Research Article

Multiscale Dense Cross-Attention Mechanism with Covariance Pooling for Hyperspectral Image Scene Classification

Runmin Liu ^{1,2}, Xin Ning ³, Weiwei Cai ⁴, and Guangjun Li ¹

¹College of Sports Engineering & Information Technology, Wuhan Sports University, Wuhan 430079, China

²School of Graduate, Wuhan Sports University, Wuhan 430079, China

³Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China

⁴Central South University of Forestry and Technology, Changsha 410004, China

Correspondence should be addressed to Guangjun Li; liguangjun@whsu.edu.cn

Received 19 March 2021; Accepted 24 April 2021; Published 4 May 2021

Academic Editor: Fazlullah Khan

Copyright © 2021 Runmin Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In recent years, learning algorithms based on deep convolution frameworks have gradually become the research hotspots in hyperspectral image classification tasks. However, in the classification process, high-dimensionality problems with large amounts of data and feature redundancy with interspectral correlation of hyperspectral images have not been solved efficiently. Therefore, this paper investigates data dimensionality reduction and feature extraction and proposes a novel multiscale dense cross-attention mechanism algorithm with covariance pooling (MDCA-CP) for hyperspectral image scene classification. The multisize convolution module can detect subtle changes in the hyperspectral images' spatial and spectral dimensions between the pixels in the local areas and are suitable for extracting hyperspectral data with complex and diverse types of structures. For traditional algorithms that assign attention weights in a one-way manner, thus leading to the loss of feature information, the dense cross-attention mechanism proposed in this study can jointly distribute the attention weights horizontally and vertically to efficiently capture the most representative features. In addition, this study also uses covariance pooling to further extract the features of hyperspectral images from the second order. Experiments have been conducted on three well-known hyperspectral datasets, and the results thus obtained show that the MDCA-CP algorithm is superior compared to the other well-known methods.

1. Introduction

Hyperspectral remote sensing technology can generate a set of images with rich information. It can significantly improve data analysis quality, with significant improvements in the method's detail, reliability, and credibility. In general, because different categories of the same feature target exhibit different optical behavior and shown obvious differences in the band space, the feature type's pixel level can be used for identification and classification. This technology is nowadays being used in agriculture [1–3], environmental remote sensing [4–6], physics [7, 8], ocean [9–11], and other fields, among which the research on hyperspectral image (HSI) classification is extremely important. The internal information of an image contains various types of features. It is precisely because of the difference in the various forms of features inside an image that they can be extracted for

analysis. Finally, the method of distinguishing the target into different categories is called image classification. HSI classification enables identifying different targets of the ground features based on the obtained remote sensing information. In addition, it can also analyze the inherent laws of the ground features during the classification process. As HSI acquisition costs are relatively high, this will lead to an insufficient number of training samples. Further, the image itself has a relatively high spectral dimension and a relatively large amount of data, which increases its complexity.

In the early research, spectral information for classification was a popular procedure [12–14]. The methods of feature selection and dimensionality reduction [15, 16] were often used to ease the spectral dimension's high dimensionality. With advances in this research topic, the complex space and spectral feature distribution of HSI have become the main problems that plague their classification. Many

researchers have chosen to add local spatial connections to improve the models [16, 17] and have also achieved positive results to a certain extent. However, these methods are primarily based on manual and shallow models, which rely heavily on expert knowledge and have poor generalization ability, making it difficult to extract the representative discriminating features.

In recent years, deep learning models have been fruitful and have emerged as a new research hotspot among the different methods used for HSI classification. Although deep learning methods can obtain deep-level feature representations from the original HSIs, detection and recognition usually rely on many training samples. However, it is time-consuming and labor-intensive to label the data used for HSI classification and to use it for training. The annotation data are very limited. In view of the above-mentioned problems, maintaining deep learning methods with fewer training samples is still a significant challenge. In 2014, Chen et al. [18] proposed a deep learning framework that combines spatial and spectral features. The deep learning framework combines the principal component analysis (PCA) and deep learning architecture employing stacked autoencoders (SAE) to obtain the deep features using logistic regression (LR) results in classification (SAE-LR). Although the SAE-LR deep learning method has shown great potential in HSI classification, the autoencoder model usually flattens the local image patches into vectors and then feeds them into the model. However, this method destroys the image's two-dimensional structure, resulting in the loss of spatial information and the training time is also insufficient. In 2015, Makantasis et al. [19] used convolutional neural networks (CNNs) to develop a deep-supervised HSI classification algorithm. Their algorithm uses random PCA to reduce the original input data's dimensionality, a CNN to extract the deep-level features, and multilayer perceptron (MLP) for classification. In 2016, Zhao and Du proposed the spectral-spatial feature classification (SSFC) method. The framework uses a balanced local discriminant embedding (BLDE) algorithm to extract the spectral features based on the classification's multi-feature classifiers.

In addition, three-dimensional spectral-spatial feature extraction methods have also been developed to discover useful but difficult-to-extract information from HSIs. Tsai et al. [20] proposed a three-dimensional gray-level co-occurrence feature extraction method for HSI classification based on the traditional two-dimensional morphological operators, and Weeks et al. [21] proposed a three-dimensional discrete wavelet transform, which is a series of one-dimensional wavelet transforms along the three-direction sequence of HSI processing and classification. Feature extraction from HSIs is important for improving the classification performance of the method. The related model of deep learning provides a new solution to the problems in HSI classification. However, applying deep learning to such a classification task also faces some difficulties.

- (1) The high-dimensional problem of HSIs is different from ordinary natural images. HSI data present a three-dimensional structure. The spectral information is very rich, and the spatial information is relatively small. The application of deep learning for the classification of HSIs is the first step to designing a network structure suitable for HSI structure, which can effectively use the HSIs to enrich the spectral information and the limited spatial information.
- (2) In the HSI data, there is often a serious imbalance between the number of classes and the number of classes. If the weight constraints of various samples are not imposed, the trained network's robustness is low.

Based on the above limitations, a novel multiscale dense cross-attention mechanism algorithm with covariance pooling (MDCA-CP) has been proposed in this work for HSI scene classification. Multiscale convolution can detect the spatial dimensions of HSIs. The subtle changes between the pixels in the local areas in the spectral dimension can be applied to the feature extraction of the hyperspectral data of complex and diverse types and structures. Traditional algorithms only assign attention weights in a one-way manner, resulting in the loss of feature information. The dense cross-attention mechanism proposed in this work can simultaneously distribute the attention weights horizontally and vertically to efficiently fetch the most representative hyperspectral data. In addition, covariance pooling has been used in this study to further extract the second-order features of HSIs. Experiments have been conducted on three well-known hyperspectral datasets, and the results thus obtained prove the effectiveness of the MDCA-CP algorithm.

The main contributions of this work are as follows:

- (1) A novel dense cross-attention (horizontal and vertical) mechanism algorithm has been proposed, and weight addition and maximum weight strategies have been constructed, which can mine more representative features in the hyperspectral data.
- (2) A novel attention-guided covariance pooling method has been proposed to make full use of the second-order information in the hyperspectral data feature map. It allows the neural network to learn more representative features when dealing with remote sensing scene classification problems.
- (3) Many experiments and ablation studies have been conducted on three public hyperspectral datasets of Indian Pines, Salinas, and Pavia University, and visual analysis of the results has been done, proving its effectiveness superiority of the MDCA-CP algorithm.

The rest of the paper has been organized as follows: Section 2 briefly describes other works related to this article. Section 3 introduces the principle and implementation details of the MDCA-CP algorithm. Section 4 presents the experimental results and visual analysis. Section 5 provides a summary of the conclusions based on this research.

2. Related Work

In recent years, deep learning has been mainly based on multilayer neural networks that infinitely approximate the nonlinear functions by using three or more neural networks to learn large amounts of abstract feature information in the image. Because it is considerably difficult to obtain the labels for HSIs, the use of deep learning methods to achieve better classification results while using a small number of training samples has always been the focus of research in the field of hyperspectral remote sensing. The classic neural network models based on deep learning include the autoencoder (AE) [22], stacked autoencoder (SAE) [23], and restricted Boltzmann machine (RBM) [24, 25]. Although these methods have relatively good classification performance, scholars have designed deep neural network models based on CNNs and others due to many model parameters and shallow image features. The methods are combined and applied to the classification of HSIs and provide superior feature extraction effect and high classification performance.

Chen et al. [18] proposed deep feature extraction and HSI classification method based on CNNs. The CNN is used for extracting the deep features from the HSIs that are nonlinear, discriminative, and invariant. A combination of the CNN and a regularization method is used for extracting the spatial-spectral features of HSIs for classification. Xu et al. [26] proposed a multisource remote sensing data classification method based on CNN, which mainly uses two channels. The CNN extracts the spatial-spectral joint features of the HSI, fuses them with the features of other sources of remote sensing data, and classifies them. Luo et al. [27] proposed a new HSI classification method based on CNN in which the spatial-spectral features from the target pixel and its neighboring pixels are extracted. The convolution operations are performed on them, and the convolution results are superimposed into a two-dimensional matrix as the input of the standard CNN. Finally, the XGBoost classification model, proposed by Wang et al. [28], is an HSI classification method based on a combination of the random forest technique and CNN. It regards the CNN as an individual classifier that is used for extracting the discriminating features of HSIs. The random forest technique involves a random selection of the extracted features and training samples to establish a multiclassification system for performing a classification task.

The abovementioned models based on CNNs are all constructed with standard convolutional layers and applied to HSI classification. In order to enable the models to gain the huge advantages of the standard convolutional layer while saving a considerable amount of computing resources, some scholars have introduced the attention mechanism into the classification models. Mei et al. [29] proposed a spectral-spatial attention network for HSI classification. This method can learn the internal spectral correlation in the continuous spectrum, and the attention model focuses on the neighboring pixels in the spatial dimension. Experimental results show that this method can make full use of spectral information. In order to extract the spectral and

spatial features, Ma et al. [30] proposed a dual-branch multiattention mechanism network (DBMA) for HSI classification. The network has two branches that extract the spectral and spatial features, respectively, reducing the two features' interference. In addition, because of the two branches' different characteristics, two types of attention mechanisms are applied in the two branches, respectively, which ensures more distinctive spectral and spatial features. Excellent results have been achieved using this method. The abovementioned research studies show that the attention mechanism [31, 32] is extremely effective in hyperspectral data classification tasks and provides outstanding results.

3. Methodology

Figure 1 shows a schematic of the overall architecture of our MDCA-CP algorithm. First, locality preserving projections (LPPs) are used for reducing the dimensionality of the HSI to eliminate redundant information and reduce computational costs. Second, the reduced-dimensional features are inputted into our MDCA module, and the second-order features are extracted through CP. Finally, the SoftMax function is used for predicting each pixel to achieve the classification of the hyperspectral data.

3.1. Locality Preserving Projections. The spectral image of each pixel of the HSI has multiple endmember spectral images at the same time. In a low-dimensional subspace, the high-dimensional data can be projected onto this low-latitude subspace when the HSI is classified. The contained pixels represent a single endmember, which represents a characteristic of a ground feature. The local preserving dimensionality reduction method proposed in this study can find a linear map M . The adjacent pixels of the original HSI can obtain relatively close projection space distances, therefore achieving effective retention of the relevant information in the local neighborhood and efficiently protecting the diversified local structure of the original HSI. Assuming that the original data's training sample is x_i , the class label x_j is the number of sample categories, x^{th} is the total number of training samples, and the training samples are x_i and x_j . Then, the close relationship equation between x_i and x_j can be expressed as follows:

$$A_{ij} = \exp\left(-\frac{x_i - x_j^2}{\gamma_i - \gamma_j}\right), \quad (1)$$

where $\gamma_i = x_i - x_i^{(m)}$ represents the local scale of the sample x_i and $x_i^{(m)}$ represents the m th adjacent sample of the pixel x_i . The local interclass scatter matrices $L^{(lb)}$ and $L^{(lw)}$ in the local preservation dimensionality reduction are defined as

$$L^{(lb)} = \frac{1}{2} \sum_{i,j=1}^n W_{i,j}^{(lb)} (x_i - x_j)(x_i - x_j)^T, \quad (2)$$

$$L^{(lw)} = \sum_{i,j=1}^n W_{i,j}^{(lw)} (x_i - x_j)(x_i - x_j)^T.$$

FIGURE 1: Schematic showing our MDCA-CP algorithm's overall architecture. $\otimes$ represents the feature fusion operation, whereas $\oplus$ represents the feature concatenation operation.

Both $L^{(lb)}$ and $L^{(lw)}$ are $n \times n$ -dimensional square matrices.

Equation (1) provides the weights of the adjacent pixels of the same kind, and the interclass scatter matrices $L^{(lb)}$ and $L^{(lw)}$ hardly affect the same nonadjacent pixels. By locally spreading the matrix, the maximized Fisher ratio is given by

$$L^{(lb)}T = \delta L^{(lw)}T. \quad (3)$$

The transformation matrix obtained by deformation calculation is

$$T_{\text{LFDA}} = \arg \max_{T \in R^{d \times r}} \text{tr} \left[\left(T_{\text{LFDA}}^T L^{(lw)} T_{\text{LFDA}} \right)^{-1} T_{\text{LFDA}}^T L^{(lb)} T_{\text{LFDA}} \right]. \quad (4)$$

LPPs reduce the adjacent data of the same kind via the transformation matrix, T . They effectively separate the adjacent data of different types while retaining the data's local characteristics. Using the local retention and dimensionality reduction to preprocess the original HSI, the redundant information is eliminated. The distribution of similar classes is made more compact, the noise is reduced, and the classification accuracy is improved.

3.2. Multiscale Convolution. In this work, the use of a multiscale convolution kernel has two major advantages:

The biggest advantage of multiscale convolution kernels is that different sizes of the convolution kernels can extract features of different scales from the HSIs. Thus, the filter can extract and learn richer information from them.

When the CNN training model is used, training is achieved by learning the parameters of the filter (weights and offsets), i.e., by continuously learning the parameters of the filter to achieve an optimal value that is closest to the label. In this study, a multiscale convolution kernel has been used to enable every convolutional layer to have a variety of filters, thus achieving diversification of the weight and bias learning and a complete and effective extraction and learning of the useful information from the HSI.

In computer vision models, multiscale inference methods are usually used for obtaining the best results. Generally, the finer details are efficiently predicted at larger sizes, the larger objects are efficiently predicted at smaller sizes, and the receiving field of the network can accurately understand the scene at smaller sizes. In contrast to the traditional multiscale structure, a novel multiscale dense convolution model (as shown in Figure 2) has been proposed in this study. In particular, we apply the convolution process to extract the features in four sizes: 11×11 , 7×7 , 5×5 , and 3×3 . The corresponding calculation formulae are as follows:

FIGURE 2: Schematic showing the multiscale layer that includes four different scale convolutions. Y_1 , Y_2 , Y_3 , Y_4 , and Y_5 are the feature maps from the different scale convolutions.

$$\begin{aligned}
 Y_1 &= \phi \left\{ \sum_{i=0}^{n1} w_{ij}^1 * x_i^1 + b_j^1 \right\}, \\
 Y_2 &= \text{add} \left[\phi \left\{ \sum_{i=0}^{n1} w_{ij}^1 * x_i^1 + b_j^1 \right\}, \phi \left\{ \sum_{i=0}^{n2} w_{ij}^2 * x_i^2 + b_j^2 \right\} \right], \\
 Y_3 &= \phi \left\{ \sum_{i=0}^{n3} w_{ij}^3 * x_i^3 + b_j^3 \right\}, \\
 Y_4 &= \text{add} \left[\left(\phi \left\{ \sum_{i=0}^{n1} w_{ij}^1 * x_i^1 + b_j^1 \right\}, \phi \left\{ \sum_{i=0}^{n2} w_{ij}^2 * x_i^2 + b_j^2 \right\} \right), \phi \left\{ \sum_{i=0}^{n3} w_{ij}^3 * x_i^3 + b_j^3 \right\} \right], \\
 Y_5 &= \text{add} \left[\left(\left(\phi \left\{ \sum_{i=0}^{n1} w_{ij}^1 * x_i^1 + b_j^1 \right\}, \phi \left\{ \sum_{i=0}^{n2} w_{ij}^2 * x_i^2 + b_j^2 \right\} \right), \phi \left\{ \sum_{i=0}^{n3} w_{ij}^3 * x_i^3 + b_j^3 \right\} \right), \phi \left\{ \sum_{i=0}^{n4} w_{ij}^4 * x_i^4 + b_j^4 \right\} \right],
 \end{aligned} \tag{5}$$

where x_i^1 , x_i^2 , x_i^3 , x_i^4 , w_{ij}^1 , w_{ij}^2 , w_{ij}^3 , and w_{ij}^4 represent the neurons and weights from the kernel of the multiscale convolutional layer and n represents the number of filters in CONV_i. As the output of the convolutional layer, Y is fed to the cross-attention-guided CP module.

3.3. Cross-Attention Mechanism

3.3.1. Attention Mechanism. In essence, deep learning's attention mechanism is similar to the visual attention mechanism humans use when making choices. Both aim to filter out the key information, which is more helpful to the task they are dealing with, from a large amount of information, and suppress the useless information that is not relevant to the task at hand. The attention mechanism is prominent in two aspects: on the one hand, it can select the locally important information that is required to be focused on the overall input by itself; on the other hand, it can reasonably allocate a small amount of important information to the critical task objectives via computing resources. Because these two aspects, which enable the current attention mechanism network to be widely applied in image recognition, stand out, the attention mechanism can strengthen the characteristics of the local information. Further, the attention area position will change depending

on the mission objectives; by understanding the target's local information, the most useful information corresponding to the target can be put to good use. Figure 3 shows the framework of the attention mechanism network. The output of the decoder part can be expressed as follows:

$$S_t = f(s_{t-1}, y_{t-1}, C_t), \tag{6}$$

where S_t is the state output of the decoder at a time t , s_{t-1} is the state output of the decoder at a time $t-1$, y_{t-1} is the label at a time $t-1$, and f is the dense layer.

$$C_t = \sum_{j=1}^{T_x} a_{tj} h_j, \tag{7}$$

where C_t is the output of the next state, h_j is the output of the j th input in the decoder, and a_{tj} is the attention weight.

$$a_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T_x} \exp(e_{tk})}, \tag{8}$$

$$e_{tj} = g(s_{t-1}, h_j) = v \times \tanh(W \times h_j + u \times S_{t-1} + b),$$

where a_{tj} represents the degree of alignment between the current decoder and the j th input, which is the attention weight. g is used for calculating the relationship score

FIGURE 3: Schematic of the attention mechanism model.

between S_{t-1} and h_j . It can be seen that if the score is higher, the attention distribution is focused on the input.

3.3.2. Cross-Attention. Considering that the traditional attention mechanism assigns attention weight from only a single direction, it often leads to the loss of feature information. This study proposes a cross-attention mechanism, which constructs two novel weight distribution strategies: weight addition and maximization strategies. First, the cross-attention mechanism calculates the feature weight coefficients along with the horizontal and vertical directions. These two weight coefficients are added in order to enhance the feature, and the largest weight coefficient is obtained by applying the maximization strategy. Finally, the output of the two strategies is obtained via fusion. Figure 4 shows a schematic diagram of the cross-attention mechanism.

The equations for calculating the addition and maximization of weights are as follows:

$$S_{tj} = \sum_{j=1}^{T_x} \frac{\exp(e_{tj})}{\sum_{k=1}^{T_x} \exp(e_{tk})} h_j, \quad (9)$$

$$\text{add} = (S_1 + S_2),$$

$$\text{max} = (S_1, S_2),$$

$$\text{Cross_att} = \text{concatenate}[S_1, S_2, \text{add}, \text{max}],$$

where S_1 and S_2 represent the weight coefficients of the horizontal and vertical attention mechanisms, respectively, add represents the addition of the weight coefficients, and max represents the maximization operation of the weight coefficients.

FIGURE 4: Schematic of the cross-attention mechanism.

3.4. Covariance Pooling. The convolutional layer of a traditional CNN uses max/average pooling, and the fully connected layer can only capture the first-order information. Although the rectified linear unit (ReLU) introduces non-linearity, it is only limited to the single-pixel level. We believe that the covariance matrix pooling (as shown in Figure 5) can extract the features from HSIs with higher efficiency than the first-order statistics. If it is a set of features, its covariance matrix is

$$CP = \frac{1}{n-1} \sum_{i=1}^n (f_i - \bar{f})(f_i - \bar{f})^T, \quad (10)$$

where $\bar{f} = (1/n) \sum_{i=1}^n f_i$. Only when the number of linear independent components in $\{f_1, f_2, \dots, f_n\}$ is greater than d , the resulting matrix is a symmetric positive definite (SPD) matrix. The covariance matrix must be SPD to use the SPD manifold network's geometric structure to maintain the layer. However, even if the matrix is only positive semi-definite, it can be normalized by adding a multiple of the trace to the diagonal term of the covariance matrix as follows:

$$CP^+ = CP + \lambda \text{trace}(CP)I, \quad (11)$$

where λ is the regular parameter and I is the identity matrix.

3.5. Feature Fusion. The multiscale convolution's output features were spliced, and the CP that is guided by the cross-attention mechanism was used for obtaining the final deep features. The calculation equation for the same is written as follows:

$$O = \text{add}[CP^+(Y_1), CP^+(\text{Cross_att}(Y_2)), CP^+(Y_3), CP^+(\text{Cross_att}(Y_4)), CP^+(\text{Cross_att}(Y_5))], \quad (12)$$

FIGURE 5: Illustration of the concept of covariance pooling.

where O represents the output of the CP module guided by the multiscale cross-attention mechanism.

4. Experiments

4.1. Evaluation Index. The results of the classification of HSIs require an evaluation of the classification ability of the MDCA-CP algorithm. For the same, evaluation indicators are required as a measurement standard. In this work, three evaluation indicators have been used for HSI classification: the kappa coefficient, the average accuracy (AA), and the overall accuracy (OA).

4.1.1. OA. This evaluation index is expressed as the ratio of the number of correctly classified pixels to the total number of marked pixels. If n represents the number of categories of the image feature objects, N_i represents the number of pixels in the i th category, and h_{ii} represents the number of pixels that are correctly classified in the i th category, then the OA is expressed as

$$OA = \frac{\sum_{i=1}^n h_{ii}}{\sum_{i=1}^n N_i}. \quad (13)$$

4.1.2. AA. This evaluation index is defined as follows. The ratio of the number of correctly classified pixels of each category to the category's total number of pixels is calculated. The ratios of the overall categories are summed. The obtained sum is further divided by the overall categories, thus giving a number, which is the average classification accuracy. If N represents the number of sample pixels that are to be tested in the overall training sample, n represents the number of categories, and h_{ii} represents the number of correctly classified pixels of the i th category, then AA is expressed as

$$AA = \frac{1}{n} \sum_{i=1}^n \frac{h_{ii}}{N_i}. \quad (14)$$

4.1.3. Kappa. This coefficient is a performance index used for measuring the accuracy of classification via the classification confusion matrix and is used for consistency testing. It can be expressed as follows:

$$\text{kappa} = \frac{P_o - P_e}{1 - P_e}. \quad (15)$$

4.2. Datasets. This paper will test the proposed MDCA-CP on three benchmark datasets: IP (Indian Pines), PU (University of Pavia), and SA (Salina) to verify its effectiveness. An AVIRIS sensor over northwestern Indiana took the IP dataset. It comprises $145 \text{ pixels} \times 145 \text{ pixels}$ in the spatial domain. The spectral domain is composed of 224 spectral reflectance bands, and the wavelength range is $0.4\text{--}2.45 \mu\text{m}$. Among them, the available ground truth is classified into 16 categories. In this paper's experiment, 20 water absorption bands are removed, and the final image size is $145 \text{ pixels} \times 145 \text{ pixels} \times 200 \text{ pixels}$. A ROSIS sensor took the PU dataset over northern Italy. It consists of $10 \text{ pixels} \times 340 \text{ pixels}$ in the spatial domain; after removing the water absorption band, the spectral domain consists of 103 spectral bands, and the spectral coverage ranges from 430 to 860 nm. Among them, the ground truth is classified into 9 categories, and the size of the image used in the experiment of this article is $610 \times 340 \times 103$.

The SA dataset was taken over California by the AVIRIS sensor. It has 224 bands in the spectral domain and $512 \text{ pixels} \times 217 \text{ pixels}$ in the spatial domain. It also has a high spatial resolution (3.7 m pixels). **Features.** After removing the 20 absorbent bands, the experimental image's size is $512 \text{ pixels} \times 217 \text{ pixels} \times 204 \text{ pixels}$, and the available ground truth is classified into 16 categories. Tables 1–3 describe the sample numbers of these three datasets in detail.

4.3. Hyperparameter Setting. The MDCA-CP proposed in this paper is based on Python language and Keras deep learning framework. The experimental environment was the Windows 10 operating system, 16 GB RAM, and NVIDIA GeForce GTX 1080 8 GB GPU. In order to prevent the deviation caused by different training samples, the average value of more than 20 experimental results under the same conditions was analyzed in this paper. In this model, the random gradient descent method is adopted to update the weight. The learning rate is 0.01, the proportion of disconnected neurons in the dropout layer of the whole connection is set as 0.5, and the activation function is ReLU. In this paper, the bilateral fusion block network was trained for small-batch gradient descent. The number of training samples was set as 1000, and the EPOCH was set as 600, and Table 4 shows the detailed hyperparameter settings.

4.4. Experimental Results Obtained from Different Methods. To verify the validity and correctness of our proposed method, a comparison of the results obtained using MDCA-CP was done with those obtained using AlexNet [23], ResNet [33], DenseNet [34], PRAN [35], FSSNet [36], and SAGP [37]. In order to ensure fairness in the experiment, all hyperparameters in the comparison network were set to the same value. The results obtained by applying the different models on the three datasets are given in Tables 5–7.

TABLE 1: Subject category distribution of the IP dataset.

Type	Species	Samples
C1	Alfalfa	46
C2	Corn-notill	1428
C3	Corn-mintill	830
C4	Corn	237
C5	Grass-pasture	483
C6	Grass-trees	730
C7	Grass-pasture-mowed	28
C8	Hay-windrowed	478
C9	Oats	20
C10	Soybean-notill	972
C11	Soybean-mintill	2455
C12	Soybean-clean	593
C13	Wheat	205
C14	Woods	1265
C15	Buildings-grass-trees-drives	386
C16	Stone-steel-towers	93

TABLE 2: Subject category distribution of the PU dataset.

Type	Class	Samples
C1	Asphalt	6631
C2	Meadows	18649
C3	Gravel	2099
C4	Trees	3064
C5	Painted metal sheets	1345
C6	Bare soil	5029
C7	Bitumen	1330
C8	Self-blocking bricks	3682
C9	Shadows	947

TABLE 3: Subject category distribution of the SV dataset.

Type	Class	Samples
C1	Brocoli_green_weeds_1	2009
C2	Brocoli_green_weeds_2	3726
C3	Fallow	1976
C4	Fallow_rough_plow	1394
C5	Fallow_smooth	2678
C6	Stubble	3959
C7	Celery	3579
C8	Grapes_untrained	11271
C9	Soil_vinyard_develop	6203
C10	Corn_senesced_green_weeds	3278
C11	Lettuce_romaine_4wk	1068
C12	Lettuce_romaine_5wk	1927
C13	Lettuce_romaine_6wk	916
C14	Lettuce_romaine_7wk	1070
C15	Vinyard_untrained	7268
C16	Vinyard_vertical_trellis	1807

4.4.1. Experimental Results Obtained Using the IP Dataset. We randomly selected 5% of the training samples from the IP dataset for training and the remaining 95% samples for testing. It can be seen from Table 5 and Figure 6 that the AlexNet model has the worst classification performance, with considerable noise in its classification graph. This is

TABLE 4: Experimental parameter settings.

Type	Hyperparameter
Adam	Learning rate = 0.01
	$\beta_1 = 0.9$
	$\beta_2 = 0.999$
	Epsilon = 1e-08
	Decay = 3e-8

because its model is not deep enough to extract representative discriminant features. It fails to make corresponding optimization adjustments for overfitting and resolution decline in the training process. In the other classification models, the model's classification results with the attention mechanism algorithm are better than those obtained by other methods. In addition, compared with PRAN, SAGP has a better classification performance, which is due to the extreme imbalance of category samples in the IP dataset. The attention mechanism overcomes this disadvantage, so the MDCA-CP proposed in this paper achieves the optimal results.

4.4.2. Experimental Results Obtained Using the PU Dataset. From Table 6 and Figure 7, it can be seen that, in the Pavia University dataset, among the accuracy rates obtained by various methods, the MDCA-CP method proposed in this work exhibits the highest classification accuracy. As observed in Figure 7, the method proposed in this work has a larger number of correctly classified pixels in the classification map and has less noise than the other methods in some areas. This is because SAGP contains an attention mechanism network that focuses on the features that play an important role in the current task goal by strengthening the local features and selects the current state from many features. The network model can use the analysis of the key features to replace the overall image features analysis and enhance the weight of the features related to the current task goal. The targeted features will not help the classification as they are weakened. When performing the classification process, attention can be focused on finding the useful features associated with the current output to obtain better feature representations for different categories, which helps the final classification performance. From the above description, we can see that the method proposed in this study is better than the other methods.

4.4.3. Experimental Results Obtained Using the SV Dataset. For the SV dataset, this paper randomly selects 1%, 5%, and 10% of training samples, and the remaining samples are used as test samples. Table 7 and Figure 8 show the comparative experimental results. Similarly, in terms of visual effects, the ground object classification map displayed on the SV dataset by the method proposed in this paper has the slightest noise. In quantitative analysis, the OA on the SA data set of the MDCA-CP model reaches 96.15%, which is higher than other comparison methods on OA, KA, and AA. The proposed MDCA-CP model can effectively identify most of the ground features and perform better in the SV dataset.

TABLE 5: Classification results of different methods on the IP data (%).

Methods	5%			10%			15%		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
AlexNet [23]	68.69	56.29	64.05	74.32	65.17	70.94	81.76	79.44	79.70
ResNet [33]	70.67	69.69	66.04	78.40	79.22	75.10	83.33	80.76	80.95
DenseNet [34]	71.23	67.55	66.66	78.37	75.29	75.87	84.66	81.37	82.58
PRAN [35]	72.45	73.37	69.48	77.66	73.28	74.47	82.72	76.51	80.28
FSSFNet [36]	73.75	67.95	69.98	78.66	71.26	75.47	82.61	74.48	80.05
SAGP [37]	73.49	76.58	73.61	78.36	80.89	76.72	81.59	86.50	82.89
MDCA-CP	76.85	79.11	73.72	82.56	81.87	80.03	87.32	87.65	85.55

TABLE 6: Classification results of different methods on the PU data (%).

Methods	1%			5%			10%		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
AlexNet [23]	87.15	84.24	82.66	92.81	91.77	90.43	92.85	93.46	91.32
ResNet [33]	84.06	82.38	77.59	90.81	91.32	87.76	94.15	94.78	93.19
DenseNet [34]	82.32	80.61	75.69	90.25	89.49	87.08	92.27	93.46	91.21
PRAN [35]	89.27	88.61	85.63	93.45	92.32	91.48	94.71	93.55	93.02
FSSFNet [36]	86.63	84.23	82.11	93.72	92.35	91.65	94.37	92.65	92.52
SAGP [37]	84.32	83.84	78.83	91.09	89.75	88.14	93.73	92.97	91.65
MDCA-CP	89.40	87.97	86.81	93.55	91.91	92.21	94.99	93.02	93.11

TABLE 7: Classification results of different methods on the SV data (%).

Methods	1%			5%			10%		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
AlexNet [23]	90.88	93.72	89.77	94.05	96.64	93.42	94.26	95.67	95.29
ResNet [33]	87.21	92.07	85.88	91.70	95.15	90.74	93.52	96.73	93.11
DenseNet [34]	85.52	90.81	84.12	91.03	94.36	90.05	93.14	96.12	92.17
PRAN [35]	78.86	76.17	76.49	90.28	88.17	89.16	91.81	89.32	91.17
FSSFNet [36]	91.04	94.74	90.02	93.56	96.69	92.82	95.85	98.03	95.37
SAGP [37]	90.87	94.68	89.82	92.64	95.69	91.81	94.76	97.45	94.16
MDCA-CP	91.45	94.84	89.98	94.51	97.11	93.24	96.15	98.11	96.21

FIGURE 6: Classification results on the IP dataset using 5% training samples. (a) Ground truth. (b) AlexNet. (c) ResNet. (d) DenseNet. (e) PRAN. (f) FSSFNet. (g) SAGP. (h) MDCA-CP.

FIGURE 7: Classification results on the PU dataset using 1% training samples. (a) Ground truth. (b) AlexNet. (c) ResNet. (d) DenseNet. (e) PRAN. (f) FSSFNet. (g) SAGP. (h) MDCA-CP.

4.5. Results of Ablation Experiments on Different Weighting Strategies. Table 8 shows the experimental results obtained by using different weight allocation strategies. From the table, it can be seen that the combined addition and maximization strategy is significantly better than the individual strategies of addition only or maximization only. The combined strategy enhances the model classification ability, proving that the two strategies of MDCA-CP are effective. Further, the weight addition strategy is superior to the weight maximization strategy, thus proving that only using the maximization strategy will lead to a loss of some features. Figures 9–11 also show the limit pie chart of each indicator on the three datasets with different weighting strategies.

4.6. Results of Ablation Experiments on Different Submodules. This section presents the results of the ablation experiments on the cross-attention module and the CP module. Table 9 and Figure 12 show the corresponding experimental results from which it is observed that a combination of the two models (MDCA-CP) provides superior results. Similarly, we found that the cross-attention module has better performance than the covariance pooling module, which further shows the importance of feature mining in hyperspectral image classification. The MDCA module can effectively select the most representative features. In contrast, the covariance module is better than the baseline model, proving that data mining from HSIs and extracting second-order

FIGURE 8: Classification results on the SV dataset using 1% training samples. (a) Ground truth. (b) AlexNet. (c) ResNet. (d) DenseNet. (e) PRAN. (f) FSSFNet. (g) SAGP. (h) MDCA-CP.

TABLE 8: Results of ablation experiments on different weighting strategies (%).

Modules	IP			PU			SV		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
Add	74.77	75.48	71.01	92.02	90.58	89.20	92.48	94.85	92.99
Max	74.12	74.36	70.25	91.58	89.74	89.02	91.52	94.44	91.85
Add + Max (MDCA-CP)	76.85	79.11	73.72	93.55	91.91	92.21	94.51	97.11	93.24

FIGURE 9: Classification results on the IP dataset using different weighting strategies. (a) OA. (b) AA. (c) Kappa.

FIGURE 10: Classification results on the PU dataset using different weighting strategies. (a) OA. (b) AA. (c) Kappa.

FIGURE 11: Classification results on the SV dataset using different weighting strategies. (a) OA. (b) AA. (c) Kappa.

TABLE 9: Results of ablation experiments on different submodules (%).

Modules	IP			PU			SV		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
No-CrossAttention	74.78	76.54	72.24	90.81	89.17	90.10	91.78	94.25	91.36
No-CovariancePooling	76.13	71.13	71.77	93.25	90.65	91.11	93.89	95.25	92.15
MDCA-CP	76.85	79.11	73.72	93.55	91.91	92.21	94.51	97.11	93.24

FIGURE 12: Classification results on the IP dataset using different submodules. (a) Ground truth. (b) No MDCA. (c) No CP. (d) MDCA-CP.

FIGURE 13: Classification results on the IP dataset using different submodules.

FIGURE 15: Classification results on the SV dataset using different submodules.

FIGURE 14: Classification results on the PU dataset using different submodules.

features are also effective. This further proves the superiority of our MDCA-CP model. Figures 13–15 also show the histograms of the indicators of the different submodules on the three datasets.

5. Conclusion

This paper presents the data dimensionality reduction and feature extraction and presents the novel MDCA-CP model for HSI scene classification. Multiscale convolution has been proven to effectively detect the subtle changes between the pixels in the local area of spatial and spectral dimensions of an HSI, and it can be applied to the feature extraction of hyperspectral data of complex types and structures. Traditional algorithms assign the attention weights in a one-way manner, thus resulting in the loss of feature information. This work's dense cross-attention mechanism can jointly distribute the attention weights horizontally and vertically. As a result, it can efficiently mine the most representative features of the hyperspectral data. A further enhancement in

mining the features can be achieved by employing the weight addition and maximization strategies. This paper also proves the effectiveness and superiority of the MDCA-CP algorithm by comparing results obtained from different models and conducting ablation experiments.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] T. Adão, J. Hruška, L. Pádua et al., "Hyperspectral imaging: a review on UAV-based sensors, data processing and applications for agriculture and forestry," *Remote Sensing*, vol. 9, no. 11, Article ID 1110, 2017.
- [2] T.-H. Hsieh and J.-F. Kiang, "Comparison of CNN algorithms on hyperspectral image classification in agricultural lands," *Sensors*, vol. 20, no. 6, Article ID 1734, 2020.
- [3] B. Lu, P. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sensing*, vol. 12, no. 16, Article ID 2659, 2020.
- [4] S. Jay, M. Guillaume, A. Minghelli et al., "Hyperspectral remote sensing of shallow waters: considering environmental noise and bottom intra-class variability for modeling and inversion of water reflectance," *Remote Sensing of Environment*, vol. 200, pp. 352–367, 2017.
- [5] A. Song and Y. Kim, "Deep learning-based hyperspectral image classification with application to environmental geographic information systems," *Korean Journal of Remote Sensing*, vol. 33, no. 6_2, pp. 1061–1073, 2017.
- [6] C. Weber, R. Aguejidad, X. Briottet et al., "Hyperspectral imagery for environmental urban planning," in *Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1628–1631, IEEE, Valencia, Spain, July 2018.
- [7] X. Dong, J. Dong, A. K. Yetisen et al., "Characterization and layer thickness mapping of two-dimensional MoS₂ flakes via hyperspectral line-scanning microscopy," *Applied Physics Express*, vol. 12, no. 10, Article ID 102004, 2019.
- [8] L. Windrim, R. Ramakrishnan, A. Melkumyan, and R. J. Murphy, "A physics-based deep learning approach to shadow invariant representations of hyperspectral images," *Institute of Electrical and Electronics Engineers Transactions on Image Processing*, vol. 27, no. 2, pp. 665–677, 2017.
- [9] A. Ibrahim, B. Franz, Z. Ahmad et al., "Atmospheric correction for hyperspectral ocean color retrieval with application to the Hyperspectral Imager for the Coastal Ocean (HICO)," *Remote Sensing of Environment*, vol. 204, pp. 60–75, 2018.
- [10] S. P. Garaba, J. Aitken, B. Slat et al., "Sensing ocean plastics with an airborne hyperspectral shortwave infrared imager," *Environmental Science & Technology*, vol. 52, no. 20, pp. 11699–11707, 2018.
- [11] S. Serranti, R. Palmieri, G. Bonifazi, and A. Cózar, "Characterization of microplastic litter from oceans by an innovative approach based on hyperspectral imaging," *Waste Management*, vol. 76, pp. 117–125, 2018.
- [12] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *Institute of Electrical and Electronics Engineers Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4085–4098, 2010.
- [13] L. Samaniego, A. Bárdossy, and K. Schulz, "Supervised classification of remotely sensed imagery using a modified k -NN technique," *Institute of Electrical and Electronics Engineers Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2112–2125, 2008.
- [14] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *Institute of Electrical and Electronics Engineers Transactions on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.
- [15] S. Jia, G. Tang, J. Zhu, and Q. Li, "A novel ranking-based clustering approach for hyperspectral band selection," *Institute of Electrical and Electronics Engineers Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 88–102, 2015.
- [16] G. Licciardi, P. R. Marpu, J. Chanussot, and J. A. Benediktsson, "Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles," *Institute of Electrical and Electronics Engineers Geoscience and Remote Sensing Letters*, vol. 9, no. 3, pp. 447–451, 2011.
- [17] P. Chen, J. D. Nelson, and J. Y. Tournier, "Toward a sparse Bayesian Markov random field approach to hyperspectral unmixing and classification," *Institute of Electrical and Electronics Engineers Transactions on Image Processing*, vol. 26, no. 1, pp. 426–438, 2016.
- [18] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *Institute of Electrical and Electronics Engineers Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.
- [19] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 4959–4962, IEEE, Milano, Italy, July 2015.
- [20] F. Tsai and J.-S. Lai, "Feature extraction of hyperspectral image cubes using three-dimensional gray-level cooccurrence," *Institute of Electrical and Electronics Engineers Transactions on Geoscience and Remote Sensing*, vol. 51, no. 6, pp. 3504–3513, 2013.
- [21] M. Weeks and M. A. Bayoumi, "Three-dimensional discrete wavelet transform architectures," *Institute of Electrical and Electronics Engineers Transactions on Signal Processing*, vol. 50, no. 8, pp. 2050–2063, 2002.
- [22] H. Luo, Y. Y. Tang, X. Yang, L. Yang, and H. Li, "Autoencoder with extended morphological profile for hyperspectral image classification," in *Proceedings of the 2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*, pp. 1–4, IEEE, Exeter, UK, June 2017.
- [23] C. Zhao, X. Wan, G. Zhao, and Y. Yan, "Spectral-spatial classification of hyperspectral images using trilateral filter and stacked sparse autoencoder," *Journal of Applied Remote Sensing*, vol. 11, no. 1, Article ID 016033, 2017.
- [24] S. Hemissi and I. R. Farah, "Efficient multi-temporal hyperspectral signatures classification using a Gaussian-Bernoulli RBM based approach," *Pattern Recognition and Image Analysis*, vol. 26, no. 1, pp. 190–196, 2016.

- [25] Z. Zhao, T. Xu, and C. Dai, "Classifying images using restricted Boltzmann machines and convolutional neural networks," in *Proceedings of the Ninth International Conference on Digital Image Processing (ICDIP 2017)*, vol. 10420, July 2017, Article ID 104202U.
- [26] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *Institute of Electrical and Electronics Engineers Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 937–949, 2017.
- [27] Y. Luo, J. Zou, C. Yao et al., "HSI-CNN: a novel convolution neural network for hyperspectral image," in *Proceedings of the 2018 International Conference on Audio, Language and Image Processing (ICALIP)*, Shanghai, China, July 2018.
- [28] A. Wang, Y. Wang, and Y. Chen, "Hyperspectral image classification based on convolutional neural network and random forest," *Remote Sensing Letters*, vol. 10, no. 11, pp. 1086–1094, 2019.
- [29] X. Mei, E. Pan, Y. Ma et al., "Spectral-spatial attention networks for hyperspectral image classification," *Remote Sensing*, vol. 11, no. 8, p. 963, 2019.
- [30] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sensing*, vol. 11, no. 11, p. 1307, 2019.
- [31] W. Cai, B. Liu, Z. Wei, M. Li, and J. Kan, "TARDB-Net: triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 11291–11312, 2021.
- [32] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *Institute of Electrical and Electronics Engineers Geoscience and Remote Sensing Letters*, vol. 2020, p. 5, 2020.
- [33] M. Salman and S. E. Yüksel, "Fusion of hyperspectral image and LiDAR data and classification using deep convolutional neural networks," in *Proceedings of the 2018 26th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, Izmir, Turkey, July 2018.
- [34] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral-spatial hyperspectral image classification," *Institute of Electrical and Electronics Engineers Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 740–754, 2019.
- [35] G. Yang, U. B. Gewali, E. Ientilucci, M. Gartley, and S. T. Monteiro, "Dual-channel densenet for hyperspectral image classification," in *Proceedings of the IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2595–2598, IEEE, Valencia, Spain, July 2018.
- [36] H. Gao, Y. Yang, D. Yao, and C. Li, "Hyperspectral image classification with pre-activation residual attention network," *Institute of Electrical and Electronics Engineers Access*, vol. 7, pp. 176587–176599, 2019.
- [37] H. You, S. Tian, L. Yu, and Y. Lv, "Pixel-level remote sensing image recognition based on bidirectional word vectors," *Institute of Electrical and Electronics Engineers Transactions on Geoscience and Remote Sensing*, vol. 58, no. 2, pp. 1281–1293, 2019.