IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 9, 2021, accepted August 24, 2021, date of publication August 26, 2021, date of current version September 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3107976

Spectral-Spatial Classification of Hyperspectral Images Using Label Dependence

ZHUANGZHUANG HE, (Student Member, IEEE), HAO WU, (Student Member, IEEE), AND GUODONG WU, (Member, IEEE)

College of Information and Computer, Anhui Agricultural University, Hefei, Anhui 230036, China

Corresponding author: Guodong Wu (gdwu1120@qq.com)

This work was supported by the National Science Foundation of China under Grant 31671589.

ABSTRACT Hyperspectral images are rich in both spectral information and spatial dependence information between pixels; however, hyperspectral images are characterized by the high dimensionality of small data sets and the spectral variance. Facing these problems, spatial dependence information as supplementary information is a relatively effective means to solve them. And the label dependence characteristic of hyperspectral images is excellent spatial dependence information. Therefore, to address the above issues, based on residual network and spatial information extractor(RAS), which is based on a residual network, pixel embedding(PE), and a spatial information extractor(SIE). At the stage of mining spectral information, we use the residual network to mine spectral features; At the stage of mining spatial information, we utilize the label dependency characteristic to feed the set of pixels containing the target pixels into PE. Then, a pixel vector with location information and self-defined dimensionality is obtained. Next, this vector is fed into our proposed SIE to mine the spatial dependency information. In multi-group ablation experiments, our proposed model achieves overall accuracy (OA) scores of 79.16% on the 5% Indian Pines test set, 90.82% on the 1% Pavia University test set, and 92.17% on the 1% Salinas test set. Especially, the experimental results demonstrate that the joint spectral-spatial approach is effective in improving the accuracy of hyperspectral image classification.

INDEX TERMS Deep learning, hyperspectral image, image classification, pattern recognition.

I. INTRODUCTION

Hyperspectral images (also known as remote sensing data) have continuous, multiband narrow spectral bands [1]. The wide spectral range carries substantial spatial and spectral information [2]. Due to the abundant information in hyperspectral images, the technology is used in many fields, such as agriculture [3], medicine [4], and food safety [5]. In recent years, the recognition and classification of target objects in hyperspectral images have become an important direction for research in the hyperspectral image field [6]. Hyperspectral image classification is the classification of pixel points in an image, and the usual method is to use a priori information in the image, such as a small number of labeled training samples, to learn to discriminate the classes corresponding to other pixels in the hyperspectral image. Because the spectral range of hyperspectral images is wider than that of ordinary images, they carry more useful information in the continuous bands but also include large amounts of redundant information. Therefore, hyperspectral image classification is a challenging task. In many methods to classify the category of pixels, they employ exclusively spectral signatures for classification. This approach has two advantages, the concept is easy to understand and can be easily implemented. To retain the useful information while eliminating the redundant information, Mahesh and Foody [13] used a support vector machine (SVM) to perform feature selection, and the experimental results showed good performances even under conditions with limited samples. Zhong *et al.* [14] used a conditional random field (CRF) algorithm trained on samples; the trained model was able to eliminate most of the redundant information. To further improve the classification accuracy, Han *et al.* [16] proposed using a pretrained AlexNet neural network model to deeply mine the image feature information and significantly improve the classification accuracy.

However, this pixel-level classification approach has two limiting factors, the high dimensionality of small data sets, and the spectral variance [28]. For the first issue, researchers

The associate editor coordinating the review of this manuscript and approving it for publication was Dominik Strzalka.

usually view it from the perspective of Hughes, which usually raises two problems. First, a small number of the hyperspectral image labeled samples may lead to singularities in the sample covariance matrix, which leads to problems of unavailability of some classification methods. Second, the high dimensional characteristics of the spectrum lead to many parameters in the model needing to be estimated. And this raises the problem of reduced model generalization ability and overfitting. For the second issue, Regarding the spectral variation, which is brought about by many factors such as atmospheric effects, unwanted shadows, and shading, and instrument failures [7]–[9]. This causes a significant problem in that distinguishing some pixel categories becomes difficult. All these may hinder the classification of hyperspectral images.

So how to weaken these problems and thus improve the accuracy of hyperspectral classification has become crucial for researchers to study. Considering the hyperspectral image itself, since hyperspectral images are inherently 3-D and pictorial, the spatial information that complements spectral behavior naturally makes them a useful source of information in addition to the spectrum. A concept closely related to spatial information is spatial dependency, which refers to the spatial relationship between neighboring pixels. According to Tobler's first law of geography, the similarity between two objects on the same geographical surface is inversely proportional to their distance [10]. Therefore, spatially related pixels are called neighboring pixels, and all these neighboring pixels are in the same neighborhood. Therefore, the introduction of spatial dependence offers the possibility to improve pixel classification. Early attempts to incorporate spatial information into hyperspectral classification date back more than a decade, and some successful studies have shown its ability to improve the classification. In addition, spatial dependence is associated with label dependence(see Figure 1), which refers to the correlation of labels of neighboring pixels, where the labels of pixels in a small area are likely to be the same. So we can utilize this factor to improve the classification effect. To the best of our knowledge, we are the first to apply deep learning methods to extract the spatial information underlying the label dependencies.

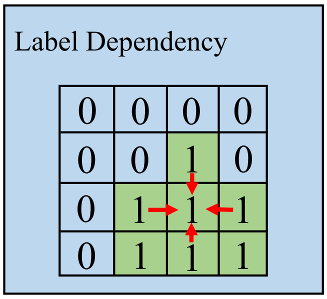

FIGURE 1. The label dependency diagram. If a pixel's category is 1, then its surrounding labels are also 1 with a high probability, and then the surrounding pixels can be used as spatial information (context information).

Based on this knowledge, it is necessary to extract the spatial information in the image as a complement to the spectral information and thus enhance the classification ability. Extracting both the spatial and spectral information from hyperspectral images. Then fusing the two pieces of information in a certain way to complete a classified task is deserving of attention. Tarabalka *et al.* [35] proposed a spectral–spatial classification scheme. He segmented the images by clustering, the segmentation provides an adaptive neighborhood for each pixel and uses an pixel wise SVM method. Tao *et al.* [36] built two different learning procedures for spatial and spectral information, sparse spectral feature learning and multi-scale spatial feature learning. Zhao *et al.* [37] introduced game theory into hyperspectral image classification and uses conditional random fields to model the images while taking spatial background information into account. However, these models are complex and have more parameters than our proposed model, and are not easy to fit. Other researchers have mined the spatial-spectrum information from some interesting perspectives. Ma *et al.* [38] proposed a spatial-spectum kernel generation module, and the experimental results show that the module is effective. Zhu *et al.* [39] proposed a triple-branch progressive fusion residual network for classification. Ma *et al.* [40] proposed a dual-branch interactive spatial-channel collaborative attention enhancement network (SCCA-net) for classification. And many researchers have investigated this issue. convolutional neural networks(CNN) are not only outstanding in general image processing, but also perform well in hyperspectral images and are often used to capture spatial information. Slavkovikj *et al.* [11] presented a CNN framework for HSI classification where the proposed model is capable of able to learn spatial

information. 2D and 3D CNNs are widely used for hyperspectral classification because of the excellent information mining performance. Gao *et al.* [31] proposed a two-dimensional spectral image method that makes full use of spectral values and spatial information. The problem of heterogeneous noise caused by the traditional data processing method with small area pixel blocks or one-dimensional spectral vectors as input units is solved. Liu *et al.* [32] extracts features from spectral and spatial dimensions by applying 3D convolution, thus capturing important identification information encoded in multiple adjacent frequency bands. Makantasis *et al.* [12] combined random principal component analysis (RPCA) and CNN into a new model (named RPCA-CNN) for the joint extraction of spectral and spatial information. Li *et al.* [30] proposed a new CNN-based method to encode hyperspectral image features and predict them by a voting strategy. According to their experimental results, the CNN model does work well. However, CNN-based models still have some problems, such as limited perception fields and difficulty in generalization. On one hand, a large convolutional kernel limits the depth of information extracted by the CNN, while a small convolutional kernel (3 × 3, 5 × 5) limits the perceptual domain. On the other hand, CNN-based models cannot adapt to different sizes of the same shaped region. For example, If the model is trained on a 48 × 48 square area, it must be retrained on a 36 × 36 area if it is desired to predict a 36 × 36 square area.

In conclusion, in order to solve the problem of using exclusively spectral information and limited CNN receptive fields. we propose a combined deep learning model (RAS). Our model consists of two modules, the spectral information module, and the spatial information module. The spectral information extraction module is mainly composed of residual networks, and the spatial information module is mainly composed of many spatial information extractor(SIE) and pixel embedding(PE). And to abbreviate the name of the model, so we call it RAS.

The proposed spatial information extractor is based on the transformer in the field of natural language processing. transformer performs extremely well in the area of natural language processing(NLP), both in speech and in semantic extraction. Based on two aspects, we migrated the transformer to the field of hyperspectral image processing. First of all, from a spectral aspect, a pixel can be analogous to a sentence(see Figure 2). Because both can be considered as a vector and both represent certain meanings. The pixel vector represents the land situation, while the word vector represents a word. In addition, from a spatial aspect, the pixels around a pixel can be analogous to the context of a word in the corpus. Based on these two analogies, we can input pixel sequences into the language model like sentences. Motivated by these analogies, we propose an improved transformer-SIE based on a language model.

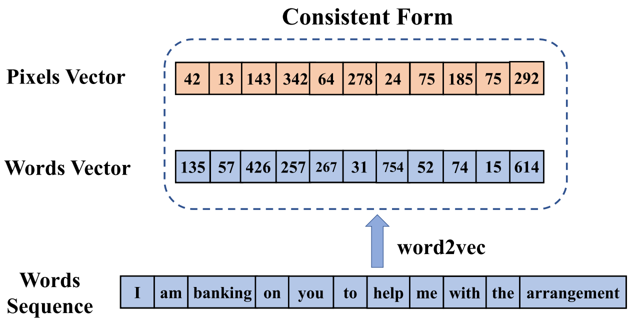

FIGURE 2. A pixel vector (data vector) with a sequence of words is converted into a word vector (data vector) of the same form as a pixel vector.(word2vec is a method in the field of natural language processing; transforming word sequences into word vectors.)

Therefore, SIE is an approach to adapt transformer to HSI feature extraction and classification. SIE acquires the global receptive fields through the self-attention(SA) mechanism. Thus it can capture richer global background information. It enables better extraction of spatial dependency information from label dependencies. Given an input region, the SA mechanism can capture the relationship between two different pixels without caring about their spatial distance. Compared with the mentioned CNN models, SA is more flexible, and this feature makes it possible to dynamically select the context, which is the crux of obtaining spatial information. In addition, 1 × 1 convolution network layers are used to increase the information interaction between channels while reducing the feature dimension and the number of parameters, which is conducive to improved training.

This model not only uses SIE with global receptive domain to mine spatial dependencies from the labels, but also combines the spatial information with the spectral information to do the final classification task. The results of various comparison experiments and ablation experiments show that RAS is relatively effective in solving the problems of poor generalization ability and limited receptive fields using exclusively spectral information. The main contributions of this article are as follows:

- We take advantage of the label dependence characteristic in hyperspectral images to obtain spatial dependence information from them.
- To better extract the spatial dependence information, we draw inspiration from NLP domain methods to design SIE and PE for hyperspectral images.

The rest of this paper is organized as follows. Section II describes the RAS algorithm, and Section III describes the data set and evaluation method with parameter settings. Section IV reports the experimental results and analyzes them. Section V gives a discussion of some issues in this paper, and Section VI draws a conclusion of this paper.

II. PROPOSED APPROACH

Figure 3 shows the RAS model. First, we prepare two identical training datasets and feed them into the spectral information module and the spatial information module, respectively. In the upper module of the figure, we use two residual blocks to extract useful information in the image and eliminate redundant information in the image, thus retaining the spectral information in it. In the lower module of the figure. As it is described in Figure 1. Suppose we want to predict a given target pixel, then we simultaneously flat it and the pixel region (context) around it into a pixel sequence as well. This is the input pixel sequence. Then the pixel sequence is transformed using pixel embedding(PE) to convert that pixel vector to a pixel vector of self-defined dimensions(We take five neighboring pixels as a group and change the pixel dimension to their average dimension by PE.). After that, the pixel sequence is fed into a multilayer SIE to automatically mine the contextual information between pixels and obtain spatial dependence information. Finally, the obtained spatial information is stitched with the spectral information after complete concatenation processing to obtain the spectral-spatial information. The prediction is performed by the softmax function.

Therefore, based on the above process, four main works are carried out in this setion.

- In subsection *A*. We introduce the inspiration for ResNet to be used for spectral information extraction, and the formula deduction.
- In subsection *B*. We introduce why PE is used, and the specific formulas, diagrams, and computer process charts for PE.

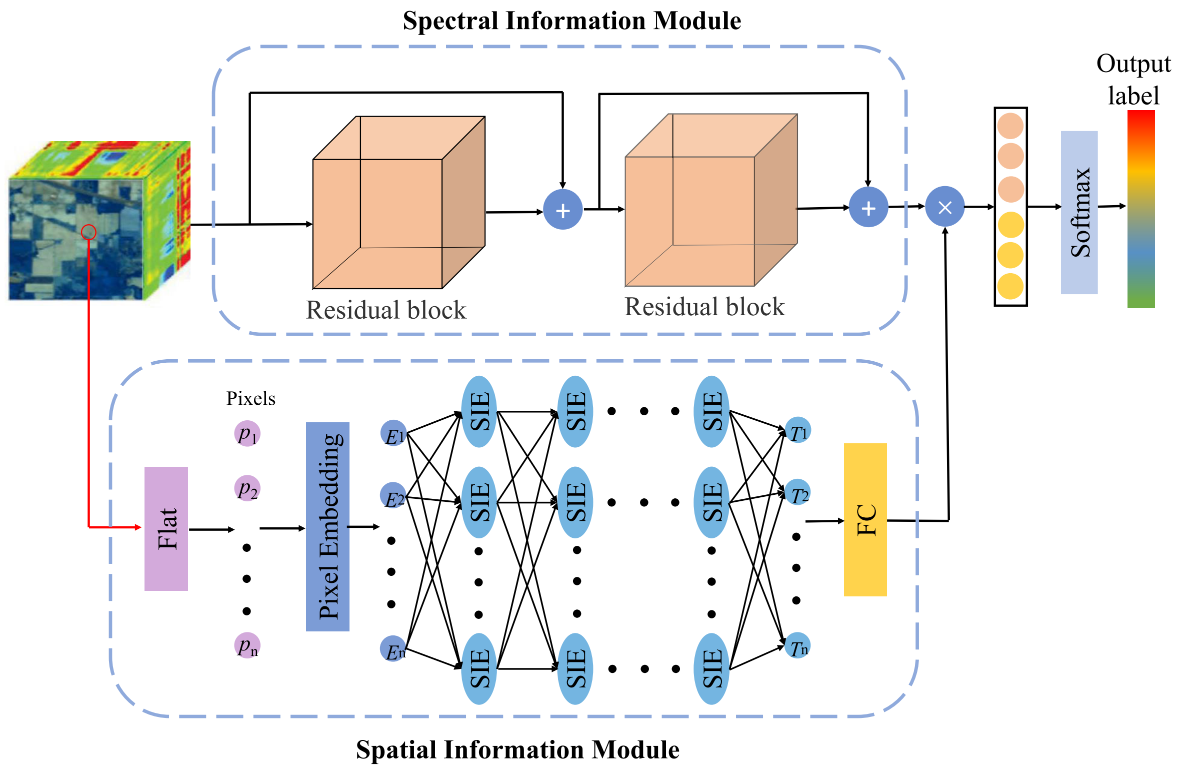

FIGURE 3. RAS algorithm model. $P_1, P_2, \ldots, P_n$ are is the pixel vector after flattening. $E_1, E_2, \ldots, E_n$ are the embedded pixels after the pixel embedding(PE) process. $T_1, T_2, \ldots, T_n$ are the encoded pixel sequence processed through multiple SIE layers. n represents the maximum pixel sequence length. In the spectral information model, the ? represents two feature summing operations and the × represents two feature concatenating operations. FC is a fully connected layer.

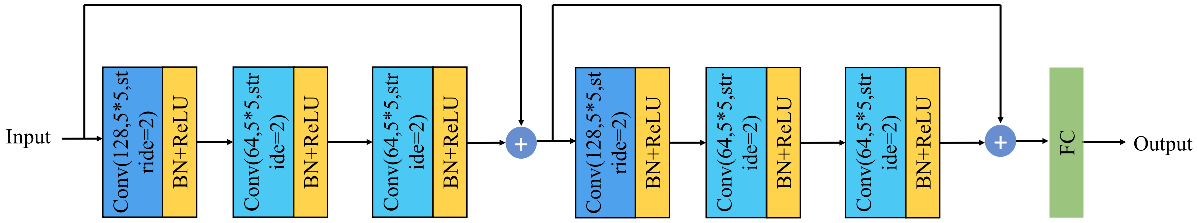

FIGURE 4. The process of obtaining spectral information.

- In subsection C. We introduce the spatial information extractor, which details the action of the self-attention mechanism, and its diagram of the global attention mechanism. The reason why we use 1 × 1 convolution is introduced.
- In subsection D. We introduce feature fusion, and their formulas.

A. RESIDUAL NETWORK

Convolution networks plays a very important role in the field of image processing, which can deeply explore the spectral information contained in images as well as eliminate the redundant information in images. We refer to the residual network proposed by He *et al.* [18] to design the residual block. Its structure in detail is shown in Figure 4. The residual network consists of mainly two residual blocks, which contain three convolutional units, and the convolutional units are divided into two categories, one contains one convolutional kernel of size (128, 5 × 5), a Batch Normalization, and activation function ReLU. the other contains one convolutional kernel of size (64, 5 × 5), a Batch Normalization, activation function ReLU. The formula of residual networks is as follows.

$$H(x)^{(k+1)} = x^{(k-1)} + F(x)^{k}(1 \leq k \leq 2). \qquad (1)$$

where, $H(x)$ represents the input of the next residual block. $F(x)$ represents the output of the current residual block, and x represents the output of the previous residual block. k represents that kth residual block.

B. PIXEL EMBEDDING

The purpose of PE is to transform the original pixel vector into a vector of new self-defined dimensions. (see Figure 6) It was added to the spatial information module for two reasons.

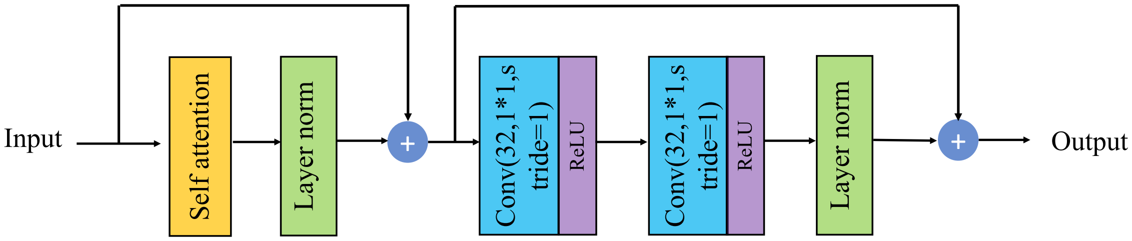

FIGURE 5. Spatial information extractor(SIE) details.

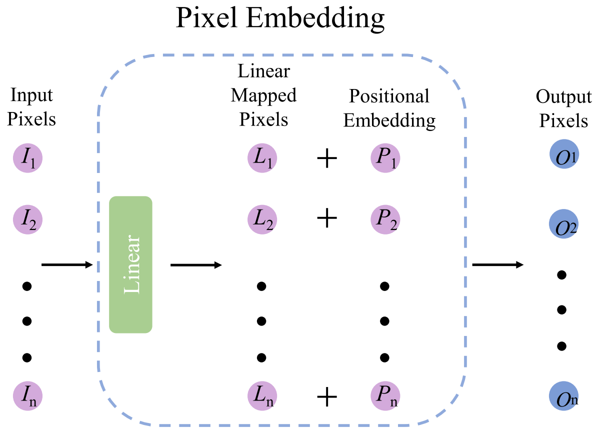

FIGURE 6. The detailed process of pixel embedding. The linear mapped pixels are pixels processed through the linear transformation (LT), while positional embedding is the learned positional embedding(LPE). The capital letters in the circles are abbreviations of the names above their circles. *n* represents the maximum pixel sequence length.

First, PE allows different input dimensions to be used for this model and can transform the initial vector dimensions into dimensions that fit the model. Second, PE can also be used to reduce the input dimensions to speed up the training. PE consists of an Linear mapped pixels(LMP) and a positional embedding [19], [20].

1) LINEAR MAPPED PIXELS

We obtain LMP by linear transformation(LT) of the input pixels. As shown in Figure 6.

$$LT(X) = W^T X. \tag{2}$$

where, LT is an abbreviation for linear transformation. LMP can be obtained after LT. W is the weight of the learned linear information extractor. X is the original pixel vector.

2) POSITION EMBEDDING

Position embedding is a common technique used in NLP to encode the location of words. Here it is used to encode the position information of each pixel to generate a sequence of pixels. Position encoding works in the self-attention stage. Suppose a pixel vector is to be computed with two identical pixel vectors for attention, if there is no position encoding to show the difference, then the same attention value will be obtained, but in general the pixel is not associated with the same two pixels.

$$LPE(X) = X + P. \tag{3}$$

where, LPE is the abbreviation for learned position embedding, and P is a position matrix having the same shape as X.

The P is concretely calculated as follows.

$$P(pos, 2i) = sin(\frac{pos}{10000^{\frac{2i}{d}}}). \tag{4}$$

$$P(pos, 2i+1) = cos(\frac{pos}{10000^{\frac{2i}{d}}}). \tag{5}$$

where, pos represents the position of a value in the pixel vector, $2i$ represents an even position and $2i + 1$ represents an odd position. d represents the dimensionality of a pixel vector.

C. SPATIAL INFORMATION EXTRACTOR

Since the birth of the Transformer [27], Transformer the NLP world has received great praise, such as in machine translation [21], question answer [22], [23], language understanding [24] and other areas of top performance. Based on transformer architecture, we propose a spatial information extrator model as shown in Figure 5. This has the advantage that the spatial information in the hyperspectral images can be extracted efficiently.

1) SELF ATTENTION

Attention is a technique that allows models to focus on important information and learn to absorb it fully [29]. The self-attention mechanism is a variant of the attentional mechanism that is less dependent on external information. It is more adept at capturing the internal relevance of features [15]. The self-attention mechanism obtains global receptive fields and contextual information by capturing a broader range of information. Figure 8 shows the difference between the global receptive field and the limited receptive field. The global receptive field can capture the global information, while the limited receptive field can only capture the limited information. The specific computation process of the attention mechanism can be summarized in two processes: the first process is to compute the weighting coefficients based on Query and Key, and the second process is to compute the sum of the weights of Value based on the weighting coefficients [26]. SIE uses self-attention, while [17] uses the Multiple Head self-attention mechanism (MHSA). We use self-attention

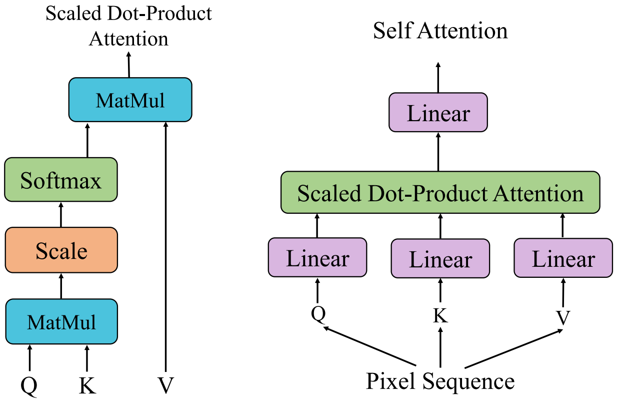

FIGURE 7. (Left) Scaled dot-product attention and (Right) SA mechanism for the pixel sequence.

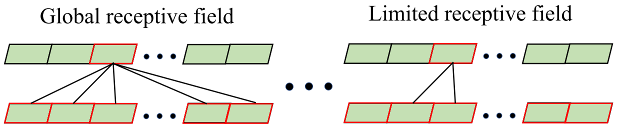

FIGURE 8. (Left) Global receptive field versus (Right) limited receptive field in a pixel sequence.

instead of MHSA for two reasons. First, the original authors set up multiple heads because they wanted to mine more feature information, but when Li *et al.* [25] used regularization to test the attention of each head, they found that they could not extract their ''jurisdiction'' by initialization as we thought. In addition, we only need spatial information mining, so we need to ensure the purity of information extraction. Second, MHSA requires more training time than single-headed. The specific calculation process, as shown in Figure 7.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

where, Q, K, V represent query, key, value. d is the dimension of the input data.

2) CONVOLUTION 1 × 1

We use 1 × 1 convolution to integrate the spatial information in hyperspectral images for the following reasons: First, without changing the spatial structure of the image, using the full connection layer will destroy the spatial structure of the image, while 1 × 1 convolution layer will not destroy the spatial structure of the image. Second, cross-channel information transfer, for example, a filter having a convolution kernel size of 64 and a filter having a convolution kernel size of 1 × 1 and a convolution kernel number of 28, wherein that output layer size is equal to the output layer size obtained by a filter having a convolution kernel size of 3 × 3 and a convolution kernel number of 28, and the original 64 channels can be understood as cross-channel linear combination into 28 channels, which is the information interaction between channels. Thirdly, reducing the parameter, and reducing the dimension is reducing the parameter. Since that feature map is small, the parameter is reduced accordingly, which is equivalent to convolution on the channel number of the feature map, compressing the feature map to extract the feature twice, so that the feature expression of the new feature map is better.

$$CN(x) = \sigma(\sigma(W_1x + b_1)W_2 + b_2) \tag{7}$$

where, CN represent that convolution result of two layers 1 × 1, σ representing the activation function ReLU, W represents the weight, and b represents the bias.

D. FEATURE FUSION

In order to maintain the integrity of the spectral and spatial information, we use feature concatenation to connect the two features.

$$Output = concat([Output_{ResNet}, Output_{SIE}]) \tag{8}$$

where, *Output* represent that result of feature fusion, *concat* is an abbreviation for concatenate. The classification of the output is finally done by the softmax function.

III. DATA SETS, EVALUATION METRICS AND PARAMETER SETTINGS

In this section, we detail three popular public datasets, namely Indian Pines, Pavia University, and Salinas. all experimental metrics and training parameter settings in the experiments are also detailed.

A. DATA SETS

Here we will introduce three popular datasets and describe in detail their number of pixels and number of categories.

1) INDIAN PINES DATA SET

The first test dataset used for hyperspectral image classification in this paper is Indian Pines. In 1992, the air-borne visual infrared imaging spectrometer (AVIRIS) imaged 145 × 145 hyperspectral images of an Indian pine tree in Indiana, USA. AVIRS continuously imaged 220 wavebands of ground objects at an imaging wavelength of 0.4-2.5 μm and a spatial resolution of approximately 20 m. However, because 20 of these bands are not reflected by water, researchers generally use only the remaining 200 bands as subjects. The dataset includes a total of 21,025 pixels, but only 10,249 pixels represent features; the remaining 10,776 pixels are limited to background. The dataset contains 16 ground object categories, as shown in Table 1.

2) PAVIA UNIVERSITY DATA SET

The Pavia University hyperspectral data were imaged by an airborne reflectance optical spectral imager (Re-flective Optics Spectrographic Imaging System, ROSIS-03) in 2003 at 610 × 340 over Pavia, Italy. ROSIS continuously imaged 115 wavebands of the ground surface at an imaging wavelength of 0.43-0.86 μm and a spatial resolution of approximately 1.3 m, but 12 wavebands were deleted due to noise. The remaining 103 wavebands are

TABLE 1. The number of samples of each class in the Indian Pines dataset.

Serial Number	Classification	Number of samples
1	Alfalfa	54
2	Corn-notill	1434
3	Corn-mintill	834
4	Corn	234
5	Grass-pasture	497
6	Grass-trees	747
7	Grass-pasture-moved	26
8	Hay-windrowed	489
9	Oats	20
10	Soybean-notill	968
11	Soybean-mintill	2468
12	Soybean-clean	614
13	Wheat	212
14	Woods	1294
15	Building-mass-trees-drives	380
16	Stone-steal-towers	95

TABLE 2. The number of samples of each class in the Pavia University dataset.

Serial Number	Classification	Number of samples
1	Asphalt	6631
2	Meadows	18649
3	Gravels	2099
4	Trees	3064
5	Painted metal sheets	1345
6	BareSoil	5029
7	Bitumen	1330
8	Self- Blocking Bricks	3682
9	Shadows	947

generally used as research objects. The dataset has a total of 2,207,400 pixels, but only 42,776 of these pixels are available as feature pixels; the remaining 2,164,624 pixels are background pixels. In addition, the dataset includes nine ground object categories, as shown in Table 2.

3) SALINAS DATASET

The Salinas dataset was also captured by the AVIRIS Imaging Spectrometer and includes a 512 × 217 pixel hy-perspectral image of the Salinas Valley in California, USA. AVIRIS continuously imaged 224 bands of surface fea-tures with an imaging spatial resolution of approximately 3.7 but 20 of the bands were affected by noise and deleted. The remaining 204 bands are generally used as subjects. The dataset has a total of 111,104 pixels, but only 54,129 pixels are available as feature pixels; the remaining 56,975 pixels are background pixels. In addition, the dataset includes 16 ground object categories, as shown in Table 3.

B. EVALUATION METRICS

The OA, AA, and the Kappa coefficient (Kappa) were used to evaluate the classification ability of the proposed model. The specific calculations are as follows.

$$OA = \frac{\sum_{i=1}^{n} h_{ii}}{\sum_{i=1}^{n} N_i} \tag{9}$$

TABLE 3. The number of samples of each class in the Salinas dataset.

Serial Number	Classification	Number of samples
1	Brocoli_green_weeds_1	2009
2	Brocoli_green_weeds_2	3726
3	Fallow	1976
4	Fallow_rough_plow	1394
5	Fallow_smooth	2678
6	Stubble	3959
7	Celery	3579
8	Grapes_untrained	11271
9	Soil_vinyard_develop	6203
10	Corn_seneseed_green_weeds	3278
11	Lettuce_romaine_4wk	1068
12	Lettuce_romaine_5wk	1927
13	Lettuce_romaine_6wk	916
14	Lettuce_romaine_7wk	1070
15	Vinyard_untrained	7268
16	Vinyard_vertical_trellis	1807

where, n is the number of classes of the hyperspectral image feature object, N_i is the number of the ith category image element, and h_{ii} is the number of the ith category image element that is correctly classified.

$$AA = \frac{1}{n}\sum_{i=1}^{n}\frac{h_{ii}}{N_i} \tag{10}$$

where, N represents the number of sample pixels for testing in the total training sample, n represents the number of categories, and h_{ii} is the number of pixels correctly classified in the ith category.

$$kappa = \frac{N\sum_{i=1}^{N} m_{ii} - \sum_{i=1}^{N} m_{i+}m_{+i}}{N^2 - \sum_{i=1}^{N} m_{i+}m_{+i}} \tag{11}$$

where, m_{ii} represents the value of the i-th row and i-th column. m_{i+} denotes the sum of row i, m_{+i} denotes the sum of column i. N is the total number of samples.

C. PARAMETER SETTINGS

In order to ensure the rigour of the experiment, we randomly divided the experimental data set and the test data set, and set the random division status to 5. We set the value of batch size to 128 and we used a rigorous research approach, running 200 epochs in each experiment. We experimented with multiple groups of learning rates, 0.1, 0.001, 0.005, 0.0001; finally we chose the best result of 0.01 for the whole experiment. To speed up the model convergence, we choose the Adam optimizer. Also the loss function we use is the cross-entropy loss function.

IV. EXPERIMENTS

Four main works were carried out in this section.

- In subsection A. To ensure fair comparisons in our experiment, our model is compared with three recent models (BERT, 2D-CNN, and 3D-CNN) and three classic models (ResNet, AlexNet, DenseNet) on three classic datasets. The experimental results are reported in Tables 4–6, which show the results on the Indian

TABLE 4. The classification accuracies of different methods on the Indian Pines images tested on training sets of 5%, 10%, and 15% of the total sample size. Red entries indicate the best results. Blue entries indicate the second-best results.

Methods	OA (5%)	AA (5%)	Kappa (5%)	OA (10%)	AA (10%)	Kappa (10%)	OA (15%)	AA (15%)	Kappa (15%)
AlexNet [16]	0.6904	0.6286	0.6437	0.7353	0.7124	0.6861	0.7964	0.7887	0.7633
ResNet [33]	0.6909	0.6024	0.6445	0.7764	0.7421	0.7592	0.8078	0.8147	0.7834
DenseNet [34]	0.7308	0.6820	0.6961	0.7848	0.7911	0.7586	0.8518	0.8345	0.8125
BERT [17]	0.7627	0.7584	0.7496	0.8231	0.8434	0.8078	0.8823	0.8745	0.8549
2D-CNN [31]	0.7487	0.7757	0.7243	0.8062	0.8141	0.7949	0.8194	0.8456	0.7993
3D-CNN [32]	0.7427	0.7154	0.7384	0.8131	0.7735	0.7892	0.8357	0.8122	0.8563
RAS	0.7916	0.7797	0.7694	0.8497	0.8374	0.8231	0.9077	0.8829	0.8935

TABLE 5. The classification accuracies of different methods on the Pavia University images tested on training sets of 1%, 5%, and 10% of the total sample size. Red entries indicate the best results. Blue entries indicate the second-best results.

Methods	OA (1%)	AA (1%)	Kappa (1%)	OA (5%)	AA (5%)	Kappa (5%)	OA (10%)	AA (10%)	Kappa (10%)
AlexNet [16]	0.8442	0.8603	0.7887	0.9178	0.9212	0.8943	0.9233	0.9315	0.8857
ResNet [33]	0.8361	0.8583	0.7753	0.8902	0.9145	0.8861	0.9378	0.9084	0.9187
DenseNet [34]	0.8537	0.8245	0.8071	0.9101	0.9128	0.9001	0.9348	0.8973	0.9122
BERT [17]	0.9014	0.8747	0.8633	0.9264	0.9301	0.8867	0.9495	0.9053	0.9042
2D-CNN [31]	0.8752	0.8683	0.8357	0.8927	0.9189	0.8913	0.9175	0.9234	0.8853
3D-CNN [32]	0.8857	0.8864	0.8763	0.9152	0.8936	0.8874	0.9153	0.9186	0.8861
RAS	0.9082	0.8884	0.8719	0.9312	0.9392	0.9265	0.9576	0.9412	0.9346

TABLE 6. The classification accuracies of different methods on the Salinas images are tested on training sets of 1%, 5%, and 10% of the total sample size. Red entries indicate the best results. Blue entries indicate the second-best results.

Methods	OA (1%)	AA (1%)	Kappa (1%)	OA (5%)	AA (5%)	Kappa (5%)	OA (10%)	AA (10%)	Kappa (10%)
AlexNet [16]	0.8645	0.8919	0.8403	0.9213	0.9458	0.9043	0.9317	0.8963	0.8722
ResNet [33]	0.8818	0.9148	0.8685	0.9303	0.8918	0.8845	0.9265	0.8876	0.8870
DenseNet [34]	0.8746	0.9217	0.8392	0.8978	0.9143	0.9012	0.8943	0.9157	0.9086
BERT [17]	0.9142	0.9249	0.8983	0.9337	0.9422	0.9311	0.9443	0.9214	0.9028
2D-CNN [31]	0.9041	0.9153	0.8953	0.9056	0.9156	0.8931	0.9142	0.9163	0.8975
3D-CNN [32]	0.9139	0.9186	0.8956	0.9175	0.9258	0.9163	0.9357	0.9142	0.8637
RAS	0.9217	0.9423	0.9189	0.9424	0.9537	0.9252	0.9575	0.9289	0.9013

TABLE 7. Experimental results of various attentional mechanisms. Red indicates the first under this evaluation criterion. Blue indicates the second under this evaluation criterion.

Methods	OA (IP)	AA (IP)	Kappa (IP)	OA (UA)	AA (UA)	Kappa (UA)	OA (Salinas)	AA (Salinas)	Kappa (Salinas)
RAS-no-SelfAttention	0.7552	0.7321	0.7189	0.8639	0.8420	0.8478	0.8829	0.8895	0.8542
RAS-Mutihead-SelfAttention	0.7837	0.7743	0.7531	0.8868	0.8733	0.8689	0.9175	0.9246	0.9128
RAS-Attention-1	0.7627	0.7637	0.7328	0.8785	0.8723	0.8645	0.8924	0.9054	0.8889
RAS-Attention-2	0.7689	0.7749	0.7329	0.8742	0.8692	0.8434	0.9053	0.9142	0.8963
RAS-Attention-1-2	0.7767	0.7752	0.7442	0.8859	0.8627	0.8567	0.9041	0.9135	0.8946
RAS-Attention-1-3	0.7653	0.7784	0.7335	0.8826	0.8465	0.8468	0.8945	0.9277	0.8912
RAS	0.7916	0.7800	0.7694	0.9082	0.8884	0.8719	0.9217	0.9423	0.9089

Pines, Pavia University, and Salinas datasets, respectively.

- In subsection *B*. We not only compare the single-headed attention mechanism with the multi-headed attention mechanism but also add some pooling layers to the attention mechanism for comparison.
- In subsection *C*. We compare feedforward with 1 × 1 convolution and also compare 2 × 2 convolution and 3 × 3 convolution to demonstrate the advantage of 1 × 1 convolution.
- In subsection *D*. We perform ablation experiments on the whole model to prove that our spatial information module, PE, SIE and spectral information module are effective. Detailed comparative data are shown in Tables 7-9. we perform a large number of ablation experiments to prove the validity of our model.

A. EXPERIMENTAL RESULTS OF DIFFERENT METHODS

1) RESULT ON THE INDIAN PINE DATASET

The Indian Pines dataset contains 16 classes of samples containing 10,249 pixels available for classification. Table 4 clearly shows that our algorithm outperforms both the traditional and recent algorithms. Our proposed model shows that fusing spatial and spectral information positively affects the classification results. BERT benefits from the information

TABLE 8. Experimental results of various SIE structures. Red indicates the first under this evaluation criterion. Blue indicates the second under this evaluation criterion.

Methods	OA (IP)	AA (IP)	Kappa (IP)	OA (UA)	AA (UA)	Kappa (UA)	OA (Salinas)	AA (Salinas)	Kappa (Salinas)
RAS-no-Convolution (1×1)	0.7550	0.7463	0.7418	0.8624	0.8635	0.8341	0.8939	0.9042	0.8978
RAS-FeedForward	0.7892	0.7737	0.7659	0.8841	0.8724	0.8552	0.9028	0.9154	0.9052
RAS-Convolution (3×3)	0.7793	0.7800	0.7532	0.8622	0.8745	0.8483	0.9124	0.9256	0.8887
RAS-Convolution (2×2)	0.7721	0.7849	0.7453	0.8723	0.8529	0.8587	0.9079	0.9175	0.8842
RAS	0.7916	0.8068	0.7694	0.9082	0.8884	0.8719	0.9217	0.9423	0.8989

TABLE 9. Experimental results of different submodules. Red indicates the first under this evaluation criterion. Blue indicates the second under this evaluation criterion.

Methods	OA (IP)	AA (IP)	Kappa (IP)	OA (UA)	AA (UA)	Kappa (UA)	OA (Salinas)	AA (Salinas)	Kappa (Salinas)
RAS-no-SpatialInformationModule	0.8417	0.8300	0.7956	0.9263	0.9178	0.9236	0.9117	0.9024	0.8828
RAS-no-SIE	0.8642	0.8568	0.8129	0.8587	0.9252	0.9262	0.9183	0.9063	0.8929
RAS-no-PE	0.8847	0.8429	0.8421	0.9343	0.9134	0.9144	0.9083	0.9131	0.9057
RAS-no-SpectralInformationModule	0.8243	0.8565	0.8374	0.9042	0.9054	0.9186	0.9229	0.8941	0.8842
RAS	0.8953	0.8942	0.8759	0.9452	0.9583	0.9441	0.9357	0.9430	0.9347

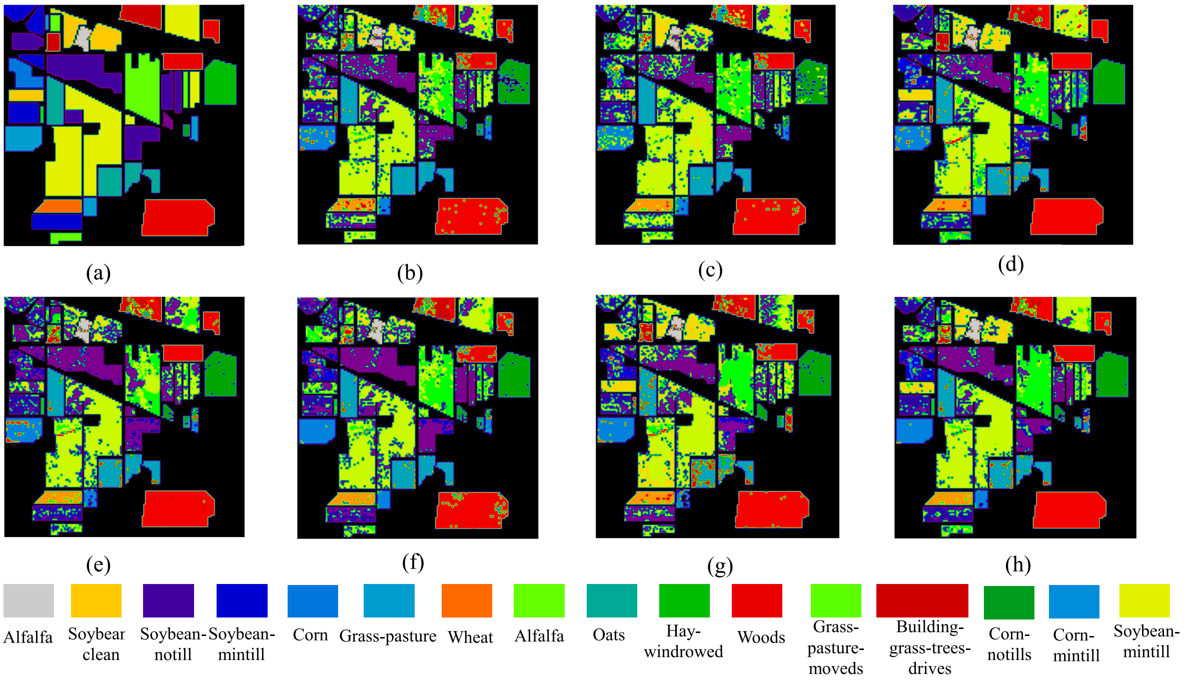

FIGURE 9. Indian Pine dataset classification results.(a) Ground truth (b) AlexNet (c) ResNet (d) DenseNet (e) Bert (f) 2D-CNN (g) 3D-CNN (h) RAS.

mining of the transformer in the full field of view. The final results are also very good and exceed our AA on the 10% training set by 0.6%, indicating that the transformer structure is quite important in image processing. RAS plays a prominent role in the feature mining and feature fusion processes. Figure 9 shows the classification results on the 1% training set. The experimental results demonstrate the strong performance of our proposed algorithm for spectral-spatial feature mining.

2) RESULT ON THE PAVIA UNIVERSITY DATASET

The largest difference between the Pavia University and Indian Pines datasets is that although the former includes only nine categories, the total amount of data is much larger than in the Indian Pines dataset. Table 5 shows that the OA, AA, and Kappa of our algorithm outperform those of the other algorithms. In addition, the prediction accuracy of our algorithm increases as the training set increases. 3D-CNN relies on powerful spatial feature extraction and thus performs well. Figure 10 shows the classification results on the 1% training set. These figures clearly show that the accuracy of classification increases as the number of training samples increases. The classification accuracy of the RAS algorithm is better than that of the other models.

3) RESULT ON THE SALINAS DATASET

The number of pixels in the Salinas dataset reaches 1,111,044, and there are 54,129 datasets available for

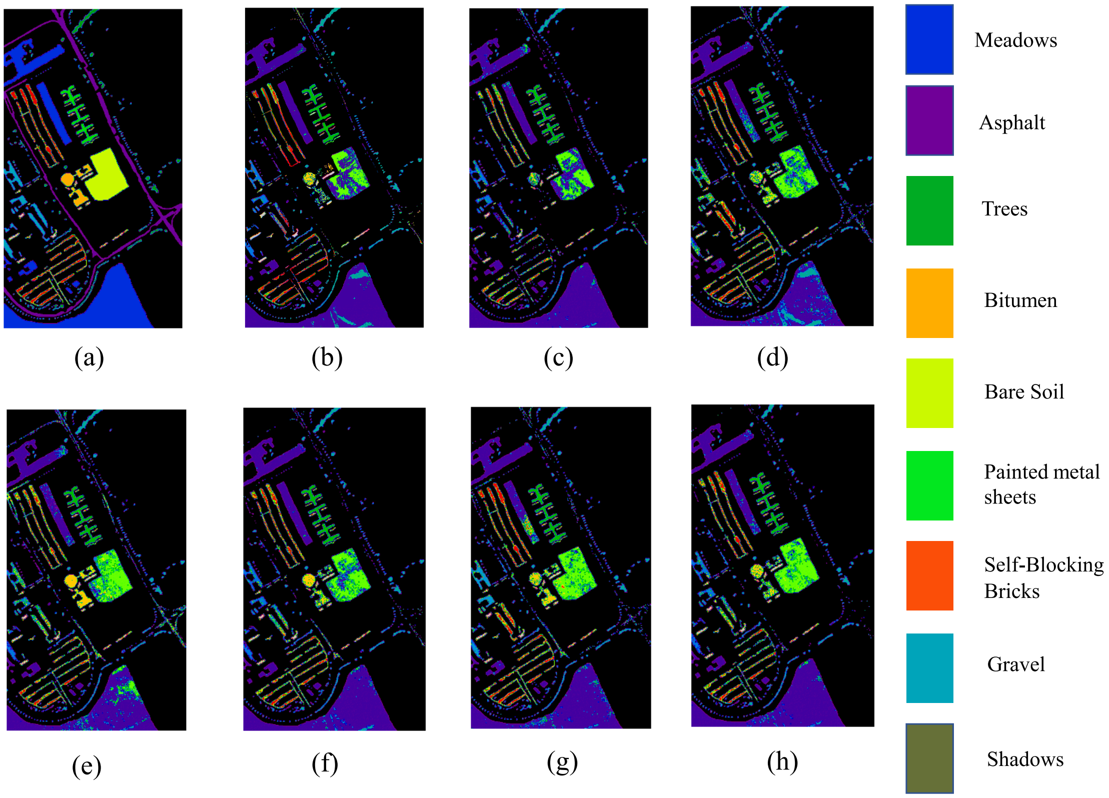

FIGURE 10. Indian Pine dataset classification results.(a) Ground truth (b) AlexNet (c) ResNet (d) DenseNet (e) Bert (f) 2D-CNN (g) 3D-CNN (h) RAS.

classification. Even using a 1% dataset as the training set, as clearly shown in Table 6, our algorithm still has a strong ability to mine information. Figure 11 shows the results of training set classification using only 1% of the total samples, clearly demonstrating that our algorithm is very powerful at mining and fusing spectral information with spatial information.

B. EXPERIMENTAL RESULTS OF DIFFERENT ATTENTIONAL MECHANISMS

In this section, we compare the effects of different types of attention mechanisms on the experimental results. We used 5% (Indian Pines), 1% (University of Pavia), and 1% (Salinas) from the three datasets as training sets. Here, a "1" represents the traditional attention mechanism; a "2" represents average pooling as the attention mechanism, and a "3" represents global average pooling as the attention mechanism. Multiple numbers represent operations using multiple simultaneous attention mechanisms. Table 7 clearly shows that the models with attention mechanisms are more accurate than models without attention mechanisms. The experimental results of the combined attention mechanism model are better than those of the single attention mechanism model and surpass the base model by 0.39% in Kappa (Salinas). The traditional attention mechanism model and the global average pooling attention mechanism model function better alone than combined because the global average pooling attention mechanism can mine only one feature; thus, the generalizability of the model is limited. The experimental results show that the self-attention mechanism we adopted performs sufficiently well for information mining.

C. EXPERIMENTAL RESULTS OF 1 × 1 CONVOLUTION AND FEEDFORWARD

In this section, we compare two spatial information mining structures: FeedForward and convolution. Table 8 shows the effects of several different sizes of convolution kernels and FeedForward on the results, showing that adding either Convolution or FeedForward has a positive effect on the results. The 3 × 3 convolution works better than does the 2 × 2 convolution: we hypothesize that odd convolution sizes can place the features in the center after convolution, while even convolution sizes force a trade-off for features. Because the 1 × 1 convolution does not destroy the spatial structure of the image, the experimental results of the 1 × 1 convolution are better than FeedForward.

D. EXPERIMENTAL RESULTS OF DIFFERENT SUBMODULES

To demonstrate whether each part of the model is valid, we do ablation experiments on the spatial information module, spectral information module, PE, and SIE, respectively. Table 9 shows the experimental results of the different

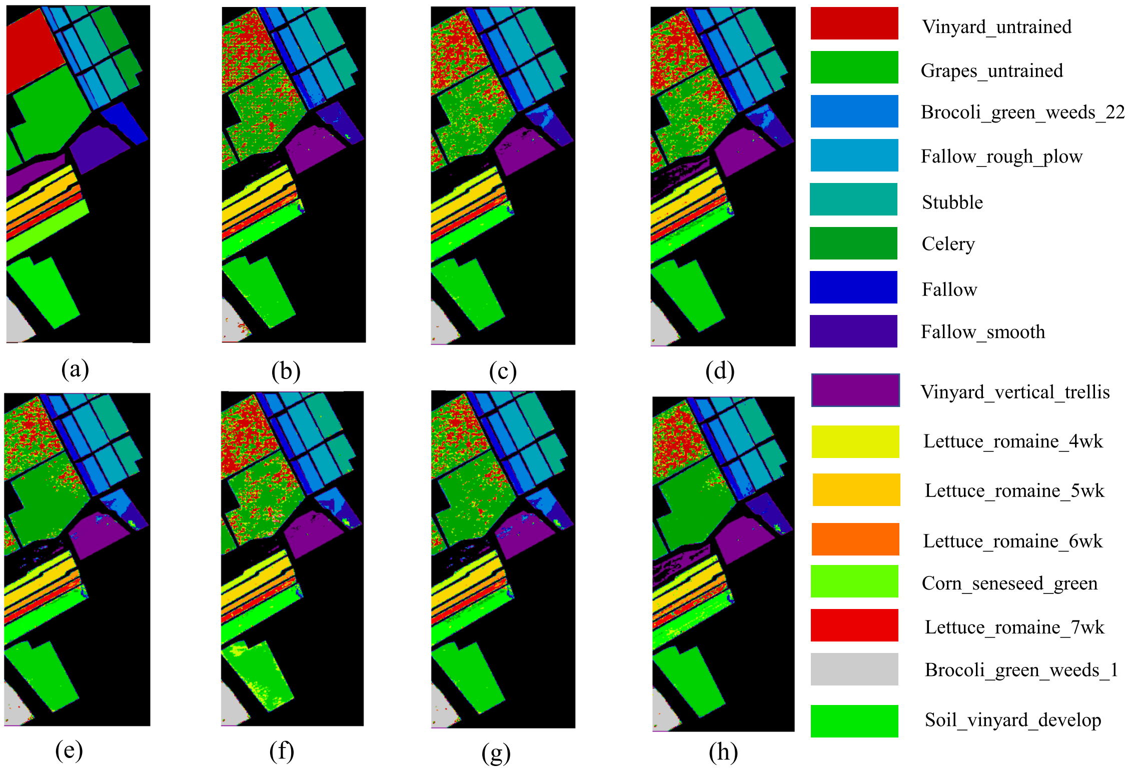

FIGURE 11. Indian Pine dataset classification results.(a) Ground truth (b) AlexNet (c) ResNet (d) DenseNet (e) Bert (f) 2D-CNN (g) 3D-CNN (h) RAS.

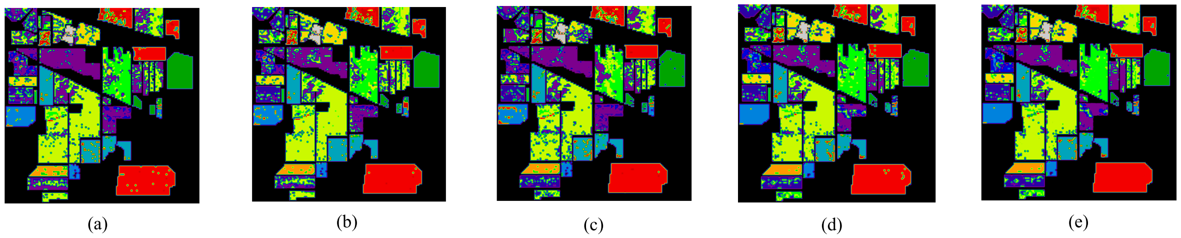

FIGURE 12. Indian Pine dataset classification results (a) RAS-no-SpatialInformationModule (b) RAS-no-SIE (c) RAS-no-PE (d)RAS-no-SpectralInformationModule (e) RAS.

submodules on the three datasets. In this case, we used 15% of the total Indian Pines samples, 10% of the total Pavia University samples, and 10% of the total Salinas samples for training. The experimental results indicate that each submodule plays an important role. In the Indian Pines dataset, our proposed SIE is important for constructing the relationships between pixels. On the Pavia University dataset. PE has little effect, and we speculate that it may be related to the fact that the vectors of hyperspectral pixels do not overlap. The experimental results do not differ substantially between the submodules, which we speculate may be due to the larger training dataset. In Salinas, useful information was retained and redundant information was eliminated. SIE plays a critical role in mining spatial information and the ResNet model played also an extremely important role in this process. The experimental results show that the prediction accuracy declines significantly when the SIE is removed. The sample classification results are shown in the following images (Figure 12: Indian Pines; Figure 13: Pavia University; Figure 14:Salinas dataset).

V. DISCUSSION

In this study, we fused spectral and spatial information to build a powerful model. Effectively, the problem of inaccurate pixel classification caused when only spectral information is used is addressed. However, some aspects still need further improvement. For example, for dataset dividing, although random dividing is more natural. However, it may cause some problems, for example, a category with little data is divided into fewer training parts, which can lead to imperfect features learned by the model and thus be unfavorable for classification.

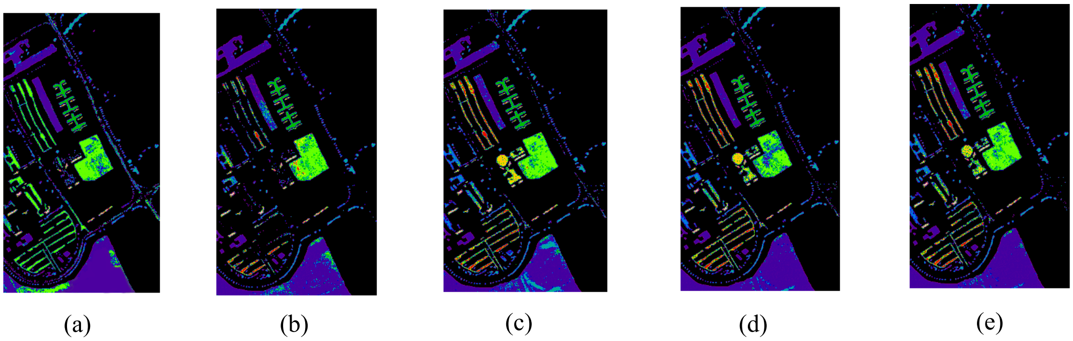

FIGURE 13. Pavia University data set classification results (a) RAS-no-SpatialInformationModule (b) RAS-no-SIE (c) RAS-no-PE (d)RAS-no-SpectralInformationModule (e) RAS.

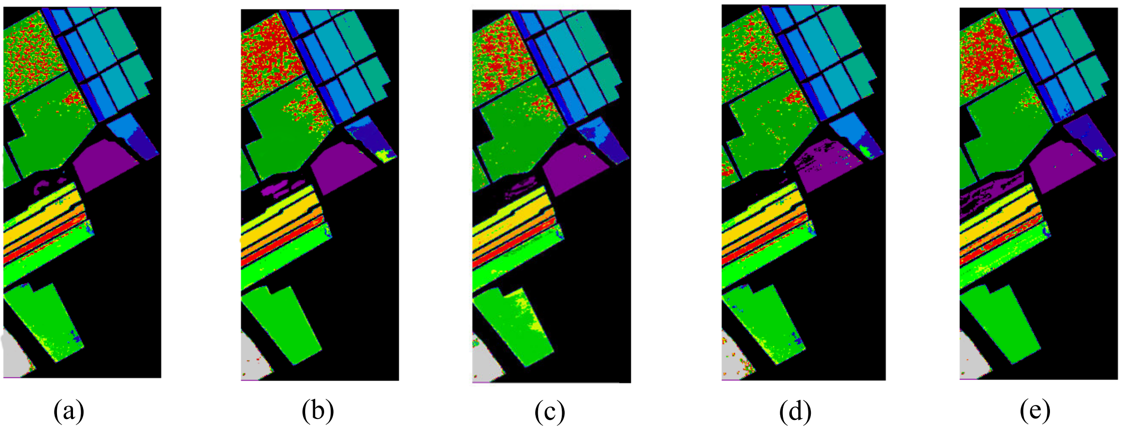

FIGURE 14. Salinas dataset classification results (a) RAS-no-SpatialInformationModule (b) RAS-no-SIE (c) RAS-no-PE (d)RAS-no-SpectralInformationModule (e) RAS.

VI. CONCLUSION

In this paper, we propose a joint spectral-spatial model called RAS. The spectral features are mined using a spectral information extractor. We exploit the label dependence property in hyperspectral images and use the surrounding pixels of pixels to be classified as spatial information. These pixels are then fed into SIE together for training, thus mining the association between pixels used as a complement to the spectral information. The two kinds of information are fully retained by feature fusion and used for the final classification. Ablation experiments are conducted by performing on multiple datasets. Our model relatively effectively solves the problems of weak classification ability using only spectral information and limited receptive fields. Finally, our algorithm achieves relatively good results on three widely used public datasets.

ACKNOWLEDGMENT

The authors would like to thank the Editor-in-Chief, the Associate Editor, and the Reviewers for their insightful comments and suggestions. The authors would also like to thank Lijing Tu for her teaching and support, and the lab members for their company and encouragement (Fang, Ji, Su, Wu, and so on).

REFERENCES

[1] G. A. Shaw and H.-H. K. Burke, "Spectral imaging for remote sensing," *Lincoln Lab. J.*, vol. 14, no. 1, pp. 3–28, 2003.

[2] J. J. Liu and J. F. MacGregor, "On the extraction of spectral and spatial information from images," *Chemometric Intell. Lab. Syst.*, vol. 85, no. 1, pp. 119–130, Jan. 2007.

[3] T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, and J. J. Sousa, "Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry," *Remote Sens.*, vol. 9, no. 11, p. 1110, 2017.

[4] G. Lu and B. Fei, "Medical hyperspectral imaging: A review," *J. Biomed. Opt.*, vol. 19, no. 1, 2014, Art. no. 010901.

[5] Y.-Z. Feng and D.-W. Sun, "Application of hyperspectral imaging in food safety inspection and control: A review," *Crit. Rev. Food Sci. Nutrition*, vol. 52, no. 11, pp. 1039–1058, 2012.

[6] W. Lv and X. Wang, "Overview of hyperspectral image classification," *J. Sensors*, vol. 2020, pp. 1–13, Jul. 2020.

[7] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 12–16, Jan. 2002.

[8] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. M. 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.

[9] A. Zare and K. Ho, "Endmember variability in hyperspectral analysis: Addressing spectral variability during spectral unmixing," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 95–104, Jan. 2014.

[10] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Econ. Geogr.*, vol. 46, pp. 234–240, Jun. 1970.

[11] V. Slavkovikj, S. Verstockt, W. De Neve, S. Van Hoecke, and R. Van de Walle, "Hyperspectral image classification with convolutional neural networks," in *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2015, pp. 1159–1162.

[12] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, vol. 9142. Bellingham, WA, USA: International Society for Optics and Photonics, Jul. 2015, pp. 4959–4962.

[13] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297–2307, May 2010.

[14] P. Zhong and R. Wang, "Learning conditional random fields for classification of hyperspectral images," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1890–1907, Jul. 2010.

[15] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O. L. García-Navarrete, and J. Blasco, "Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment," *Food Bioprocess Technol.*, vol. 5, no. 4, pp. 1121–1142, 2012.

[16] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.*, vol. 9, no. 8, p. 848, 2017.

[17] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020.

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

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[20] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1747–1756.

[21] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T.-Y. Liu, "Incorporating BERT into neural machine translation," 2020, *arXiv:2002.06823*. [Online]. Available: http://arxiv.org/abs/2002.06823

[22] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin, "End-to-end open-domain question answering with BERTserini," 2019, *arXiv:1902.01718*. [Online]. Available: http://arxiv.org/abs/1902.01718

[23] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "CommonsenseQA: A question answering challenge targeting commonsense knowledge," 2018, *arXiv:1811.00937*. [Online]. Available: http://arxiv.org/abs/1811.00937

[24] H. Xu, B. Liu, L. Shu, and P. S. Yu, "BERT post-training for review reading comprehension and aspect-based sentiment analysis," 2019, *arXiv:1904.02232*. [Online]. Available: http://arxiv.org/abs/1904.02232

[25] J. Li, Z. Tu, B. Yang, M. R. Lyu, and T. Zhang, "Multi-head attention with disagreement regularization," in *Proc. EMNLP*, 2018, pp. 2897–2903.

[26] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser, "Universal transformers," 2018, *arXiv:1807.03819*. [Online]. Available: http://arxiv.org/abs/1807.03819

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[28] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2017.

[29] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, nos. 10–12, pp. 1489–1506, Jun. 2000.

[30] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[31] H. Gao, S. Lin, Y. Yang, C. Li, and M. Yang, "Convolution neural network based on two-dimensional spectrum for hyperspectral image classification," *J. Sensors*, vol. 2018, pp. 1–13, Aug. 2018.

[32] B. Liu, X. Yu, P. Zhang, X. Tan, R. Wang, and L. Zhi, "Spectral–spatial classification of hyperspectral image using three-dimensional convolution network," *Proc. SPIE*, vol. 12, no. 1, 2018, Art. no. 016005.

[33] Y. Jiang, Y. Li, and H. Zhang, "Hyperspectral image classification based on 3-D separable ResNet and transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1949–1953, Dec. 2019.

[34] G. Yang, U. B. Gewali, E. Ientilucci, M. Gartley, and S. T. Monteiro, "Dual-channel densenet for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2595–2598.

[35] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.

[36] C. Tao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2438–2442, Dec. 2015.

[37] J. Zhao, Y. Zhong, T. Jia, X. Wang, Y. Xu, H. Shu, and L. Zhang, "Spectral-spatial classification of hyperspectral imagery with cooperative game," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 31–42, Jan. 2018.

[38] W. Ma, H. Ma, H. Zhu, Y. Li, L. Li, L. Jiao, and B. Hou, "Hyperspectral image classification based on spatial and spectral kernels generation network," *Inf. Sci.*, vol. 578, pp. 435–456, Nov. 2021.

[39] H. Zhu, M. Ma, W. Ma, L. Jiao, S. Hong, J. Shen, and B. Hou, "A spatial-channel progressive fusion ResNet for remote sensing classification," *Inf. Fusion*, vol. 70, pp. 72–87, Jun. 2021.

[40] W. Ma, J. Zhao, H. Zhu, J. Shen, L. Jiao, Y. Wu, and B. Hou, "A spatial-channel collaborative attention network for enhancement of multiresolution classification," *Remote Sens.*, vol. 13, no. 1, p. 106, Dec. 2020.

ZHUANGZHUANG HE (Student Member, IEEE) was born in Anhui, China. He is currently pursuing the degree in computer science with Anhui Agriculture University. His research interests include deep learning and recommendation systems.

HAO WU (Student Member, IEEE) is currently pursuing the degree in computer science and technology with the School of Information and Computer Science, Anhui Agricultural University. His research interests include machine learning, natural language processing, and human–computer interaction systems.

GUODONG WU (Member, IEEE) received the master's degree in computer application technology from Hefei University of Technology and the Ph.D. degree in intelligent decision making and knowledge management from Donghua University. He is currently an Associate Professor and the Head of the Department of Computer Science, Anhui Agricultural University. He is mainly engaged in the teaching and research of deep learning (graph neural networks), recommendation systems, and intelligent business. He has presided over and participated in nearly 20 provincial, ministerial, and national research projects, published more than 40 articles in academic journals at home and abroad, authorized more than 20 invention patents or registered software copyrights, edited two 13th Five-Year Plan textbooks, and participated in three textbooks. He is a member of China Computer Society, the Director of Anhui Computer Society, the Executive Director of Anhui Agricultural Informatization Association, and a Post Expert of Anhui Agricultural Informatization Industry Technology System.

• • •