IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 May 2024, accepted 30 June 2024, date of publication 8 July 2024, date of current version 16 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3424674

RESEARCH ARTICLE

Hyperspectral Image Classification Using Attention-Only Spatial-Spectral Network Based on Transformer

WEIYI LIAO, FENGSHAN WANG, AND HUACHEN ZHAO
College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China

Corresponding author: Weiyi Liao (lgdxlwy@163.com)

This work was supported by the National Social Science Fund of China under Grant 2022-SKJJ-C-007.

ABSTRACT Hyperspectral image (HSI) classification has drawn increasing attention in the last decade. HSIs accurately classify terrestrial objects by capturing approximately contiguous spectral information. Owing to their excellent performance in image classification and semantic segmentation, many of the latest deep learning approaches, which can extract complex spatial and spectral characteristics compared to traditional machine learning methods, have been applied in HSI classification. The paper proposes a new HSI classification network based on pure multihead attention mechanisms based on a vision transformer. Due to the unique spatial and spectral attention modules, the network can derive long-range spatial and spectral contextual relations between pixels in images. The spatial and spectral features are effectively fused and interacted through the cross-field gating module. The paper evaluates the classification performance of the proposed network on three HSI datasets by conducting extensive experiments, showing its superiority over standard convolutional neural networks and achieving a significant improvement in comparison with other networks. In addition, due to the complete abandonment of the convolution layer and the application of multihead attention mechanisms, the number of parameters of the network is greatly reduced.

INDEX TERMS Hyperspectral image classification, remote sensing, multihead attention mechanism, vision transformer.

I. INTRODUCTION

Hyperspectral images (HSIs) are composed of hundreds of approximately contiguous wavelength bands and play an important role in remote sensing. Compared to traditional images, HSIs can provide rich spatial and spectral information simultaneously from the same area on the surface of the earth. The values of each pixel can be regarded as a high-dimensional vector whose entries correspond to the spectral reflectance at a specific wavelength. Due to the rich spatial and spectral information, HSIs can distinguish subtle differences between similar terrestrial objects and achieve excellent performance. With their advantages in the classification of objects, HSIs have been widely applied in many fields [1], [2], [3].

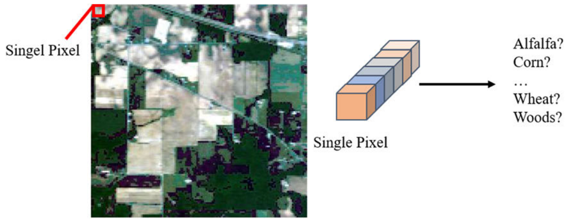


FIGURE 1. Hyperspectral images classification is to assign a single pixel to a certain class instead of the whole image.

HSI classification assigns each pixel to a certain class based on its spatial and spectral characteristics, as shown in Fig. 1. In the early stages of HSI classification research, conventional image classification methods, such as support vector machines (SVMs) and neural networks, which are widely used in computer vision, were applied in the remote

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang.

sensing field. Melgani and Bruzzone assessed the potential of SVM classifiers in HSI classification [4]. Farag et al. used support vector machines (SVMs) to conduct density estimation based on class conditional probability and increased the accuracy by up to 10% [5].

In [6], Jimenez et al. introduced a neural network to conduct feature fusion and decision making based on project pursuit and majority voting. Benediktsson and Kanellopoulos proposed multisource classification methods based on neural networks and achieved high accuracies compared with conventional classification schemes [7].

Feature-extraction or dimension-reduction techniques are also applied in HSI classification. In [8], Chang and Du used ratio-based principal component analysis (SINR-PCA) and interference-annotated noise-whitened principal component analysis (IANW-PCA) to adequately represent image quality. Experiments showed that modified PCA improved the estimation of the noise covariance matrix. To capture all the discriminant information of hyperspectral images, Chulhee Lee modified PCA with pre-encoding discriminant information. The proposed method improved the classification accuracy compared to that of conventional compression methods [9].

With the development of deep learning, deep learning has demonstrated superior performance in traditional computer vision fields, including image classification [10], object localization [11], semantic segmentation [12], and instance segmentation [13]. Lecun et al. proposed LeNet-5, which applied a back-propagation algorithm and convolutional neural network (CNN) [14], to provide record accuracy on business and personal chipes. LeNet-5 is considered the prototype of modern neural networks. However, its performance in real-world tasks cannot compete with that of traditional algorithms such as SVM and boosting. In 2012, Krizhevsky et al. improved LeNet-5 and proposed AlexNet [15]. AlexNet achieved a 15.3% error rate in the ILSVRC-2012 competition and far exceeded the second place. Compared to LeNet-5, AlexNet has introduced many new novelties. First, AlexNet uses a rectified linear unit as an activation function instead of a sigma0d function. Second, a dropout layer is used to mitigate the overfitting problem. In addition, AlexNet also uses overlapping max pooling layers and data augmentation. Due to these novelties, AlexNet obtained great results and is considered the earliest deep learning model. Since the success of AlexNet, many advanced networks based on CNNs have been developed and achieved great success. VGGNet employs 3 × 3 filters whose stride is one pixel [16]. Small filters can reduce the number of weights and the training complexity. GoogLeNet employs an inception module and auxiliary classification to improve the results of image classification [17]. In the wake of great success in the employment of CNNs, an increasing number of effective networks, such as ResNet [18], MobileNet [19], and EfficientNet [20], have emerged and achieved great success.

In addition to traditional CNN-based methods, several classical neural networks that have been previously applied in natural language processing, such as recurrent neural networks (RNNs) [21] and long short-term memory (LSTM) [22], have also been employed in computer vision fields. In 2017, Google proposed a new network architecture called the transformer architecture, which is based on attention mechanisms and entirely abandons recurrence and convolutions. The experiments showed that the transformer required less time to train and improved upon the existing best results. Inspired by the great success of the transformer, Dosovitskiy proposed the vision transformer (ViT), which is also based solely on attention mechanisms [23]. Vision transformers have achieved excellent results compared to former convolutional networks. ViT showed the great potential of attention mechanisms in computer vision fields and proved that CNNs are not necessary. Due to the great success of neural networks in computer vision fields, deep learning methods have been considered alternatives in HSI classification. Hu et al. proposed a unidimensional CNN based on individual spectra and achieved better classification performance than traditional deep learning methods [24]. To avoid information loss in representing hyperspectral pixels, L Mou first proposed a novel RNN model to analyze HSIs as sequential data. The experimental results demonstrated that the proposed RNN model can efficiently process hyperspectral data [25]. Given that HSI data are presented in the format of 3D cubes, Li designed a 3D convolutional neural network to extract spatial-spectral features [26]. Compared to other deep learning-based methods, the experiments showed that the 3D-CNN-based model outperformed the previous methods. Due to the high dimensions, Sigirci and Bilgin introduced BERT-based (bidirectional encoder representations from transformers) models for classification, which have been widely applied in natural language processing [27]. The model can also accept spatial features because of its structure. Experiments showed that the BERT-based models outperformed conventional 1D/2D convolutional models.

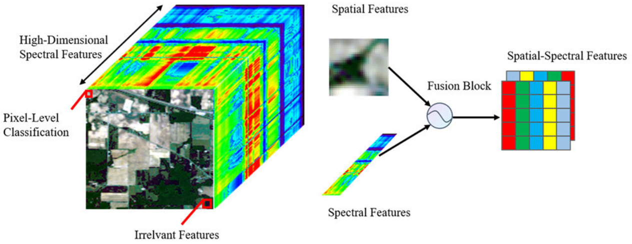


FIGURE 2. Specificities of hyperspectral images classification. Compared with typical image classification tasks, HSIs classification is more concerned with how to handle high-dimensional spectral features and fuse the different dimensional features.

Compared to typical image classification tasks, HSI classification has the following main differences, as shown in Fig. 2:

1) Pixel-level classification: HSI classification aims to label each pixel in remote sensing images with a semantic class. Sometimes pixel-level classification is also called semantic segmentation. However, typical image classification tasks seek to classify each image into a semantic class. Due to the pixel-level classification, traditional image classification models, such as ResNet and EfficientNet, are not suitable for HSI classification.
2) High-dimensional spectral features: Each pixel in HSIs has hundreds of contiguous wave-length bands, whereas typical images generally have three bands. Traditional CNNs can hardly capture the potential features among continuous spectral bands. In addition, different orders of magnitude of spectral bands cause HSI classification models to be unable to use pretrained models in other image classification datasets, such as ImageNet or CoCo, and must train the whole model from the beginning.
3) Spatial and spectral feature fusion: Different locations and contiguous wavelength bands of each pixel provide rich spatial and spectral features. To carry out HSI classification tasks effectively, it is necessary to integrate spatial and spectral features. Apart from this, the different features cannot equally contribute to the task. Due to their different importance, useful spatial and spectral features should be emphasized compared to other features.

To address these differences, inspired by the great success of the attention mechanism, this paper proposes a spatial-spectral pure attention network (SSPAN), which is totally composed of attention mechanism and aborts convolutional structures. The attention mechanism mimics human visual and cognitive processes by dynamically and selectively focusing on one part of the spectral or spatial features while ignoring the rest, helping the network to focus on the significant parts of band or neighborhoods when processing the HSI classification.

In terms of pixel-level classification, this paper takes the neighborhood block of each pixel as a 3D cube to capture the spatial features. The proposed spatial attention module is more concerned with adjacent and significant pixels and ignores distant and redundant. In terms of high-dimensional spectral features, a global spectral attention module can capture the contiguous wavelength band features and highlight important spectral band features. These modules effectively capture the key spatial and spectral features and reduce redundant information and computations to accelerate the model training process. The paper designs a fusion gating module to integrate the different dimension features, which emphasizes significant and suppresses unnecessary features. This paper validates the proposed SSPAN model on three HSI datasets. The experimental results demonstrate that SSPAN can achieve superior classification results.

The main contributions of the paper can be concluded as follows:

1) The spectral features of the contiguous wavelength bands in HSIs are considered using a pure attention mechanism, which can capture the partial and global correlations simultaneously for HSI classification.
2) A 3D cube mechanism and neighborhood crossover attention mechanism for HSIs are designed to emphasize the adjacent and effective pixels and weaken the distant and redundant pixels in the spatial context, which can reduce the computation and model training time.
3) A joint fusion block is proposed to integrate the captured spatial and spectral features, which can fully utilize the spatial-spectral contexts for HSI classification.

The remainder of the paper is organized as follows. Section II introduces the proposed SSPAN model for HSI classification in detail. Section III describes the datasets and experimental results. Finally, Section IV concludes the paper.

II. PROPOSED METHOD

A. ARCHITECTURE OF SSPAN

As shown in Fig. 3, the proposed SSPAN model has three main components: a spectral attention module, a spatial attention module, and a cross-field adaptive gating module. This paper will illustrate the parts in the following sections in detail.

B. SPECTRAL ATTENTION MODULE

CNN-based methods have been proposed for traditional image classification tasks. Compared to other conventional image classification methods, CNNs can effectively capture localized spatial and spectral information via different convolution layers. Among the layers, trainable kernels play a vital role. The trainable kernels are designed to be of a specific size. Kernels of different sizes scan across the whole image and extract the local features. Different convolutional layers can extract different levels of image features. Due to its proximity to the input layer, the shallow convolutional layer can extract simple features such as color and texture. The deeper convolutional layer can capture abstract and high-level features.

Despite the great success of convolutional neural networks in traditional image classification, there are many problems with the application of convolutional structures to HSI classification. The first problem is the large amount of computation. Convolutional neural networks use a convolutional kernel to slide the image to obtain local features of the image. In traditional computer vision tasks, the size of the convolution kernel is typically 3 × 3, 5 × 5, or 7 × 7, and the number of channels usually corresponds to the input image. For a three-channel RGB image, the image is compressed into a single-channel image after a convolution kernel calculation.

The compressed image contains all the channel information from the original image. In a convolutional layer, multiple convolutional kernels are generally used for computation to obtain a multichannel image.

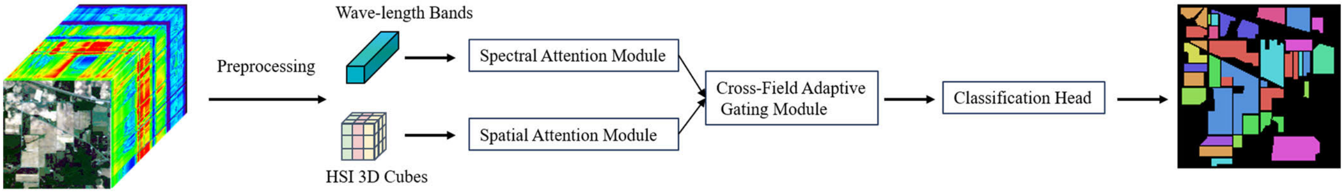


FIGURE 3. Overview of the proposed SSPAN. After preprocessing, HSIs first pass through spectral and spatial attention modules in parallel. Then the cross-field adaptive gating module fuses the features. Finally, the classification head generates classification maps by exploiting fused features.

When utilizing a convolutional neural network for HSI classification, the channels of the convolution kernel match the channels of the HSIs. This leads to the depth of each convolutional kernel reaching hundreds. Excessive depth leads to a significant increase in computation, especially when multiple convolution kernels are used. In addition, due to the characteristics of image convolution calculations, all the channels of HSIs are compressed to one after convolution calculations. The calculation results in the loss of a massive amount of spectral information. It is also difficult for convolutional neural networks to capture the dependence between arbitrary bands and global features.

To address these problems and extract hyperspectral information effectively, a novel spectral module is necessary. The bands of each pixel in HSIs can be considered sequence data. The spectral information of a single pixel consists of a succession of bands, which is fully consistent with the characterization of sequence data. There are many ways to process sequence data, such as RNN, GRU, LSTM, etc. However, these methods make it difficult to handle the information of HSIs and easily lead to the problem of vanishing gradients. Therefore, this paper proposes a spectral attention module for spectral feature extraction.

Attention mechanism can improve the expressive ability of the network. It can significantly reduce the computation time because of parallel computing. In this paper, a spectral module is designed and conducted to selectively emphasize significant spectral information and ignore the redundant, thereby avoiding gradient disappearance or explosion. ViT references the structure of transformer and converts image data to sequence data, which consists of embedding layer, transformer encoder and multi-layer perceptron (MLP) head. Using the structure of the ViT as a reference, a novel spectral attention module is designed to calculate the dependence between arbitrary positional bands and capture the global spectral features, as shown in Fig. 4.

The attention mechanism is generally modeled as follows:

$$O = X \otimes F(X)$$

X is the input, and **F**(·) is the weight. "$\bigotimes$" and **O** represent the elementwise product and the output, respectively. The weight **F**(·) is calculated by specific attention modules, such as squeeze-and-excitation [28], convolutional block attention [29], and efficient channel attention [30]. In the spectral attention module, multihead attention [31] is used, where multiple self-attention (SA) layers are stacked and integrated. Compared to other attention mechanisms, the SA mechanism is better at capturing the global spectral features of a single pixel band. The SA mechanism can be calculated according to the following four steps.

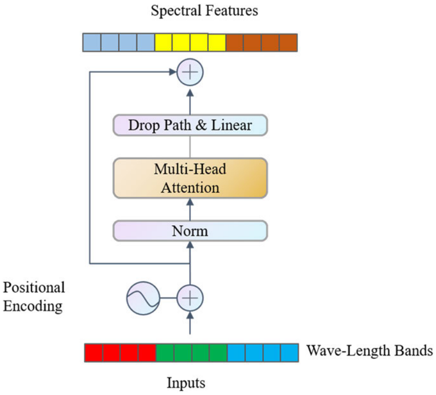


FIGURE 4. Structure of spectral attention module. The position encoding contains learnable parameters with the same size as wave-length bands. The norm layer is batch normalization. The spectral features are obtained by the skip connection after the drop path and linear layer.

Step 1: The sequence data x with a length of m, where $x_i, i = 1, \ldots, m$, are multiplied by a shared matrix W to obtain the feature embedding, denoted as a_i.

Step 2: Each feature embedding is multiplied by three different matrices W_q, W_k, W_v to obtain three vectors, i.e., query ($Q = [q_1, .., q_m]$), key ($K = [k_1, .., k_m]$), and value ($V = [v_1, .., v_m]$).

Step 3: Q and K are multiplied in the form of an inner product to compute the attention score s, e.g., $q_i \cdot k_j$. Then, the scaled score is obtained by normalization to stabilize the gradients, i.e., $s_{i,j} = q_i \cdot k_j/\sqrt{d}$, where d is the dimension of q_i or k_j.

Step 4: The attention score is obtained by the softmax activation function to generate the attention weights $z = [z_1, \ldots, z_m]$.

In summary, the SA module can be integrally calculated as follows:

$$z = \text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

To make full use of the sequence information, the feature embedding is formulated as$a_i + e_i$, where e_i denotes a manual positional vector and can be optimized by the whole network.

This paper proposes a novel and generic spectral feature extraction module (i.e., the spectral attention module). To focus on spectrometric characteristics, the proposed spectral attention module contains multihead attention layers. As shown in Fig. 5, it contains multiple different SA modules and can extract spectral features from multiple perspectives. To improve the detail-capturing capacity of subtle spectral discrepancies, the module also applied the SE mechanism.

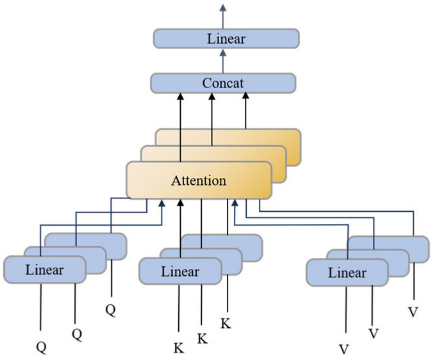


FIGURE 5. **Structure of multi-head attention mechanism.**

C. SPATIAL ATTENTION MODULE

Classic image classification or segmentation networks, including ViT, widely apply convolutional layers. This is because the calculation method of convolution layers can effectively capture the local features, and the deeper convolution layers can capture the global features.

For HSIs, the application of convolutional layers has several problems. The first problem is the redundancy of partial local features. Shallow convolutional layers take advantage of sliding computations to capture all local features. For traditional image classification tasks, these local features are essential for the final image classification tasks. In semantic segmentation tasks, distant local features also play an important role in single-pixel classification due to the small scene. In HSI classification tasks, the final goal is the classification of a single pixel. HSIs typically encompass vast areas, often spanning tens of square kilometers in size. Distant local features could be tens of kilometers away from the pixels to be classified. These local features are redundant and do not contribute to the classification task, thereby increasing the complexity and computational requirements of the model.

To avoid the redundancy and computational complexity triggered by distant local features, this paper proposes a spatial attention module based on HSI 3D cubes. The cubes consist of the neighborhood of the pixels to be classified. The specific size of the neighborhood depends on the resolution of the HSIs (the sizes of the cubes are $h * w * d_m$). Different from other methods for constructing HSI 3D cubes, the proposed method constructs a global multihead attention mechanism to extract spatial features, while other methods still use CNNs [32]. To improve the detail-capturing capacity of subtle spatial discrepancies, the module also applied a partial SE mechanism.

Similar to the spectral attention module, the depth of each convolutional kernel will reach hundreds, leading to a large amount of computation, despite calculations in HSI 3D cubes. In addition, CNNs will also result in the loss of spectral information of the neighborhood of the pixels to be classified.

The proposed spatial attention module calculates the attention representations $o_{i,j}$ between all the pixels in the cubes and the pixel to be classified. Different from the SA mechanism applied in the spectral attention module, the spatial attention module applies the neighborhood crossover attention mechanism, and the formulation can be concluded as follows:

$$o_{i,j} = \text{Multi-head Attention}(x_{i,j}, x_{\text{center pixel}})$$
$$x_{i,j} \text{ represents the vector with coordinates } (x, y).$$

$x_{\text{center pixel}}$ denotes the pixel to be classified.

The attention representations form a spatial attention representation cube that contains all neighborhood spatial features. Fig. 6. illustrates an overview of the proposed spatial attention module in the whole network.

D. CROSS-FIELD ADAPTIVE GATING MODULE

After the spectral and spatial module, spectral and spatial features are captured. However, the features are different in the performance dimension. Spectral features target wavelength bands and are one-dimensional. The spatial features target the pixel and its surrounding neighborhood and are two-dimensional features. The difference in dimensions makes it difficult to fuse and interact information between two features, which also makes it difficult to filter features effectively.

To enhance the information interaction between the spectral and spatial domains and to remove redundant feature information, this article designs a cross-field adaptive fusion mechanism. This mechanism applies a gating mechanism to fuse spectral and spatial domain features of different dimensions, which can remove redundant feature information and enhance the feature expression capability.

To fuse the spectral and spatial features, the gating module diffuses the spectral features and aligns them with the spatial features dimensionally. Let $z \in \mathbb{R}^{1*1*d_m}$ and $o_{h*w} \in \mathbb{R}^{h*w*d_m}$ represent the outputs of the spectral attention module and spatial attention module, respectively. The cross-field adaptive gating module can be expressed by

$$z \in \mathbb{R}^{1*1*d_m} \xrightarrow{\text{diffusion}} \hat{z} \in \mathbb{R}^{h*w*d_m}$$
$$\begin{bmatrix} \hat{z} \\ o_{h*w} \end{bmatrix} [\ddot{w}] \to \hat{o}_{h*w}$$

First, the module diffuses the output of the spectral attention module to the same size as the output of the spatial attention module. The two outputs $\hat{z}$ and o_{h*w} are concatenated and multiply with the weight coefficient $\ddot{w}$. $\ddot{w} \in \mathbb{R}^{d*w*2}$ is the learnable network parameter for adaptive fusion. There are two main reasons for this. The first is to ensure that the two

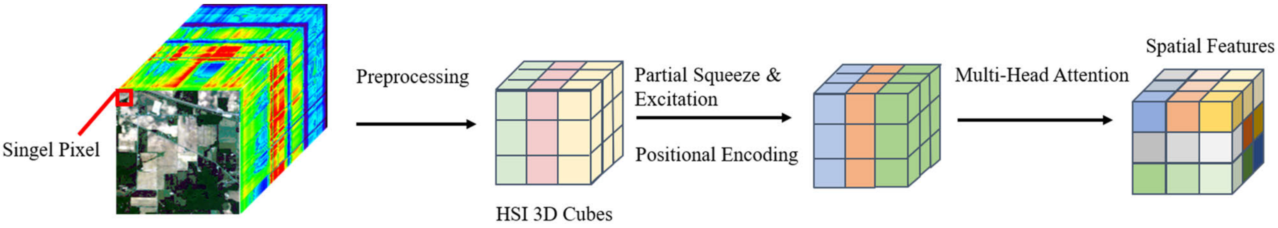


FIGURE 6. Structure of spatial attention module. The preprocessing mainly contains normalization and sampling. HSI 3D cubes are made up of neighborhoods of a particular size around the pixel to be classified. The position encoding contains learnable parameters with the same size as neighborhoods. Partial squeeze and exaction are applied in HSI 3D cubes. The spatial features are obtained by the multi-head attention with 8 heads.

output dimensions are aligned. Second, for adjacent pixels, the different distances result in differences in importance and contributions to the spectral and spatial features of the pixels to be classified. This module can adaptively learn the weights of different adjacent pixel features for the point to be classified with the help of learnable parameters. This also enhances the feature interaction and fusion capabilities of the model. Fig.7 shows an overview of the proposed spatial attention module in the whole network.

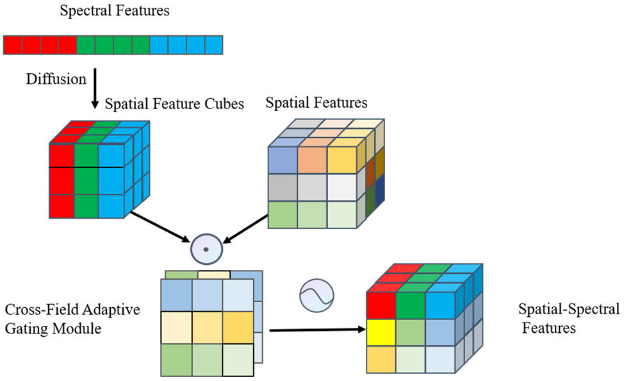


FIGURE 7. Architectural depiction of cross-field adaptive gating module. Spectral features diffuse into feature cubes of the same size as spatial features. The cross-field adaptive module contains learnable weights. The spatial-spectral features are obtained by fusing the spatial and spectral features selectively.

After the above modules, the spectrally sequential information and spatially contextual information are preserved to a great extent. These features will eventually enter the fully connected layer to obtain the final classification result of the pixels.

III. EXPERIMENTAL AND RESULTS

To evaluate the classification performance of the proposed method, the paper compares the SSPAN with other five models over three benchmark HSI datasets.

A. DATASETS

Three benchmark hyperspectral datasets are selected for this experiment, namely, Indian Pines dataset, Pavia University dataset, and Pavia Center dataset. The datasets have different sizes and can be used to verify the model's generalization.

1) Indian Pines dataset: This hyperspectral image dataset was collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines region in northwestern Indiana in 1992. This scene has a size of 145 × 145 and a spatial resolution of 20 meters per pixel. The scene comprises 220 spectral channels with wavelengths ranging from 0.4 to 2.5 μm. After 20 spectral bands are removed due to noise and water absorption phenomena, the remaining 200 radiance channels are used in the experiments. The Indian Pines dataset contains 10249 samples with 16 mutually exclusive ground-truth classes. Fig. 8 (a) shows the ground-truth map available for the scene. Table 1 shows the meaning of each class and the number of samples in each class in the Indian Pines dataset.

TABLE 1. Class name and number of each class for Indian Pines dataset.

NO.	Class	Samples
1	Alfalfa	46
2	Corn-no till	1428
3	Corn-min till	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-no till	972
11	Soybean-min till	2455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1265
15	Buildings-Grass-Trees-Drivers	386
16	Stone-Steel-Towers	93

2) Pavia University dataset. This hyperspectral dataset was collected by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over the urban area of the University of Pavia, Italy. The image size is 610 × 340, with a very high spatial resolution of 1.3 meters per pixel. The number of data channels in this image is 115, with spectra ranging from 0.43 to 0.86 μm. After 12 channels affected by noise are removed, the remaining 103 channels are used in the experiments.

Pavia University dataset contains 42776 samples with 9 classes. Fig. 9 (a) shows the ground-truth map available for the scene. Table 2 shows the meaning of each class and the number of samples in each class in the Pavia University dataset.

3) Pavia Center dataset: This hyperspectral dataset was also collected by the ROSIS optical sensor over Pavia in northern Italy. The image size is 1096 × 715, with a spatial resolution of 1.3 meters per pixel. Similar to Pavia University dataset, Pavia Center dataset has 9 classes and 102 channels. Pavia Center dataset contains 148152 samples with 9 classes. Fig. 10 (a) shows the ground-truth map available for the scene. Table 3 shows the meaning of each class and the number of samples in each class in Pavia Center dataset.

TABLE 2. Class name and number of each class for Pavia University dataset.

NO.	Class	Samples
1	Asphalt	6631
2	Meadows	18649
3	Gravel	2099
4	Trees	3064
5	Painted metal sheets	1345
6	Bare Soil	5029
7	Bitumen	1330
8	Self-Blocking Bricks	3682
9	Shadows	947

TABLE 3. Class name and number of each class for Pavia Center dataset.

NO.	Class	Samples
1	Water	65971
2	Trees	7598
3	Meadows	3090
4	Self-Blocking Bricks	2685
5	Bare Soil	6584
6	Asphalt	9248
7	Bitumen	7287
8	Tiles	42826
9	Shadows	2863

B. EXPERIMENT SETUP

This paper evaluates the classification performance of each model quantitatively in terms of three commonly used indices, overall accuracy (OA), average accuracy (AA), and Kappa coefficient (K ×100). The parameter configurations of these compared methods are detailed as follows.

1) For the SVM, the libsvm toolbox was selected, and the radial basis function (RBF) kernel was used. To obtain the optimal parameters, the model uses a grid search technique and applies an RBF, a linear kernel, and a poly kernel.
2) For the deep CNN (D-CNN), the paper uses a convolution block, which includes a set of 1D convolutional filters. a max-pooling layer and a tanh activation function. A fully connected layer is used for the final classification.
3) For the contextual deep CNN (CD-CNN), the model applies two different sizes of 3D convolution kernels to construct the inception module. Two residual blocks are designed to capture the features at different scales.
4) For the RNN, there is one recurrent layer with a gated recurrent unit (GRU) with 64 neuron units.
5) For the semi-supervised CNN (SS-CNN), the model is composed of an encoder and a decoder module. The encoder module consists of a 2D convolution layer, a max pooling layer, a BN layer and a linear layer. The decoder module is composed of several linear and BN layers.
6) For the proposed SSPAN model, due to the low spatial resolution, in the spatial attention module, the size of the 3D HSI cube is set to 2 × 2. The small size reduces calculations and parameters to accelerate the training process, which enables models to be trained on general workstations. When experiments on high spatial resolution datasets and high-performance workstations, the size can be turned up for better performance. The network consists of one spatial attention module and one spectral attention module in parallel, followed by one cross-field adaptive gating module. The dropout layer is employed after position embeddings, multihead attention, and the final linear layers for inhibiting 20% of neurons. The squeeze-and-excitation mechanism is also applied to enhance the ability to express features.

All the models are implemented on the PyTorch platform using a workstation with an i7-9750H CPU and an NVIDIA GTX 2080Ti GPU. The Adam optimizer with a minibatch size of 64 is adopted. The learning rate is firstly set to 0.01 and the epochs over the three datasets are set to 1000.

C. PERFORMANCE ANALYSIS

1) EXPERIMENTS ON INDIAN PINES DATASET

In the Indian Pines dataset, the paper corrects the dataset by removing a disproportionately small number of categories. Alfalfa, Grass-pasture-mowed, Oats, and Stone-Steel-Towers are removed. Due to uneven data across categories, to ensure the fairness of the experiment, 30% of the data are divided into training samples, and 70% are divided into test samples for each class. The classification results are shown in Table 4. The accuracies are also depicted visually. The classification maps of the compared methods for the Indian Pines dataset are displayed in Fig. 8.

According to Table 4, based on K ×100, OA, and AA, the proposed SSPAN model yields the best classification results. Compared to the traditional SVM method, SSPAN yields accuracy increases of 4.62, 4.02, and 2.91 for K ×100, OA, and AA, respectively. This result exemplifies the great advantage of SSPAN over traditional methods. Compared to the D-CNN and RNN, SSPAN also has considerable advantages. More specifically, the main reason for the poor performances of D-CNN and RNN classification is the low spatial resolution. The image size is 145 × 145, and the spatial resolution

TABLE 4. Comparisons of classification accuracies among different methods in the Indian Pines dataset.

Class	Training-Test Samples	SVM	D-CNN	CD-CNN	RNN	SS-CNN	SSPAN
Corn-notill	428/1000	84.00	75.20	87.50	79.40	89.00	92.30
Corn-mintill	249/581	74.53	62.82	87.26	67.47	75.04	81.76
Corn	71/166	75.30	60.24	81.92	71.08	81.93	81.93
Grass-pasture	144/339	97.05	88.20	96.17	92.63	84.96	95.87
Grass-trees	219/511	99.22	95.30	98.24	97.65	99.80	99.83
Hay-windrowed	143/335	99.40	100	100	100	100	100
Soybean-notill	291/681	81.20	82.97	72.98	73.86	85.02	88.55
Soybean-mintill	736/1719	88.42	82.02	91.33	83.53	88.48	89.18
Soybean-clean	177/416	90.87	80.05	87.74	88.94	82.21	92.06
Wheat	61/144	100	100	100	100	100	100
Woods	379/886	95.37	92.66	96.73	96.50	98.98	98.53
Buildings-Grass-Trees-Drivers	115/271	70.11	55.72	71.22	71.96	55.72	70.48
Kappa×100	/	86.12	79.07	87.91	82.16	86.42	90.74
OA(%)	/	87.91	81.74	89.47	84.48	88.08	91.93
AA(%)	/	87.96	81.23	89.26	85.25	86.76	90.87

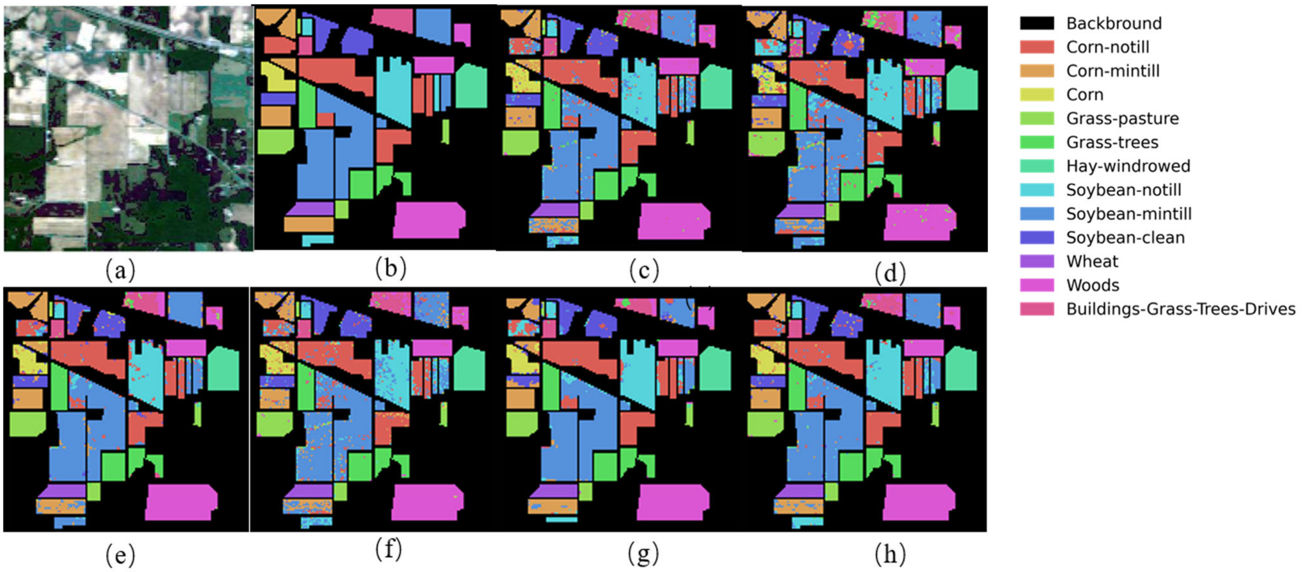


FIGURE 8. Classification maps for the Indian Pines dataset. (a) True-color image. (b) Ground-truth map. (c)-(h) Classification maps of SVM, D-CNN, CD-CNN, RNN SS-CNN and SSPAN.

is 20 meters per pixel. When using a 3 × 3 convolution kernel over the image, it yields a feature map of 60 × 60 meters per pixel, which contains considerable redundant information. This makes it difficult for D-CNN to capture features for classification. For RNNs, the model only considers spectral features and ignores spatial features, resulting in poor classification results.

Relative to the CD-CNN network, SSPAN yields accuracy increases of 2.83, 2.46, and 1.61 for K ×100, OA, and AA, respectively. Relative to the SS-CNN network, the SSPAN network yields accuracy increases of 7.32, 3.85, and 4.11 for K ×100, OA, and AA, respectively. The overall accuracy of the SSPAN model is significantly greater than that of the CD-CNN and SS-CNN models. The most important reason may be the low spatial resolution. The SSPAN model utilizes features only within a small range, which can greatly reduce the number of redundant features and improve the training accuracy. Additionally, the classification accuracy for corn is degraded by the lack of training samples. For Buildings-Grass-Trees-Drivers, the reason for the lower classification accuracy may be due to the complexity of spatial-spectral features along with the small number of training samples.

The accuracies of all models on Indian Pines dataset are not high enough, mainly because of the low spatial resolution and the few training samples. However, the SSPAN model shows great advantages over the other models. This proves that SSPAN can achieve better classification results with fewer training samples over low spatial resolution images. The classification results of the maps in Table 5 qualitatively show the differences and reveal that SSPAN performs better than the other methods.

TABLE 5. Comparisons of classification accuracies among different methods in the Pavia University dataset.

Class	Training-Test Samples	SVM	D-CNN	CD-CNN	RNN	SS-CNN	SSPAN
Asphalt	300/6331	84.85	78.88	88.58	79.51	93.16	92.29
Meadows	300/18349	92.55	86.97	94.50	90.26	89.30	95.96
Gravel	300/1799	84.32	81.93	96.39	70.21	87.27	79.32
Trees	300/2764	97.32	94.65	98.19	95.77	95.48	94.93
Painted metal sheets	300/1045	99.90	99.71	100	99.90	100	100
Bare Soil	300/4729	90.51	72.47	81.52	88.14	97.65	95.45
Bitumen	300/1030	93.40	93.20	95.05	85.34	89.81	95.63
Self-Blocking Bricks	300/3382	86.99	86.84	85.39	81.76	92.40	94.00
Shadows	300/647	100	100	100	99.85	100	100
Kappa×100	/	87.83	79.96	88.99	83.12	89.38	92.58
OA(%)	/	90.92	84.98	91.85	87.35	91.96	94.50
AA(%)	/	92.21	88.30	93.29	87.86	93.90	94.18

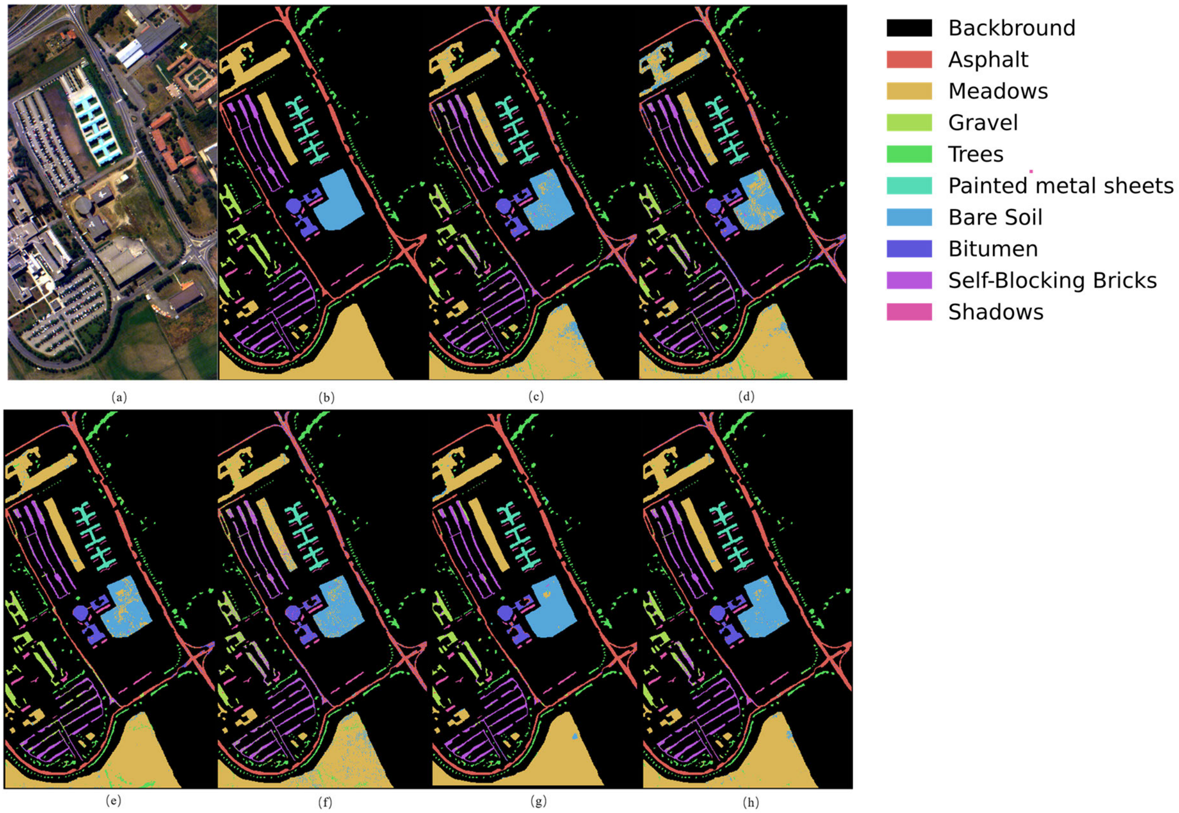


FIGURE 9. Classification maps for the Pavia University dataset. (a) True-color image. (b) Ground-truth map. (c)-(h) Classification maps of SVM, D-CNN, CD-CNN, RNN SS-CNN and SSPAN.

2) EXPERIMENTS ON PAVIA UNIVERSITY DATASET

In the Pavia University dataset, the number of categories was reduced to 9. However, the image size is 4 times larger than that of the Indian Pines dataset. The number of samples is also four times larger. To balance the training samples, 300 samples from each category were randomly selected as training samples, and the remaining samples were used as testing samples. The classification results are shown in Table 5. The accuracies are also depicted visually. The classification maps of the compared methods for the Indian Pines dataset are displayed in Fig. 3.

According to Table 5, based on K ×100, OA, and AA, the proposed SSPAN model yields the best classification results. Compared to the traditional SVM method, SSPAN yields accuracy increases of 4.75, 3.58, and 1.97 for K ×100, OA, and AA, respectively. This result exemplifies the great advantage of SSPAN over traditional methods. Compared to the D-CNN and RNN, SSPAN also has considerable advantages. Relative to the CD-CNN network, SSPAN yields accuracy increases of 3.59, 2.65, and 0.89 for K ×100, OA, and AA, respectively. Relative to the SS-CNN network, the SSPAN network yields accuracy

TABLE 6. Comparisons of classification accuracies among different methods in the Pavia Center dataset.

Class	Training-Test Samples	SVM	D-CNN	CD-CNN	RNN	SS-CNN	SSPAN
Water	500/6541	99.89	99.31	99.70	99.78	98.40	99.69
Trees	500/7098	95.87	89.53	93.04	98.42	96.76	97.30
Meadows	500/2590	97.14	89.23	94.83	69.15	89.81	95.44
Self-Blocking Bricks	500/2185	95.51	86.50	94.83	97.80	98.95	94.23
Bare Soil	500/6084	95.73	89.73	97.98	89.27	99.38	98.01
Asphalt	500/8748	98.34	95.38	98.58	98.88	97.93	98.82
Bitumen	500/6787	92.75	84.93	92.57	64.56	99.37	95.86
Tiles	500/42326	99.06	98.27	99.81	98.51	98.36	99.35
Shadows	500/2363	100	99.96	100	99.83	99.58	100
Kappa×100	/	98.17	95.45	98.23	95.11	97.47	98.58
OA(%)	/	98.73	96.83	98.77	96.59	98.24	99.01
AA(%)	/	97.14	92.54	96.82	90.69	97.61	97.63

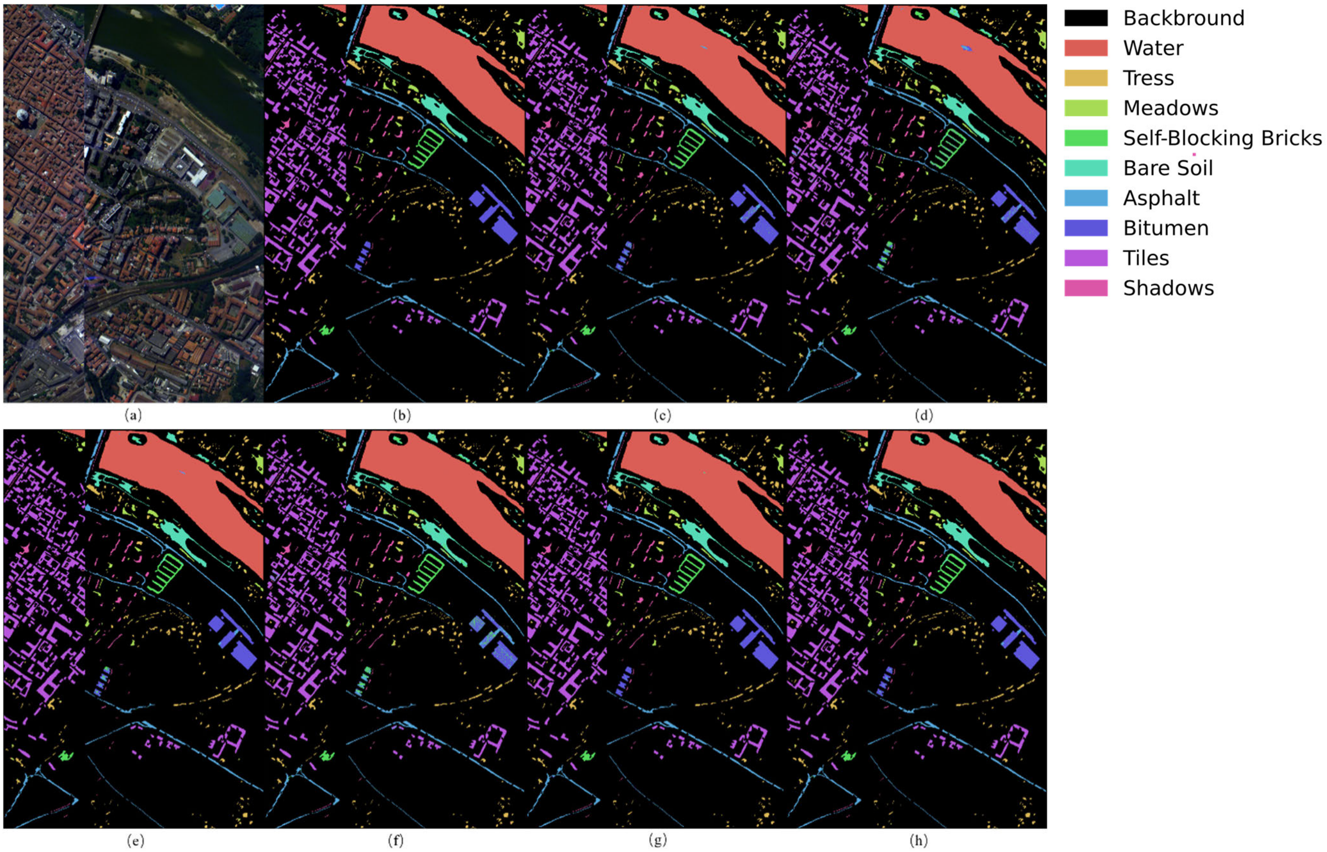


FIGURE 10. Classification maps for the Pavia Center dataset. (a) True-color image. (b) Ground-truth map. (c)-(h) Classification maps of SVM, D-CNN, CD-CNN, RNN SS-CNN and SSPAN.

increases of 3.2, 2.54, and 0.28 for K ×100, OA, and AA, respectively.

Comparing Tables 1 and 2, it can be concluded that the accuracy of each model improves in the Pavia University datasets This is due to the increase in spatial resolution from 20 to 1.3 meters per pixel. Moreover, the training samples also increase slightly.

3) EXPERIMENTS ON PAVIA CENTER DATASET

In the Pavia Center dataset, all the categories were selected for training. The size and number of each category of the Pavia Center dataset are greater than those of the other datasets. Therefore, 500 samples from each category were selected as training samples, and the remaining samples were selected as testing samples. The classification results are shown in Table 6. The accuracies are also depicted visually. The classification maps of the compared methods for the Pavia Center dataset are displayed in Fig. 10.

Compared to the other two datasets, all models achieved fairly high classification accuracy in the Pavia Center dataset due to the higher spatial resolution and greater number of training samples. According to Table 6, based on K ×100, OA, and AA, the proposed SSPAN model yields the best classification results. Compared to the traditional SVM method, SSPAN yields accuracy increases of 0.41, 0.28, and 0.49 for K×100, OA, and AA, respectively. Compared to the

D-CNN and RNN, SSPAN has considerable advantages. Relative to the CD-CNN network, SSPAN yields accuracy increases of 0.35, 0.24 and 0.81 for K ×100, OA, and AA, respectively. Relative to the SS-CNN network, the SSPAN model yields accuracy increases of 1.11, 0.77, and 0.02 for K ×100, OA, and AA, respectively.

In Pavia Center dataset, the gaps between SSPAN and the other models are smaller. The accuracies of D-CNN and RNN are more than 2% lower than those of the other models. The classification accuracies of the remaining three models are very close. To further compare the performances of the models, this paper focuses on comparing the computational complexities of the CD-CNN, SS-CNN, and SSPAN models. The floating point operations (FLOPs) and parameters are shown in Table 7.

TABLE 7. **FLOPs and parameters of different methods.**

MODELS	FLOPs	PARAS
CD-CNN	6572800	264074
SS-CNN	7899040	4354672
SSPAN	1714110	566110

Table 7 shows that although the number of SSPAN parameters is twice that of the CD-CNN, it is only 13% of the number of parameters of the SS-CNN. However, due to the complete abandonment of the CNN and the application of the attention mechanism, the FLOPs of the SSPAN model are only 26% of those of the CD-CNN and 21% of those of the SS-CNN. SSPAN ensures the best classification accuracy while significantly reducing the computational complexity.

IV. CONCLUSION

For HSI classification tasks, this paper proposes a novel SSPAN framework composed of an attention mechanism to capture spectral and spatial features. First, the original spectral bands of the pixels to be classified are directly used as the input of the spectral attention module, which can simultaneously capture the partial and global spectral correlations. Second, the 3D HSI cube is designed and used as the input of the spatial attention module. The spatial attention module can emphasize adjacent and effective pixels and weaken distant and redundant pixels. Then, the cross-field adaptive gating module fuses the spectral and spatial features selectively, which can enhance the information interaction between the spectral and spatial domains. Finally, the SSPAN model achieved state-of-the-art performance on three datasets compared to other methods.

The future direction of the proposed work is to incorporate new variants of attention mechanisms, such as deformable attention and coordinate attention, aimed at learning more discriminative spectral-spatial features. Meanwhile future research will pursue to design an attention mechanism capable of capturing both spectral and spatial features. Furthermore, how to design an efficient feature fusion module is another research direction. In addition, structural re-parameterization and is also under consideration for future research.

REFERENCES

[1] P. Ghamisi, M. D. Mura, and J. A. Benediktsson, "A survey on spectral–spatial classification techniques based on attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2335–2353, May 2015.

[2] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, Jun. 2013.

[3] C. Zhang, S. Zhu, D. Xue, and S. Sun, "Gabor filter-based multi-scale dense network hyperspectral remote sensing image classification technique," *IEEE Access*, vol. 11, pp. 114146–114154, 2023.

[4] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[5] A. A. Farag, R. M. Mohamed, and A. El-Baz, "A unified framework for MAP estimation in remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1617–1634, Jul. 2005.

[6] L. O. Jimenez, A. Morales-Morell, and A. Creus, "Classification of hyperdimensional data based on feature and decision fusion approaches using projection pursuit, majority voting, and neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1360–1366, May 1999.

[7] J. A. Benediktsson and I. Kanellopoulos, "Classification of multisource and hyperspectral data based on decision fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1367–1377, May 1999.

[8] C.-I. Chang and Q. Du, "Interference and noise-adjusted principal components analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 5, pp. 2387–2396, Sep. 1999.

[9] C. Lee, S. Youn, T. Jeong, E. Lee, and J. Serra-Sagrista, "Hybrid compression of hyperspectral images based on PCA with pre-encoding discriminant information," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1491–1495, Jul. 2015.

[10] S. Li, P. Kou, M. Ma, H. Yang, S. Huang, and Z. Yang, "Application of semi-supervised learning in image classification: Research on fusion of labeled and unlabeled data," *IEEE Access*, vol. 12, pp. 27331–27343, 2024.

[11] K. Sun and J. Zhu, "Learning consistency from high-confidence pseudo-labels for weakly supervised object localization," *IEEE Access*, vol. 11, pp. 16657–16666, 2023.

[12] Y. Li, T. Shi, Y. Zhang, and J. Ma, "SPGAN-DA: Semantic-preserved generative adversarial network for domain adaptive remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.

[13] R. Li, C. He, Y. Zhang, S. Li, L. Chen, and L. Zhang, "SIM: Semantic-aware instance mask generation for box-supervised instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7193–7203.

[14] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

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

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

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014, *arXiv:1409.4842*.

[19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[20] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

[21] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[22] J. Donahue, L. Anne Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," 2014, *arXiv:1411.4389*.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[24] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, no. 1, pp. 1–12, Jul. 2015.

[25] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

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

[27] I. O. Sigirci and G. Bilgin, "Spectral–spatial classification of hyperspectral images using BERT-based methods with HyperSLIC segment embeddings," *IEEE Access*, vol. 10, pp. 79152–79164, 2022.

[28] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, *arXiv:1709.01507*.

[29] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.

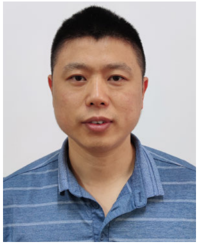
[30] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[31] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Systems(NIPS)*, 2017, pp. 5998–6008.

[32] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

WEIYI LIAO received the M.Sc. degree from the Army Engineering University of PLA, Nanjing, China. He is currently a Lecturer with the Army Engineering University of PLA. His research interests include remote sensing image processing and computer vision.

FENGSHAN WANG received the Ph.D. degree from the Army Engineering University of PLA, Nanjing, China. He is currently an Associate Professor with the Army Engineering University of PLA. His research interests include system analysis and remote sensing image processing.



HUACHEN ZHAO received the M.Sc. degree from the Army Engineering University of PLA, Nanjing, China. He is currently a Lecturer with the Army Engineering University of PLA. His research interests include digital earth and remote sensing image processing.