFIC CATEGORIES OF DIFFERENT NETWORKS ON PAVIA UNIVERSITY. THE BOLD ONE IS THE OPTIMAL RESULT

Class No.	SVM	EMAP	1-D CNN	2-D CNN	3-D CNN	SSRN	Cubic-CNN	HybridSN	SSFTT	MBDA	MATNet
1	91.22	89.74	84.91	93.28	93.30	95.35	94.48	95.51	99.33	**100.00**	99.98
2	97.78	93.50	94.14	94.90	93.99	94.69	93.96	99.49	99.92	**99.99**	**99.99**
3	34.95	42.06	47.38	75.55	90.19	96.48	93.49	94.18	98.29	**99.36**	98.59
4	81.55	81.54	82.28	93.87	91.29	96.37	98.01	**99.55**	98.49	97.93	98.35
5	98.59	97.96	99.76	97.98	95.47	99.69	99.76	96.71	99.53	**99.92**	**99.92**
6	41.50	48.18	77.67	70.05	93.85	97.49	93.48	99.43	**100.00**	**100.00**	**100.00**
7	16.71	31.58	19.40	70.92	81.45	95.36	96.48	**100.00**	99.13	99.92	**100.00**
8	89.74	91.45	70.01	90.30	92.73	91.49	92.51	95.97	98.05	99.34	**99.43**
9	**99.89**	98.78	98.55	97.89	95.46	95.90	95.35	95.22	95.44	96.45	97.78
OA (%)	82.76	84.89	83.50	90.19	92.01	95.87	95.68	98.16	99.21	99.68	**99.70**
AA (%)	72.44	76.36	74.90	87.31	90.45	95.86	95.28	97.35	98.69	99.21	**99.34**
κ*100	76.28	80.41	77.90	87.52	90.87	95.78	95.55	97.57	99.15	99.57	**99.61**

TABLE XV
CLASSIFICATION RESULTS OF SPECIFIC CATEGORIES OF DIFFERENT NETWORKS ON HOUSTON2013. THE BOLD ONE IS THE OPTIMAL RESULT

Class No.	SVM	EMAP	1-D CNN	2-D CNN	3-D CNN	SSRN	Cubic-CNN	HybridSN	SSFTT	MBDA	MATNet
1	95.65	87.85	86.20	94.02	93.75	94.98	94.46	98.85	98.84	99.15	**99.82**
2	97.52	92.53	95.13	96.30	94.91	95.60	96.65	99.73	99.38	99.93	**100.00**
3	99.84	99.84	**100.00**	89.73	95.54	97.14	96.84	99.84	99.52	99.09	99.84
4	93.66	92.63	95.43	98.31	94.91	99.45	94.56	96.07	98.39	98.15	**99.91**
5	98.30	97.26	98.98	97.88	**100.00**	99.80	99.83	**100.00**	**100.00**	**100.00**	**100.00**
6	84.25	82.16	95.15	73.46	89.86	94.98	94.59	**100.00**	**100.00**	99.08	**100.00**
7	82.65	80.47	76.60	89.63	91.84	88.49	93.48	97.63	96.67	**99.24**	99.04
8	56.61	70.51	58.12	80.96	80.36	95.86	96.49	97.95	97.68	**99.18**	97.77
9	73.91	72.69	63.16	70.98	93.58	92.78	91.21	98.67	99.29	99.71	**100.00**
10	83.51	86.78	55.57	84.65	94.28	96.49	95.89	99.00	98.73	**100.00**	**100.00**
11	68.97	64.59	70.16	90.96	92.64	97.85	97.98	99.28	99.91	**100.00**	**100.00**
12	60.72	59.18	54.74	91.46	94.37	99.32	98.12	99.46	99.55	98.56	**99.64**
13	25.69	45.29	41.93	85.65	90.97	92.69	96.36	98.10	98.10	**100.00**	99.52
14	93.25	96.48	97.05	73.71	94.58	98.46	97.46	**100.00**	**100.00**	**100.00**	**100.00**
15	99.66	98.45	99.04	99.52	99.65	99.02	99.12	99.49	99.16	**100.00**	**100.00**
OA (%)	80.92	82.39	77.64	89.05	93.73	96.89	97.11	98.80	98.92	99.46	**99.66**
AA (%)	79.61	78.49	79.15	87.82	93.56	96.19	96.23	98.93	99.01	99.47	**99.70**
κ*100	79.33	80.64	75.81	88.15	93.63	96.54	96.68	98.71	98.83	99.42	**99.63**

Overall, traditional classifiers, such as SVM and EMAP, achieve ordinary classification performance. In addition, the classical backbone network significantly outperformed the above traditional classifiers, such as SSRN and Cubic-CNN. The results largely demonstrate the value and utility of DL methods in HSI classification. For CNNs, 3-D-CNN learns

TABLE XVI
CLASSIFICATION RESULTS OF SPECIFIC CATEGORIES OF DIFFERENT NETWORKS ON THE WHU-HI DATASET. THE BOLD ONE IS THE OPTIMAL RESULT

Metric	WHU-Hi-HanChuan				
	Resnet	DPyResnet	A^2S^2K	SSFTT	MATNet
1	95.37	97.06	98.48	**99.93**	99.85
2	96.84	96.51	98.67	96.63	**99.38**
3	95.90	89.09	98.80	99.70	**99.90**
4	96.81	97.80	99.47	98.82	**99.80**
5	84.47	92,00	96.52	**99.91**	99.82
6	81.58	79.08	92.68	93.40	**97.49**
7	87.40	91.60	97.22	97.04	**99.11**
8	94.42	96.74	96.49	99.53	**99.55**
9	91.43	91.99	97.46	98.19	**99.47**
10	97.78	99.12	98.69	99.47	**99.75**
11	94.54	85.50	98.56	99.23	**99.79**
12	84.92	81.45	97.16	99.37	**99.97**
13	81.18	85.86	95.52	93.63	**96.30**
14	95.05	95.81	98.06	99.01	**99.78**
15	89.63	92.16	**98.28**	95.09	96.85
16	98.96	99.87	99.88	**99.95**	99.90
OA (%)	95.19	95.18	98.47	98.96	**99.58**
AA (%)	91.64	91.98	97.62	98.06	**99.17**
κ*100	94.37	94.37	98.21	98.78	**99.51**

and retains more correlations between space and spectra than 2-D-CNN and 1-D-CNN. The HybirdSN network mixing 3-D convolution and 2-D convolution fully preserves the spatiotemporal features of HSI for further improved accuracy. As an emerging network architecture, the transformer can extract highly sequential representations from HSI. The SSFTT demonstrated better results in experiments, demonstrating the transformer architecture's powerful feature extraction and representation capabilities. Also, MBDA showed suboptimal results which played the role of the attention mechanism.

As for our proposed MATNet network model, we can see in Tables IX, XIV, and XV that the model performs best on all three experimental datasets. For Indian Pines, our method improved by 1.33%, 2.60%, and 1.52% in OA, AA, and κ, respectively, compared with the newer SSFTT method. Not only that, more than half of the categories achieved the highest classification accuracy on each dataset. Overall, the experimental results demonstrate that our proposed MATNet performs stably in different scenarios, with relatively robust and generalization capabilities.

2) Visual Evaluation: We visually present the classification results obtained by the experiments on the three datasets, as shown in Figs. 9–11.

In addition, we can also see from the classification graph that the proposed network structure is relatively stable for a large number of categories or a small number of categories. To verify that the proposed Lpoly loss function is more suitable for the fine-grained classification task than the CE and LSCE loss functions, we present the classification results using different loss functions in Fig. 9. It is worth mentioning that MATNet using the Lpoly loss function has a high classification accuracy for every category. It is demonstrated again that the proposed MATNet performs robust.

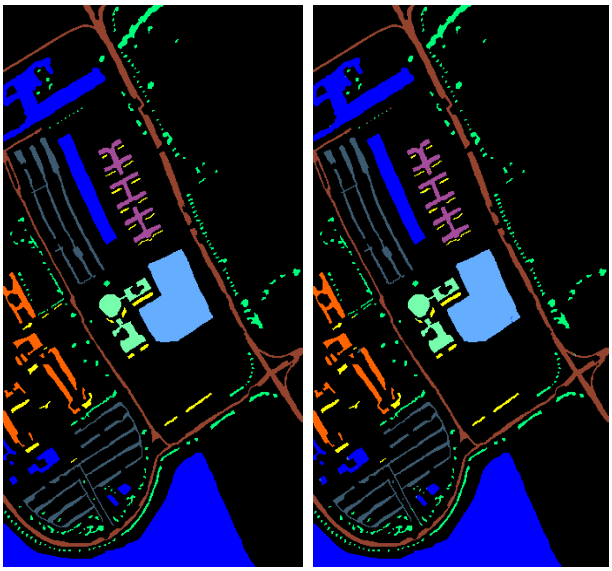

Fig. 11. Results' visualization of the Pavia University dataset. (a) Ground truth. (b) MATNet.

E. Supplementary Experiments

To further validate the robust feature extraction capabilities of our proposed MATNet, we perform some comparative experiments on dataset with larger numbers of samples. The dataset we selected was collected under an agricultural scenario, and the spatial–spectral features are very similar between different categories, which verifies the performance of our proposed MATNet.

1) WHU-Hi HanChuan Dataset: The WHU-Hi-Hanchuan dataset was collected from 17:57 to 18:46, June 2016, in Hanchuan, Hubei Province, China, with a 17-mm focal length Headwall Nano-Hyperspec imaging sensor equipped on a Leica Aibot X6 UAV V1 platform. The study area is an urban–rural fringe area with 16 different ground categories. The HSI was taken at an altitude of 250 m, with an image size of 1217×303 pixels and 274 bands between 400 and 1000 nm. The spatial resolution of the HSIs carried by the drone is about 0.109 m.

2) Experiments: We select some popular deep learning methods for comparison: Resnet [60], DPyResnet [61], A^2S^2K [62], SSFTT [57], and our proposed MATNet. These more advanced methods include both the networks with attention-based networks and networks using transformer structures. The results of the methods are all very competitive in Table XVI. By the way, we adjust the network parameters consistently in the experimental setup to ensure fairness in the contrast experiment. And the number of training samples in each category represents 5% of the total.

3) Visualization of Results: To visually see the effectiveness of our proposed MATNet, we contrast the classification results with the ground truth and the most competitive result (SSFTT) for visualization. The visual classification results are shown in Fig. 12. From the enlarged regions, it is obvious that our visualization results show less scattered misclassification regions than the classification results obtained by SSFTT.

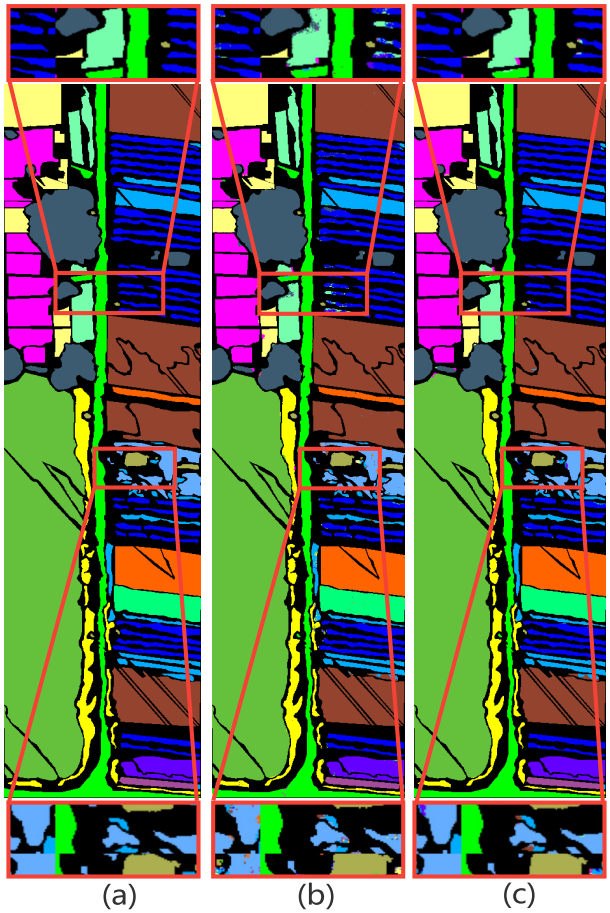

Fig. 12. Results' visualization of the WHU-Hi-HanChuan dataset. (a) Ground truth. (b) SSFTT. (c) MATNet.

This also proves that our method has a more stable and more robust performance.

V. Conclusion

In this article, we propose a network called MATNet that efficiently extracts the deep semantic features of HSI images. First, after the dimension reduction of HSI, a module containing CA, 2-D convolution layer, and SA is used to initially screen out important bands and important regions in space. Then, the feature mapping is defined as semantic tokens by the tokenizer module. These tokens represent the high-level semantic concepts of different HSI categories. Next, the abstract semantic features are modeled by a three-layer transformer encoder dense adaptive connection structure. And the multilevel fusion enriches the feature representation of different categories. Finally, the classification results are predicted by the linear layer and the softmax layer. The experimental results show that our proposed network is more effective compared with other traditional and popular methods. Meanwhile, the ablation studies also prove that the modules used in the network structure have improved the classification results to different degrees.

Based on the significant effect of attention mechanisms demonstrated in the classification task, we will explore the rich and diverse spectral information in HSI in the later work. Further improving on the basis of the existing spectral attention, we try to integrate the attention mechanisms from different angles to further learn deeper spatial–spectral information. In addition, the cross-layer fusion methods of different structures have different degrees of feature extraction, and how to reasonably design the connection methods for different tasks is also our concern in the future.

Acknowledgment

The authors thank MindSpore for the partial support of this work, which is a new deep learning computing framework (https://www.mindspore.cn/).

References

[1] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, 2010.

[2] A. Villa, J. A. Benediktsson, J. Chanussot, and C. Jutten, "Hyperspectral image classification with independent component discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4865–4876, Dec. 2011.

[3] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.

[4] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000, doi: 10.1126/science.290.5500.2323.

[5] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA, USA: MIT Press, 2001, pp. 585–591. [Online]. Available: https://proceedings.neurips.cc/paper/2001/hash/f106b7f99d2cb30c3db1c3cc0fde9ccb-Abstract.html

[6] J. Xia, N. Falco, J. A. Benediktsson, P. Du, and J. Chanussot, "Hyperspectral image classification with rotation random forest via KPCA," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1601–1609, Apr. 2017.

[7] X. Zhang, X. Jiang, J. Jiang, Y. Zhang, X. Liu, and Z. Cai, "Spectral–spatial and superpixelwise PCA for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5502210.

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

[9] E. Li, P. Du, A. Samat, Y. Meng, and M. Che, "Mid-level feature representation via sparse autoencoder for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 1068–1081, Mar. 2017.

[10] R. Kemker and C. Kanan, "Self-taught feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2693–2705, May 2017.

[11] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Apr. 2019.

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

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst., 26th Annu. Conf. Neural Inf. Process. Syst. Meeting Held*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Lake Tahoe, NV, USA, 2012, pp. 1106–1114. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[14] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[16] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional Siamese networks for object tracking," in *Computer Vision ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016, pp. 850–865.

[17] Y. Li and L. He, "An improved hybrid CNN for hyperspectral image classification," in *Proc. 11th Int. Conf. Graph. Image Process. (ICGIP)*, Jan. 2020, pp. 485–490.

[18] Z. Ge, G. Cao, X. Li, and P. Fu, "Hyperspectral image classification method based on 2D–3D CNN and multibranch feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5776–5788, 2020.

[19] S. Jia et al., "A lightweight convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4150–4163, May 2020.

[20] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, Aug. 2019.

[21] C. Yu, C. Liu, H. Yu, M. Song, and C.-I. Chang, "Unsupervised domain adaptation with dense-based compaction for hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12287–12299, 2021.

[22] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.

[23] J. Feng, G. Bai, Z. Gao, X. Zhang, and X. Tang, "Automatic design recurrent neural network for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 2234–2237.

[24] A. Sha, B. Wang, X. Wu, and L. Zhang, "Semisupervised classification for hyperspectral images using graph attention networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 157–161, Jan. 2021.

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

[26] S. Mei, X. Li, X. Liu, H. Cai, and Q. Du, "Hyperspectral image classification using attention-based bidirectional long short-term memory network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5509612.

[27] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon et al., Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–15. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[28] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[29] X. Chen, S.-I. Kamata, and W. Zhou, "Hyperspectral image classification based on multi-stage vision transformer with stacked samples," in *Proc. TENCON IEEE Region Conf. (TENCON)*, Dec. 2021, pp. 441–446.

[30] Y. Chen and X. Lu, "Deep category-level and regularized hashing with global semantic similarity learning," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 6240–6252, Dec. 2021.

[31] Y. Chen, X. Lu, and S. Wang, "Deep cross-modal image-voice retrieval in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7049–7061, Oct. 2020.

[32] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "A^2-nets: Double attention networks," 2018, *arXiv:1810.11579*.

[33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 3–19.

[34] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[36] F. Moreno-Pino, P. M. Olmos, and A. Artés-Rodríguez, "Deep autoregressive models with spectral attention," 2021, *arXiv:2107.05984*.

[37] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," 2017, *arXiv:1711.07971*.

[38] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Sep. 2019.

[39] S. Jia and Y. Wang, "Multiscale convolutional transformer with center mask pretraining for hyperspectral image classification," 2022, *arXiv:2203.04771*.

[40] J. Bai et al., "Hyperspectral image classification based on multibranch attention transformer networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5535317.

[41] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[42] Y. Xu, B. Du, and L. Zhang, "Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification," *IEEE Trans. Big Data*, vol. 6, no. 3, pp. 492–506, Sep. 2020.

[43] J. Yin, C. Qi, W. Huang, Q. Chen, and J. Qu, "Multibranch 3D-dense attention network for hyperspectral image classification," *IEEE Access*, vol. 10, pp. 71886–71898, 2022.

[44] C. Yu, S. Zhou, M. Song, and C.-I. Chang, "Semisupervised hyperspectral band selection based on dual-constrained low-rank representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[45] X. He and Y. Chen, "Optimized input for CNN-based hyperspectral image classification using spatial transformer network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1884–1888, Dec. 2019.

[46] B. Liu, A. Yu, K. Gao, X. Tan, Y. Sun, and X. Yu, "DSS-TRM: Deep spatial–spectral transformer for hyperspectral image classification," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 103–114, Dec. 2022, doi: 10.1080/22797254.2021.2023910.

[47] Y. Chen, X. Lu, and X. Li, "Supervised deep hashing with a joint deep network," *Pattern Recognit.*, vol. 105, Sep. 2020, Art. no. 107368, doi: 10.1016/j.patcog.2020.107368.

[48] Y. Chen, S. Xiong, L. Mou, and X. X. Zhu, "Deep quadruple-based hashing for remote sensing image-sound retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4705814, doi: 10.1109/TGRS.2022.3155283.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

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

[51] W. Hu, Y. Huang, W. Li, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, Jul. 2015, Art. no. 258619, doi: 10.1155/2015/258619.

[52] W. Shao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Oct. 2016.

[53] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Jul. 2016.

[54] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.

[55] J. Wang, X. Song, L. Sun, W. Huang, and J. Wang, "A novel cubic convolutional neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4133–4148, 2020.

[56] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D-2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Jun. 2020.

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

[58] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[59] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham, Switzerland: Springer, 2018, pp. 421–429.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[61] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral-spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[62] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral–spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021.

Bo Zhang received the B.S. degree in engineering from Jiangxi Normal University, Nanchang, China, in 2021. He is currently pursuing the master's degree with the School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan, China.

His research interests include pattern recognition and image classification in remote sensing scenarios.

Yaxiong Chen is an Associate Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, China. His research interests include pattern recognition, machine learning, hyperspectral image analysis, and medical imaging.

Yi Rong received the B.Sc. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2012, and the M.Sc. degree in software engineering and the Ph.D. degree in computer science and technology from the Wuhan University of Technology, Wuhan, in 2014 and 2021, respectively.

He has been a Visiting Ph.D. Student with the School of Engineering, Griffith University, Southport, QLD, Australia, and the School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. His research interests include machine learning, computer vision, pattern recognition, few-shot image classification, and intelligent agriculture.

Shengwu Xiong received the B.Sc. degree in computational mathematics and the M.Sc. and Ph.D. degrees in computer software and theory from Wuhan University, Wuhan, China, in 1987, 1997, and 2003, respectively.

Currently, he is a Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan. His research interests include intelligent computing, machine learning, and pattern recognition.

Xiaoqiang Lu (Senior Member, IEEE) is currently a Full Professor with Qiyuan Laboratory, Beijing, China. His research interests include intelligent optical sensing, pattern recognition, machine learning, and hyperspectral image analysis.