IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 5, 2022, accepted January 13, 2022, date of publication January 18, 2022, date of current version January 27, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3144393

Spatial Attention Guided Residual Attention Network for Hyperspectral Image Classification

NINGYANG LI AND ZHAOHUI WANG
School of Computer Science and Technology, Hainan University, Haikou 570228, China
Corresponding author: Zhaohui Wang (william_hig@163.com)

This work was supported in part by the Framework of the Norwegian Research Council INTPART Project under Grant 309857 International Network for Image-Based Diagnosis (INID), and in part by the Hainan Key Research and Development Plan for Scientific and Technological Collaboration Projects: Research on Medical Imaging Aided Diagnosis of Infant Brain Development Diseases.

ABSTRACT Hyperspectral image (HSI) classification has become a research hotspot. Recently, deep learning-based methods have achieved preferable performances by which the deep spectral-spatial features can be extracted from HSI cubes. However, in complex scenes, due to the diversity of the types of land-cover and the bands in high dimensional, these methods are often hampered by the irrelevant spatial areas and the redundant bands, which results in the indistinguishable features and the restricted performance. In this article, a spatial attention guided residual attention network (SpaAG-RAN) is proposed for HSI classification, which contains a spatial attention module (SpaAM), a spectral attention module (SpeAM), and a spectral-spatial feature extraction module (SSFEM). Based on the spectral similarity, the SpaAM is capable of capturing the relevant spatial areas composed of the pixels of the same category as the center pixel from HSI cube with a novel inverted-shifted-scaled sigmoid activation function. The SpeAM aims to select the bands which are beneficial to the spectral features representation. The SSFEM is exploited to extract the discriminating spectral-spatial features. To facilitate the processes of bands selection and features extraction, two well-designed spatial attention masks generated by the SpaAM are employed to guide the works of the SpeAM and the SSFEM, respectively. Moreover, a spatial consistency loss function is installed to maintain the consistency between the two spatial attention masks so that the network enables the distinction of the relevant features exactly. Experimental results on three HSI data sets show that the proposed SpaAG-RAN model can extract the discriminating spectral-spatial features and outperforms the state-of-the-arts.

INDEX TERMS Hyperspectral image classification, attention mechanism, deep learning, spatial attention, residual network.

I. INTRODUCTION

Technological advancements of hyperspectral sensors and aircraft enable hyperspectral images (HSIs) to characterize the meticulous spectral features from the visible to the short-wave infrared wavelength ranges of land-cover with hundreds of contiguous bands. Furthermore, the improved spatial resolution of sensors brings HSI with possible richer spatial structures [1]. HSI classification, which aims to assign one certain category for each pixel using its spectral and spatial features [2], has captured attention increasingly in the field of HSI analysis [3]–[6]. It has shown great importance in remote sensing applications, such as precision agriculture, mineralogy, military reconnaissance, etc.

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

During the past decades, numerous classification methods based on the spectral characteristic have been conducted to classify the pixels in HSI data, including support vector machines (SVMs) [7], [8], k-nearest neighbors [9], random forest (RF) [10], logistic regression [11], extreme learning machine [12], etc. However, the classification results tend to be unsatisfactory when these methods encounter the pixels with very similar spectral features but not the identical label, since the advantage of spatial information has seldom been considered. Afterwards, quite a few methods have employed spatial features as an auxiliary means on the basis of spectral features to enhance the representation of hyperspectral data. For instance, to refine the classification maps predicted by SVMs, Markov random field and edge-preserving filtering are applied to take spatial contextual information into consideration, respectively [13], [14]. Morphological

profile [15]–[17], which is declared to be an efficient way to explore the spatial information, has been extended to adapt the spatial feature extraction of high dimensional hyperspectral data. In [18] and [19], the spatial information of the neighborhoods of each pixel is delivered to a sparse representation model to gain the optimal representation strategy with a set of common training samples. Furthermore, other approaches, such as Gabor filtering [20], compressive sensing [21], and discriminant analysis [22], [23], are also applied for HSI classification with the aid of the spectral and spatial features.

Although the aforementioned methods have achieved acceptable performances, the classification accuracy of them heavily depend on the quality of the hand-crafted features which are considered as the shallow features. Owing to the Hughes phenomenon [24] and the limited samples, the results of above-mentioned shallow models are prone to overfitting. Moreover, there is generally intense spectral variability in HSI due to the capricious environmental factors, which causes the large intra-class distance and the serious inter-class similarity. Consequently, hand-crafted features are no longer appropriate to deal with these problems. Extracting the robust spectral and spatial features for HSI classification has been a widely recognized demand by the related industry.

With the increasing computational ability of hardware, deep learning (DL) have made tremendous breakthrough in computer vision (CV) tasks (e.g., image classification [25], [26], scene segmentation [27], target detection [28]) and natural language processing [29], etc. A variety of DL-based attempts have been performed to process hyperspectral remote sensing images, which can be divided into two general categories, i.e. spectral-based methods and spectral-spatial-based methods, according to the type of the processed information. The spectral-based methods exploit spectral information only. For example, a full dimensional spectrum was input into the artificial neural network to discover the subtle spectral differences between different classes [30]. Instead of processing each band independently, a recurrent neural network (RNN) was utilized to take full advantage of the spectral correlation exists in the particular bands [31]. In order to alleviate the computational burden from the redundant information between the neighboring bands and increase the classification accuracy, a cascaded RNN model consists of two RNN layers was proposed, in which one aims to reduce redundancy whereas the other is used to learn complementarity [32]. With the trait of long-term dependence, the gradients may fade away during the training phase of RNN. Therefore, to handle this shortage, long short-term memory networks, as an extended version of RNN, are proposed to gain the contextual spectral features effectively [33], [34]. Although these spectral-based methods have improved the classification accuracy, there is still much room for improvement in their performances in complex scenes.

Different from the former, the spectral-spatial-based methods aim to extract the spectral and spatial features simultaneously for classification. Up to now, many studies have been carried out on this thinking. In [35], principle component analysis (PCA) was used to compress the spectral dimension, then every pixel and the flatten vector of the corresponding neighborhoods were sent to the multi-layer stacked autoencoders to extract spectral and spatial features, respectively. Reference [36] proposed a spatial updated deep autoencoder in which the contextual information was considered to maximize the interclass distances during feature learning. Besides, deep belief networks were also applied to capture the representative spectral features and count the statistics of neighboring pixels [37], [38]. However, these models mostly transform the spatial inputs into the flat vectors, which may destroy the spatial structure.

With the unique advantages of local perception and parameters sharing, convolutional neural networks (CNNs) have been demonstrated the power of feature extraction and dominated the field of HSI classification. The classical two-branch CNN architectures, including 1-D CNN and 2-D CNN, were designed to extract the spectral and spatial features, then accomplished the classification via the feature fusion or decision fusion strategies [39]–[41]. To further reserve the complete spectral-spatial information, an HSI cube which contains the center pixel and its neighborhoods was picked as the training sample of network. Such an approach assumes that the label of entire HSI cube can be represented by the label of center pixel because of the intensive spectral similarity existing between the center pixel and the surrounding pixels in a small region. Supported by this hypothesis, 3-D CNN has been the most appropriate network to fully extract the spectral-spatial features [41]. Moreover, 3-D CNN was united with Jeffries-Matusita distance to select effective bands for the recognition of very similar objects [42]. Aiming to address the issues of massive parameters and long-term training, those 3-D convolutional layers at deep positions are substituted by 2-D convolutional layers to simplify networks and fuse features at different levels effectively [43], [44]. As we all know, the deeper the network is, the more abstract and representative the features extracted. However, the deeper network may result in the vanishing gradient. To resolve this problem, a residual network (ResNet) was proposed to propagate the gradient from high layers to low layers quickly via the shortcut connections in residual blocks [45]. Zhong *et al*. [46] designed a deep spectral-spatial ResNet which contains serial residual blocks to alleviate the declining-accuracy phenomenon. In [47], a pyramidal bottleneck residual block was proposed to involve more feature map locations in the deeper network. Zhang *et al*. [48] combined the spectral-spatial fractal ResNet with data balance augmentation to improve the recall rates of the small samples. In addition, to enhance the robustness of model in unusual scene, a dual-channel ResNet with a noise-robust loss function was proposed to fully utilize the useful information from mislabeled samples [49]. With the aforementioned great progress, ResNet has been the mainstream architecture of the spectral-spatial-based methods for HSI classification.

However, there still exists a common drawback that has yet to be resolved. HSI usually contains abundant spectral and spatial information, whereas not all of them are beneficial to the classification [50]. In other words, the spectral bands and the salient spatial regions, which are beneficial to feature representation and classification, are supposed to be emphasized. To this end, attention mechanism, which is well received in nature machine translation [51] and CV tasks [52], [53], has been introduced to capture the most salient bands and positions in HSIs. Among the related applications, attention mechanism generally is embedded into the networks as an independent block to refine the feature maps by weighting bands, pixels, or channels unequally. For example, in the early stages, the lightweight spectral attention modules composed of the global average pooling (GAP) layers and the convolutional layers were placed at the beginning of the networks to promote the influential spectral bands to play a primary role in the subsequent feature extraction [54], [55]. Besides the spectral attention module, the spatial attention module was also proposed to enhance the significance of the relevant spatial regions. For example, Shamsolmoali [56] *et al.* employed a spatial attention module to increase the discriminating ability of network during the feature fusion. By embedding the spectral attention and spatial attention modules into the residual blocks sequentially, the useful spectral-spatial features are obtained to improve the classification performances [50], [57]–[59]. However, the spectral and spatial attention modules in the above-mentioned methods are processed independently, which hinder the complementation of spectral and spatial properties. In order to strengthen the correlation between spectral and spatial attentions, similarity matrices generated by the spectral and spatial attention branches were distributed to all locations and bands adaptively [60]. Specially, Li [61] *et al.* proposed a spectral and spatial fused attention module to apply the attention masks crosswise, which aims to fully explore the correlation between spectral bands, spatial positions, neighborhoods and the prediction results. In addition, self-attention (SA) was also adopted to explore the correlations between pixels. An SA model was designed to extract discriminating spectral and spatial features [62]. To enhance the effect of the center pixel, reference [63] proposed a method to compute the correlations between the center pixel and its neighborhoods. By integrating multiple SA modules, the spectral and spatial transformers [64], [65] were employed to model the correlation between the spectral bands and spatial locations. The transformer was also used to gain the optimized inputs for the subsequent processing [66]. However, the computational cost is enormous as there are generally several SA modules in the transformer. In [67] and [68], the global salient spectral bands and spatial areas are extracted by the spectral and spatial nonlocal blocks. Both are embedded into the spectral and spatial modules to refine the features, respectively.

Although the aforementioned models have somehow improved the classification results, there is a common deficiency that the extracted spatial attention may not focus on the positions related to the center pixel, especially in the scene of various junction land-cover. As shown in Fig. 1, there may be more than one type of land-cover in an HSI cube. But only the pixels which belong to the same category as the center pixel (marked with a yellow dot) are worthy of highlighting. In this article, these pixels are named as homogeneous pixels which are surrounded by green polygons, whereas the rest are named as interfering pixels surrounded by purple polygons. As the spectral characteristics are fully or partially different from that of the center pixel, interfering pixels may mislead to the irrelevant spatial regions and restrict the extraction of distinguishable spectral-spatial features to some extent. On the contrary, homogeneous pixels, which express the similar spectral properties with the center pixel, can lessen the impact of large intra-class variety and promote features aggregation. Therefore, the inherent functional differences between the homogeneous pixels and the interfering pixels should be fully considered for a better classification performance.

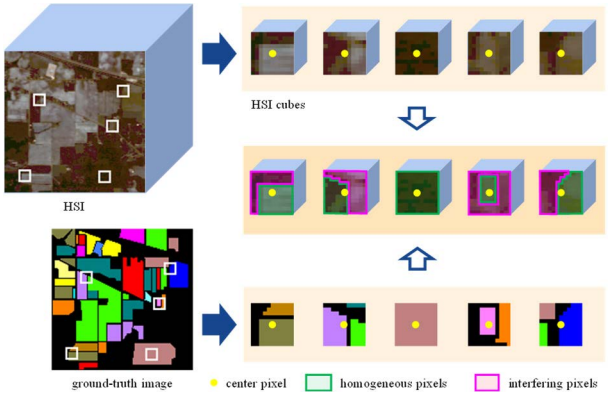

FIGURE 1. Before the 3-D CNN-based model training, each labeled pixel and its neighborhoods are cropped into an HSI cube to extract the spectral-spatial features. However, as shown in the corresponding ground truth images, there may be more than one type of land-cover in an HSI cube. Those pixels which have the same category as the center pixel are named as homogeneous pixel, whereas the rest are called interfering pixels. For convenience, the unlabeled pixels, i.e. the black pixels in ground-truth images, are assumed to the interfering pixels as well. In an HSI cube, homogeneous pixels form the relevant spatial areas, which are beneficial to the spectral-spatial feature extraction.

In order to achieve the above-mentioned purpose, a spatial attention guided residual attention network (SpaAG-RAN) is proposed to highlight the relevant spatial areas and extract the discriminating spectral-spatial features for HSI classification. The proposed model is mainly composed of the spatial attention module (SpaAM), the spectral attention module (SpeAM), and the spectral-spatial feature extraction module (SSFEM). Based upon the spectral similarities between the center pixel and its neighborhoods, the SpaAM is performed to generate the spatial attention masks efficiently which can represent the spatial distribution of homogeneous pixels and interfering pixels. Similarly, the SpeAM is designed to explore the spectral attention mask, which can be interpreted as an adaptive band selector. The SSFEM is a 3-D CNN with residual blocks, which takes charge of extracting the spectral-spatial features for classification. Among the three

modules, the SpaAM guides the other two modules so as to strengthen the effects of the relevant spatial areas. Specifically, in the SpaAM, one spatial attention mask is exploited to encourage the homogeneous pixels to contribute more for the selection of discriminating bands, whereas the other is used to suppress the interfering pixels from the feature maps extracted by the SSFEM. Besides, to identify the homogeneous pixels and the interfering pixels during the spectral-spatial feature extraction, a spatial consistency loss function is utilized to maintain the consistency of spatial attention masks generated before and after the SSFEM. Experimental results on three public HSI data sets demonstrate the effectiveness of the SpaAM and the superior classification performance of the proposal.

The main contributions of this article are as follows.

1) A lightweight spectral-similarity-based SpaAM is designed to capture the relevant spatial areas, which describes the spatial distribution of homogeneous pixels and interfering pixels implicitly. In this module, the spectral similarities between the center pixel and its neighborhoods are measured by the efficient Euclidean distance. A novel inverted-shifted-scaled sigmoid activation function is then in charge of converting the similarities to the proper spatial weights.

2) To improve the classification performance, a spatial consistency loss function is conducted to enable the SSFEM to extract effective features by preserving the specificity of homogeneous pixels and interfering pixels.

3) An end-to-end SpaAG-RAN model, which incorporates the SpaAM, the SpeAM, and the SSFEM, is proposed to stress the relevant spatial areas and extract the discriminating spectral-spatial features for HSI classification.

The remainder of this article is organized as follows. Section II introduces the proposed SpaAG-RAN model in detail. Section III presents the experimental results and analyses on three classical data sets. Finally, this article is concluded in Section IV.

II. METHODLOGY

In this section, the overview of the proposed SpaAG-RAN model is first introduced. Then, the core components of the network, including the SpaAM, the SpeAM, and the SSFEM, are described at length. Finally, the loss functions of the network and the optimization processes are given.

A. FRAMEWORK OF THE PROPOSED NETWORK

Suppose that the HSI data set $\mathcal{H} \in \mathbb{R}^{h\times w\times b}$ contains N labeled pixels $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\} \in \mathbb{R}^{1\times 1\times b}$ and their corresponding one-hot label vectors $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\} \in \mathbb{R}^{1\times C}$, where h, w, b, and C represent the height and width of spatial dimension, the number of bands, and the number of categories, respectively. Previous researches [41], [42] have demonstrated the effectiveness of spectral-spatial features for classification. Therefore, a square box with the width of ω is borrowed to crop the center pixels and their adjacent pixels to form the HSI cubes $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{\omega\times\omega\times b}$. The label of the ith HSI cube $\mathbf{x}_i$ is assumed to be $\mathbf{y}_i$, i.e. the label of the center pixel. After these preprocessing, the combination of all HSI cubes and the corresponding one-hot label vectors ($\mathbf{X}$, $\mathbf{Y}$) forms the sample set. In this work, a certain proportion of samples are selected randomly from each land-cover category to train the network, whereas the rest are used as validation set and test set.

Fig. 2 shows the workflow of the proposed SpaAG-RAN model, which is mainly composed of the SpaAM, the SpeAM, and the SSFEM. The inputs are an HSI cube $\mathbf{x} \in \mathbb{R}^{1\times\omega\times\omega\times b}$ and the corresponding true label $\mathbf{y}$. First, $\mathbf{x}$ is fed into the SpaAM to gain the incipient spatial attention mask $\mathbf{M_a^i} \in \mathbb{R}^{1\times\omega\times\omega\times 1}$, which is utilized to highlight the homogeneous pixels and suppress the interfering pixels in $\mathbf{x}$. Next, the spectral attention mask $\mathbf{M_e} \in \mathbb{R}^{1\times 1\times 1\times b}$, which contains the influential spectral bands for distinguishing the center pixel from the interfering pixels, is extracted by the SpeAM from the calibrated HSI cube $\mathbf{x_a^i} \in \mathbb{R}^{1\times\omega\times\omega\times b}$. With the uneven weighting operation of $\mathbf{M_e}$, the contributory bands in $\mathbf{x}$ are emphasized. Then, the processed HSI cube $\mathbf{x_e} \in \mathbb{R}^{1\times\omega\times\omega\times b}$ is transported to the SSFEM to extract the discriminating spectral-spatial features $\mathbf{x}' \in \mathbb{R}^{c\times\omega\times\omega\times b'}$, where c and b' are the number of channels (i.e. convolutional filters) and the number of reduced bands. Before the classification, $\mathbf{x}'$ is input to the SpaAM to acquire the terminal spatial attention mask $\mathbf{M_a^t} \in \mathbb{R}^{1\times\omega\times\omega\times 1}$, which is used to suppress the interfering pixels in $\mathbf{x}'$ and compute the spatial consistency loss L_{sc} with the incipient spatial attention mask $\mathbf{M_a^i}$. Finally, the fully connected (FC) layer map the refined features $\mathbf{x_a^t} \in \mathbb{R}^{c\times\omega\times\omega\times b'}$ to the classification space and predict the most possible label $\mathbf{y}'$ with the softmax activation function.

During the training process, the proposed network optimizes the parameters with the cross-entropy L_{ce} and the spatial consistency L_{sc} loss functions. The cross-entropy, as the universal loss function for the classification problem, is adopted to minimize the cross-validation error between the true label $\mathbf{y}$ and the predicted label $\mathbf{y}'$. Moreover, to maintain the stability of the spatial distribution of homogeneous pixels and interfering pixels, the spatial consistency loss function is installed to monitor the variation between the two spatial attention masks, i.e. $\mathbf{M_a^i}$ and $\mathbf{M_a^t}$. The details of the above three modules and the optimization of loss functions are illustrated as follows.

B. SpaAM

The SpaAM is designed to capture the spatial areas relevant to the center pixel. These spatial areas are promoted to take a leading role during the spectral attention generation and feature extraction. In order to reach these purposes, a natural idea is to analyze the spectral similarities between the center pixel and its neighboring pixels. The higher the similarity between one pixel and the center pixel is, the more possible they belong to the same category. Thus, this pixel should be assigned a greater weight to be emphasized. Considering that the spectral characteristics of land-cover in HSI data generally vary frequently under the influence of various environmental conditions (e.g. temperature and humidity), the

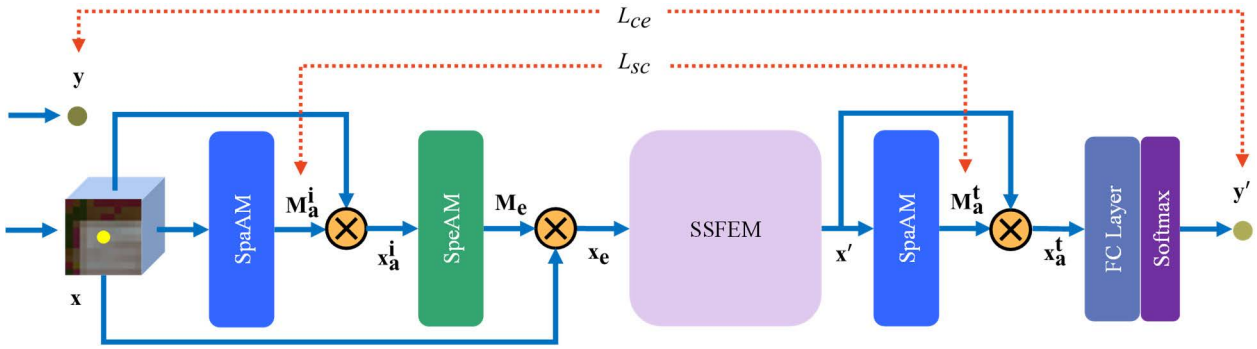

FIGURE 2. Overview of the proposed SpaAG-RAN model. The SpaAM aims to emphasize the relevant spatial areas. The SpeAM aims to discover the discriminating bands for the better spectral feature representation. The SSFEM is exploited to extract the deep spectral-spatial features. During the back propagation procedure, the cross-entropy loss function L_{ce} and the spatial consistency loss function L_{sc} are adopted to optimize the parameters of the network. "⊗" denotes the element-wise multiplication.

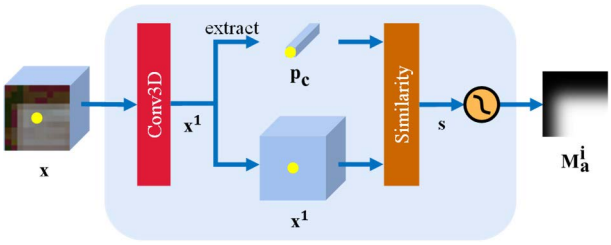

FIGURE 3. SpaAM. It contains a 3-D convolutional layer, a similarity calculation layer, and an inverted-shifted-scaled sigmoid activation function.

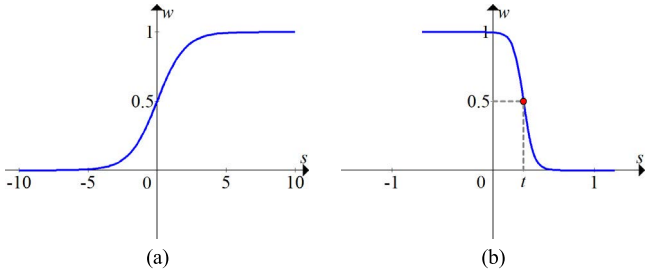

FIGURE 4. (a) Sigmoid activation function. (b) Inverted-shifted-scaled sigmoid activation function ($\alpha = 20$, $t = 0.3$). Where s and w represent the similarity and weight, respectively.

Euclidean (L2) distance, which is not sensitive to the sharp variations [69], is adopted to perform the process.

The architecture of the SpaAM is shown in Fig. 3. Given an HSI cube $\mathbf{x}$, the SpaAM aims to generate the spatial attention mask $\mathbf{M_a^i}$ (for convenience, the incipient spatial attention mask $\mathbf{M_a^i}$ is described as an example).

First, the convolutional layer with a $1 \times 1 \times 1$ kernel is employed to reduce the channels of the input to one:

$$\mathbf{x^1} = \mathbf{w^1} * \mathbf{x} + \mathbf{b^1} \tag{1}$$

where $\mathbf{x^1} \in \mathbb{R}^{1\times\omega\times\omega\times b}$ is the single-channel feature map, $\mathbf{w^1} \in \mathbb{R}^{1\times1\times1\times1}$ and $\mathrm{b}^1 \in \mathbb{R}^{1\times1\times1\times1}$ are the convolutional kernel and bias of the convolutional layer, separately. "$*$" is the convolutional operator.

Then, the L2 distance is adopted to evaluate the similarities $\mathbf{s} \in \mathbb{R}^{1\times\omega\times\omega\times1}$ between all pixels in $\mathbf{x^1}$ and the center pixel $\mathbf{p_c}$ which is copied from $\mathbf{x^1}$. The similarity at position (i, j) is calculated by

$$\mathbf{s}(i, j) = ||\mathbf{p_c} - \mathbf{x^1}(i, j)||_2 \tag{2}$$

For a pixel, the higher the similarity it gains, the more it contributes to the classification, and vice versa. Considering the range of the values of $\mathbf{s}$ is $[0, +\infty)$, the lower value represents the higher similarity. An inverted-shifted-scaled sigmoid activation function $\mathcal{R}$ is proposed to allocate a proper weight for each pixel. As shown in Fig. 4, compared with the standard sigmoid activation function, the distribution of the weights of $\mathbf{M_a^i}$ is adjusted by $\mathcal{R}$ via flipping, panning, and zooming:

$$\mathbf{M_a^i} = \mathcal{R}(\mathbf{s}) = \frac{1}{1 + e^{\alpha\cdot(\mathbf{s}-t)}} \tag{3}$$

where α is the scale to regulate the ceiling of the weights of pixels, t corresponds to the similarity value when the weight is 0.5 which is the threshold to divide the similarity intervals. The pixels, which similarity is in the range of $[0, t]$, are regarded as the homogeneous pixels, whereas the other pixels with the similarity in the range of $(t, +\infty)$ are seen as the interfering pixels.

After the above processes, an element-wise multiplication across the spatial dimension between the mask $\mathbf{M_a^i}$ and the HSI cube $\mathbf{x}$ is conducted to stimulate the homogeneous pixels to contribute more for the selection of the important spectral bands:

$$\mathbf{x_a^i} = \mathbf{M_a^i} \otimes \mathbf{x} \tag{4}$$

Similarly, following the SSFEM, the SpaAM is also used to extract the terminal spatial attention mask $\mathbf{M_a^t}$ from $\mathbf{x}'$ with the identical processes. With the aid of $\mathbf{M_a^t}$, the interfering pixels of the output of the SSFEM are weakened by

$$\mathbf{x_a^t} = \mathbf{M_a^t} \otimes \mathbf{x}'. \tag{5}$$

C. SpeAM

The SpeAM is designed to select the discriminating spectral bands which are beneficial to the spectral feature representation of the center pixel. Similar to the SpaAM, the spectral attention mask $\mathbf{M_e}$, which gives a particular weight for each band, is generated by the SpeAM.

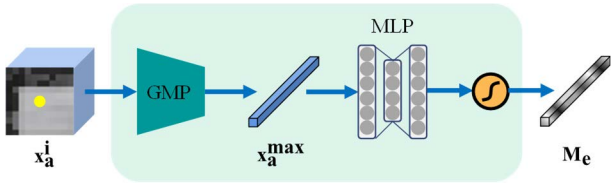

FIGURE 5. SpeAM. It contains a GMP layer, an MLP, and a sigmoid activation function.

The structure of the SpeAM which is borrowed from the previous research [70] is shown in Fig. 5. In the input of the SpeAM, the calibrated HSI cube $\mathbf{x}_\mathbf{a}^\mathbf{i}$, the homogeneous pixels are highlighted and the interfering pixels are weakened. Therefore, a global max-pooling (GMP) layer with the $\omega \times \omega \times 1$ pooling size instead of the GAP layer of the original architecture is first exploited to retain the most notable and germane information in each spectral band. The maximum element of the ith band of $\mathbf{x}_\mathbf{a}^\mathbf{i}$ is calculated by

$$\mathbf{x}_\mathbf{a}^\mathbf{max}(i) = \max(\mathbf{x}_\mathbf{a}^\mathbf{i}(i)) \tag{6}$$

where $\mathbf{x}_\mathbf{a}^\mathbf{max} \in \mathbb{R}^{1\times1\times b}$ is the output of the GMP layer.

Then, $\mathbf{x}_\mathbf{a}^\mathbf{max}$ is delivered to an MLP to explore the collaborative and exclusive relationships between spectral bands. The MLP contains two FC layers. The first FC layer aims to reduce the information redundancy and compress the critical spectral features, whereas the second one converts the abstract compressed features to the spectral attention mask with the aid of the sigmoid activation function σ

$$\mathbf{M_e} = \sigma(\mathbf{w_2} \cdot \mathrm{ReLU}(\mathbf{w_1} \cdot \mathbf{x}_\mathbf{a}^\mathbf{max} + \mathbf{b_1}) + \mathbf{b_2}) \tag{7}$$

where $\mathbf{w_1} \in \mathbb{R}^{(b/r)\times b}$, $\mathbf{b_1} \in \mathbb{R}^{(b/r)\times 1}$, $\mathbf{w_2} \in \mathbb{R}^{b\times(b/r)}$, and $\mathbf{b_2} \in \mathbb{R}^{b\times1}$ are the weight parameters and the biases of the first and second FC layers under the compression ratio of r.

Finally, an element-wise multiplication across the spectral dimension between the mask $\mathbf{M_e}$ and the HSI cube $\mathbf{x}$ is conducted to emphasize the discriminating spectral bands for the spectral-spatial feature extraction as follow

$$\mathbf{x_e} = \mathbf{M_e} \otimes \mathbf{x}. \tag{8}$$

where $\mathbf{x_e}$ is the HSI cube after the bands enhancement.

D. SSFEM

The SSFEM is built to extract the deep spectral-spatial features for classification. The architecture of the SSFEM is based on the CNN, which has been the most popular network for many CV tasks.

Generally, a classic CNN contains convolutional layers, activation functions, and pooling layers. The convolutional layers are in charge of feature extraction. The activation functions are followed to the map features to the nonlinear space. The pooling layers are used to compress the features.

In a classic CNN, the ith feature map of the lth convolutional layer can be formulated as follows:

$$\mathbf{F}_i^l = \mathcal{A}(\mathbf{w}_i^l * \mathbf{F}_j^{l-1} + \mathrm{b}_i^l) \tag{9}$$

where $\mathbf{F}_j^{l-1}$ is the jth output feature map of the $(l-1)$th layer, $\mathbf{w}_i^l$ and b_i^l are the ith convolutional kernel and the bias of the lth layer, respectively. "$*$" is the convolutional operator. $\mathcal{A}$ is an activation function, such as rectified linear unit (ReLU) [71], sigmoid function, and hyperbolic tangent function. In this article, ReLU is employed due to its advantages in efficient gradient propagation and sparse activation.

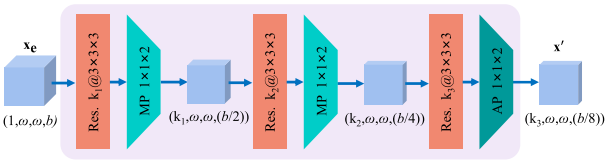

FIGURE 6. SSFEM. Where "Res.", "MP", and "AP" denote the residual block, max pooling, and average pooling, respectively.

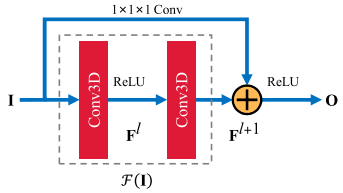

FIGURE 7. Residual block (Res.). Where "⊕" denotes the element-wise aggregation.

However, when processing the HSI data, the deeper the CNN is, the accuracy decrease. This phenomenon occurs because the classification errors from deep layers cannot be propagated back precisely, which results in the vanishing gradient. To overcome this problem, ResNet [45], is proposed to add a shortcut connection between the input volume and the output volume, which enables the network to stacks with any depth. Thus, the vanishing gradient is alleviated and the network could be optimized easily.

In view of the above-mentioned inimitable advantages of ResNet, it is adopted as the basic architecture of SSFEM. As shown in Fig. 6, by receiving the refined output of the SpeAM, i.e. $\mathbf{x_e}$, the SSFEM contains three residual blocks is exploited to extract the spectral-spatial features $\mathbf{x}'$. In this module, residual blocks equipped with the convolutional kernels of $3 \times 3 \times 3$ are connected in series to learn the deep feature representations. The numbers of convolutional kernels in three residual blocks are $\{k_1, k_2, k_3\}$, respectively. As shown in Fig. 7, the residual block can be represented briefly as follow:

$$\mathbf{O} = \mathrm{ReLU}(\mathbf{F}^{l+1}) \tag{10}$$

$$\mathbf{F}^{l+1} = \mathcal{F}(\mathbf{I}) + \mathbf{I} \tag{11}$$

$$\mathcal{F}(\mathbf{I}) = \mathbf{w}^{l+1} * \mathbf{F}^{l} + \mathbf{b}^{l+1} \tag{12}$$

$$\mathbf{F}^{l} = \mathrm{ReLU}(\mathbf{w}^{l} * \mathbf{I} + \mathbf{b}^{l}) \tag{13}$$

where $\mathbf{I}$ and $\mathbf{O}$ are the input and output of residual block, $\mathbf{F}^i$, $\mathbf{w}^i$, and $\mathbf{b}^i$ are the feature maps, the convolutional kernels, and the biases of the ith layer, respectively. $\mathcal{F}$ represents a series of operations (in dotted rectangle), including convolution and activation. Note that a convolutional layer with a kernel in size of $1 \times 1 \times 1$ is used in the shortcut connection to match the dimensions between $\mathbf{I}$ and $\mathbf{F}^{l+1}$.

Behind the first two residual blocks, the max pooling (MP) layers are followed to stress the intensive information and reduce redundancy. The average pooling (AP) layer is set after the last residual block to remain as much semantic information as possible [45]. For all pooling layers, the sizes and strides are all set to $1 \times 1 \times 2$ to preserve more spatial information.

Finally, the deep spectral-spatial features $\mathbf{x}' \in \mathbb{R}^{k_3 \times \omega \times \omega \times (b/8)}$ is obtained by

$$\mathbf{x}' = \mathcal{S}(\mathbf{x_e}) \tag{14}$$

where $\mathcal{S}$ represents the SSFEM.

Table 1 displays the details of the layers in the proposal.

TABLE 1. Detailed architecture of the proposed SpaAG-RAN model.

Module	Layer Name	Kernel Size	Strides	Connected to	Output Shape
	Input	—	—	—	$(1,\omega,\omega,b)$
SpaAM	Conv3D_1	1@1×1×1	(1,1,1)	Input	$(1,\omega,\omega,b)$
	Extract_1	—	—	Conv3D_1	$(1,1,1,b)$
	Similarity_1	—	—	Extract_1, Conv3D_1	$(1,\omega,\omega,1)$
	ISS_Sigmoid_1	—	—	Similarity_1	$(1,\omega,\omega,1)$
	Multiply_1	—	—	Input, ISS_Sigmoid_1	$(1,\omega,\omega,b)$
SpeAM	GMP	$\omega \times \omega \times 1$	$(\omega,\omega,1)$	Multiply_1	$(1,1,1,b)$
	Flatten_1	—	—	GMP	$(b,)$
	MLP_FC_1	b/r	—	Flatten_1	$(r,)$
	MLP_FC_2	b	—	MLP_FC_1	$(b,)$
	Reshape	—	—	MLP_FC_2	$(1,1,1,b)$
	Multiply_2	—	—	Input, Reshape	$(1,\omega,\omega,b)$
SSFEM	Residual_Block_1	k_1@3×3×3	(1,1,1)	Multiply_2	(k_1,ω,ω,b)
	MP_1	1×1×2	(1,1,2)	Residual_Block_1	$(k_1,\omega,\omega,b/2)$
	Residual_Block_2	k_2@3×3×3	(1,1,1)	MP_1	$(k_2,\omega,\omega,b/2)$
	MP_2	1×1×2	(1,1,2)	Residual_Block_2	$(k_2,\omega,\omega,b/4)$
	Residual_Block_3	k_3@3×3×3	(1,1,1)	MP_2	$(k_3,\omega,\omega,b/4)$
	AP	1×1×2	(1,1,2)	Residual_Block_3	$(k_3,\omega,\omega,b/8)$
SpaAM	Conv3D_2	1@1×1×1	(1,1,1)	AP	$(1,\omega,\omega,b/8)$
	Extract_2	—	—	Conv3D_2	$(1,1,1,b/8)$
	Similarity_2	—	—	Extract_2, Conv3D_2	$(1,\omega,\omega,1)$
	ISS_Sigmoid_2	—	—	Similarity_2	$(1,\omega,\omega,1)$
	Multiply_3	—	—	AP, ISS_Sigmoid_2	$(k_3,\omega,\omega,b/8)$
	Flatten_2	—	—	Multiply_3	$(k_3\times\omega\times\omega\times b/8)$
	Output(FC)	N	—	Flatten_2	$(N,)$

ISS: Inverted-Shifted-Scaled

E. LOSS FUNCTIONS AND OPTIMIZATION

To train the proposed SpaAG-RAN model effectively, the classification loss function L_{ce} is exploited together with the spatial consistency loss function L_{sc} to optimize the parameters.

Cross-entropy, as a popular loss function for the classification problems, is adopted to minimize the loss

$$L_{ce} = -\frac{1}{bs}\sum_{i=1}^{bs}\sum_{j=1}^{C}\mathbf{y}_{i,j}\log(\mathbf{y}'_{i,j}) \tag{15}$$

where $\mathbf{y}$ and $\mathbf{y}'$ are the true and predicted one-hot label vectors, respectively. C is the number of classes, bs is the number of samples in a batch, and $\mathbf{y}_{i,j}$ denotes the scalar of the jth class of the ith sample.

The spatial attention masks, $\mathbf{M}_\mathbf{a}^\mathbf{i}$ and $\mathbf{M}_\mathbf{a}^\mathbf{t}$, express the correlations between the center pixel and its neighborhoods implicitly. By preserving the consistency of the two spatial attention masks during the convolution, the feature extraction ability of the SSFEM to distinguish the homogeneous pixels from the interfering pixels is enhanced. To achieve the above-mentioned goal effectively, the mean absolute error is employed to measure the variation between the two masks, i.e. $\mathbf{M}_\mathbf{a}^\mathbf{i}$ and $\mathbf{M}_\mathbf{a}^\mathbf{t}$. The complete spatial consistency loss function L_{sc} on a batch is defined as follow:

$$L_{sc} = \frac{1}{bs}\sum_{i=1}^{bs}\sum\left|\mathbf{M}_\mathbf{a}^\mathbf{i} - \mathbf{M}_\mathbf{a}^\mathbf{t}\right| \tag{16}$$

Therefore, the total loss L can be formulated as

$$L = L_{ce} + \lambda L_{sc} \tag{17}$$

where λ controls the relative importance of the two functions. During the training procedure, the backpropagation and gradient descent algorithm are used to update parameters.

III. EXPERIMENTS AND ANALYSES

In this section, the details of three HSI data sets [72] collected by different imaging sensors, including Indian Pines (IP), University of Pavia (UP), and Botswana (BW), and the experimental configuration are described at length first. Then, the parameters setting of the network, the ablation study, and the comparison between the proposal and the state-of-the-art methods are reported and discussed. Finally, the visualization of the spatial attention masks and feature maps is presented and analyzed.

A. DATA SETS AND EXPERIMENTAL CONFIGURATION

The IP data set is gathered by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over Indian Pines test site in Northwestern Indiana in 1992. It consists of 145 × 145 pixels and 224 spectral bands in the wavelength range from 0.4 to 2.5 μm. The spatial and spectral resolutions are 20 meters/pixel (m/p) and 10 nm. After removing 20 bands covering the region of water absorption and four zero-bands, the remaining 200 bands are used for experiments. The false-color image of the IP data set and its ground-truth (GT) are shown in Fig. 8 (a) and (b). As illustrated in Table 2, 15%, 5%, and 80% of the labeled pixels are selected randomly from each of the 16 land-cover categories as the training, validation, and test sets, respectively.

The UP data set is acquired by the Reflective Optics Imaging Spectrometer (ROSIS) sensor during a flight campaign over Pavia, North Italy in 2002. It consists of 610 × 340 pixels and 115 spectral bands in the wavelength range from 0.43 to 0.86 μm. The spatial and spectral resolutions are 1.3 m/p and 4 nm, respectively. After removing 12 noisy

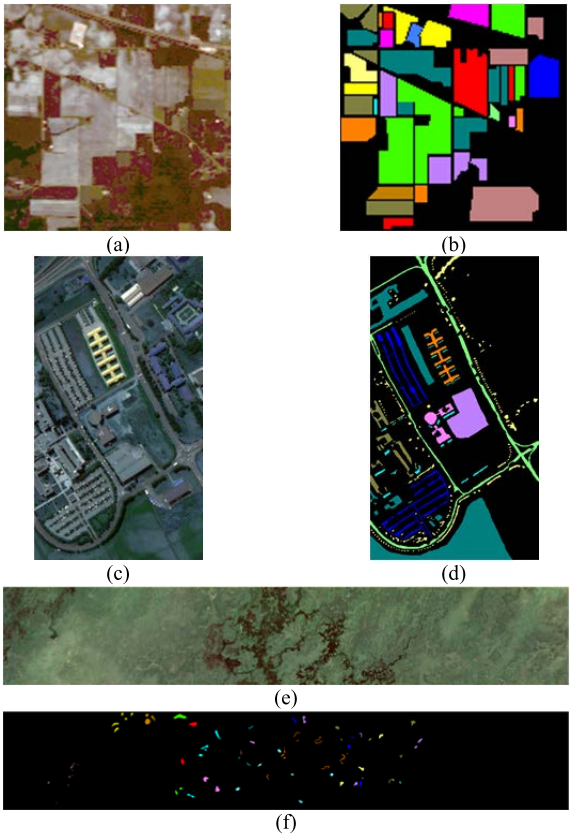

FIGURE 8. IP, UP and BW data sets: (a), (c) and (e) False-color images; (b), (d) and (f) Ground-truth maps.

bands, there are 103 bands remained for experiments. The false-color image and the corresponding ground-truth are shown in Fig. 8(c) and (d). As illustrated in Table 3, 5%, 5%, and 90% of the labeled pixels are selected randomly from each of the 9 land-cover categories as the training, validation, and test sets, respectively.

The BW data set is collected by the Hyperion sensor mounted on the Earth Observing-1 (EO-1) satellite over the Okavango Delta, Botswana, on 31st May, 2001. It consists of 1476 $\times$ 256 pixels and 242 spectral bands in the wavelength range from 0.4 to 2.5 μm. The spatial and spectral resolutions are 30 m/p and 10 nm. By removing the uncalibrated and noisy bands which cover water absorption features, there are 145 bands remained. The false-color image of the BW data set and the corresponding ground-truth are shown in Fig. 8(e) and (f). As illustrated in Table 4, 15%, 5%, and 80% of the labeled pixels are selected randomly from each of the 14 land-cover categories as the training, validation, and test sets, respectively.

The experiments on the above three data sets are performed on a computer with an AMD Ryzen 3600 at 4.07 GHz $\times$ 6 with 32-GB RAM and an NVIDIA GeForce GTX 1080Ti graphical processing unit (GPU) with 12-GB RAM. The operating system is Ubuntu 16.04. The DL frameworks are the Tensorflow and Keras. The optimization is executed by RMSprop optimizer [73] with the learning rate of 0.001

TABLE 2. Numbers of samples in the IP data set.

No.	Color	Land-cover type	Training	Validation	Test
1		Alfalfa	7	2	37
2		Corn-notill	214	71	1143
3		Corn-mintill	124	42	664
4		Corn	35	12	190
5		Grass-pasture	73	24	386
6		Grass-trees	109	37	584
7		Grass-pasture-mowed	5	1	22
8		Hay-windrowed	72	24	382
9		Oats	3	1	16
10		Soybean-notill	145	49	778
11		Soybean-mintill	368	123	1964
12		Soybean-clean	88	30	475
13		Wheat	31	10	164
14		Woods	190	63	1012
15		Building-Grass-Trees-Drives	58	19	309
16		Stone-Steel-Towers	14	5	74
		Total	1536	513	8200

TABLE 3. Numbers of samples in the UP data set.

No.	Color	Land-cover type	Training	Validation	Test
1		Asphalt	332	332	5305
2		Meadows	932	933	14919
3		Gravel	105	105	1679
4		Trees	153	153	2451
5		Painted metal sheets	67	67	1076
6		Bare Soil	251	251	4023
7		Bitumen	67	67	1064
8		Self-Blocking Bricks	184	184	2946
9		Shadows	47	47	758
		Total	2138	2139	34221

TABLE 4. Numbers of samples in the BW data set.

No.	Color	Land-cover type	Training	Validation	Test
1		Water	41	14	215
2		Hippo grass	15	5	81
3		Floodplain grasses1	38	13	200
4		Floodplain grassed2	32	11	172
5		Reeds1	40	13	216
6		Riparian	40	14	215
7		Firescar2	39	13	207
8		Island interior	31	10	162
9		Acacia woodland	47	16	251
10		Acacia shrublands	37	12	199
11		Acacia grasslands	46	15	244
12		Short mopane	27	9	145
13		Mixed mopane	40	13	215
14		Exposed soils	14	5	76
		Total	487	163	2598

and the delay factor of 0.9. The weights and biases of all layers in the proposed model are initialized by Xavier normal distribution [74]. The batch size is 16 and the total number of training iteration is 200.

In order to quantify the classification performance of the proposed SpaAG-RAN model, the overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) as the evaluation measures. The higher scores they get, the superior the performance of the model. All experiments are performed ten

times to relieve the impact of random initialized parameters and the average values are reported.

B. PARAMETERS SETTING

The structures of the DL-based models are generally complex. Models can have many hyperparameters and finding the best combination of parameters can be treated as an optimization problem. In this section, five parameters are analyzed to optimize the proposed model, including (1) the scale α and the threshold t of the inverted-shifted-scaled sigmoid activation function in the SpaAM, (2) the compression ratio r in the MLP of SpeAM, (3) the numbers $\{k_1, k_2, k_3\}$ of the convolutional kernels in the SSFEM, (4) the width ω of HSI cube, and (5) the proportion of training samples. For each parameter, the values bring the highest OAs on the test set to the network are fixed as the default settings in the next experiments.

1) SCALE α AND THRESHOLD t OF THE INVERTED-SHIFTED-SCALED SIGMOID ACTIVATION FUNCTION

The inverted-shifted-scaled sigmoid activation function aims to assign a rational spatial weight for each pixel. The scale α determines the range of weight whereas the threshold t can be regarded as the boundary between the homogeneous pixels and the interfering pixels. In order to ascertain the correlation between the two parameters and the classification performance, the two parameters, α and t, of inverted-shifted-scaled sigmoid activation function are set to {1, 2, 5, 10, 20, 50, 100, 500} and {0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}, respectively. The surface charts of the OAs on three data sets are shown in Fig. 9. Taken the IP data set as an example, from the axis of t, the OAs are not good enough when the value of t is less than 0.2 or larger than 0.6. From the axis of α, the OAs are also in an inferior position as α is less than 10. However, as t moves away from the both ends gradually, the α larger than 10 obtain the superior classification performance.

Analyzing the above-mentioned results with reference to the inverted-shifted-scaled sigmoid activation function shown in Fig. 4(b), several conclusions can be derived as follows. First, when the value of t is smaller, although most interfering pixels are shielded, a large part of homogeneous pixels are also treated as interfering pixels, which causes the inadequate feature extraction. Second, when the value of t is close to 1, the pixels involved in classification contain not only homogeneous pixels but also many interfering pixels, which also influences the classification accuracy. Last, the larger value of α enables the ceiling of the weights of homogeneous pixels to approach 1 and clarifies the boundary between the two kinds of pixels, which are both beneficial for classification. The data points {(20, 0.3), (10, 0.4), (20, 0.5)}, which are marked with the yellow ellipses corresponding to the maximum OAs on the IP, UP, and BW data sets, respectively.

2) COMPRESSION RATIO r IN THE MLP OF SPEAM

The SpeAM aims to strengthen the discriminating spectral bands. The MLP plays a key role in the dimensional reduction of features and nonlinear mapping between the spectral bands. To preserve the important spectral information and reduce the redundancy, it is necessary to adjust an apposite compression ratio. In this part, the effect of the compression ratio r in the MLP of SpeAM is analyzed. As shown in Fig. 10, the OAs of the proposed model on three data sets all reach the peak when the ratio is 2. When there is no compression (i.e. the ratio is 1), the OAs remain a lower level. In addition, as the compression ratio increases from 2 to 10, the declining-accuracy phenomenon occurs. This is because the reduced dimension leads to more spectral information to be abandoned gradually. However, the decline of the OAs on the IP data set is the most intense on three data sets. One pertinent reason is that the two hundreds of bands of the IP data set give rise to the more loss of spectral information under the same compression ratio comparing with other two data sets.

3) NUMBERS $\{k_1, k_2, k_3\}$ OF THE CONVOLUTIONAL KERNELS IN THE SSFEM

It has been approved by [25] there is a closer connection between the number of convolutional kernels and the

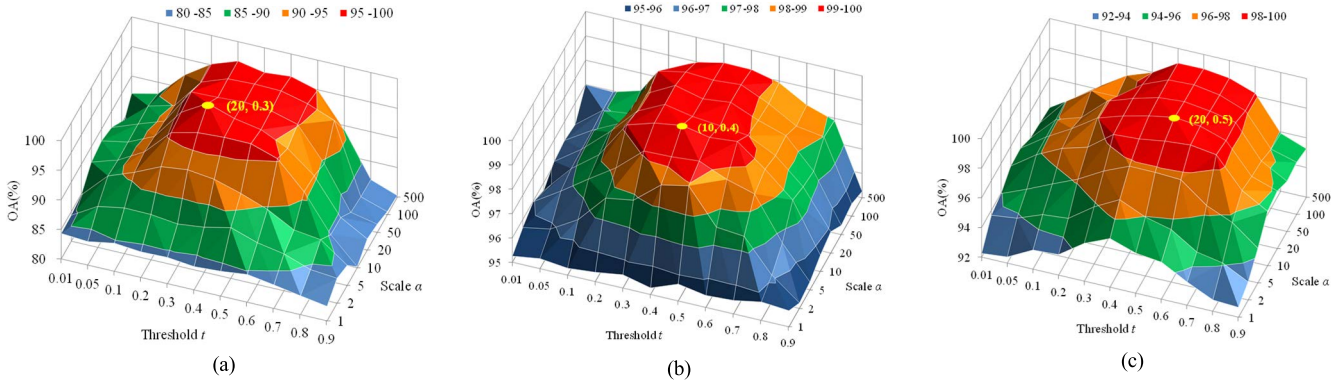

FIGURE 9. OAs (%) of different scales and thresholds of the inverted-shifted-scaled sigmoid activation function on three data sets. (a) IP. (b) UP. (c) BW.

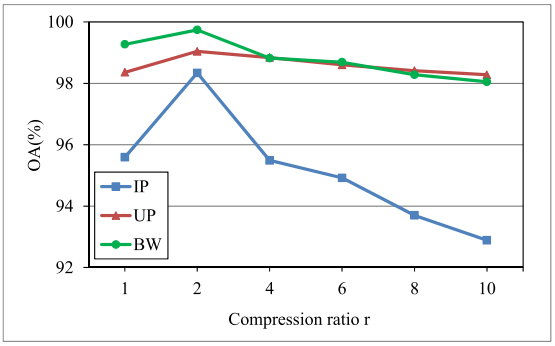

FIGURE 10. OAs (%) of different compression ratios *r* in the MLP of SpeAM on three data sets.

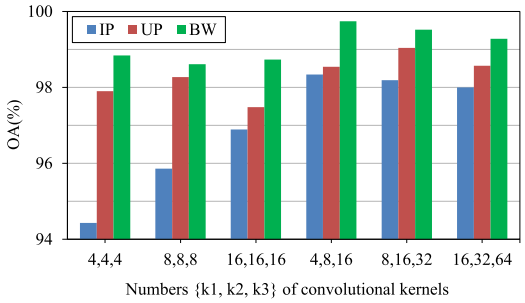

FIGURE 11. OAs (%) of different numbers $\{k_1, k_2, k_3\}$ of the convolutional kernels in the SSFEM on three data sets.

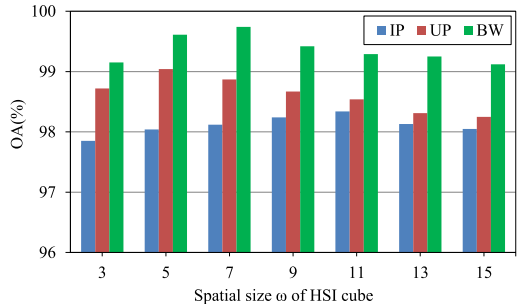

FIGURE 12. OAs (%) with different spatial sizes ω of HSI cubes on three data sets.

representational capability of features. In order to extract the sufficient spectral-spatial features efficiently, six experiments using SSFEMs with different numbers of convolutional kernels are deployed to explore their influences on the classification performance. It can be seen from Fig. 11 that the SSFEMs with the convolutional kernels of {4, 8, 16}, {8, 16, 32}, and {16, 32, 64}, which adopt the wider architecture to extract the expressive spectral-spatial features, achieve better accuracy than those of the {4, 4, 4}, {8, 8, 8}, and {16, 16, 16}. Therefore, the default numbers of convolutional kernels of the SSFEM are set to {4, 8, 16}, {8, 16, 32}, and {4, 8, 16} for the IP, UP, and BW data sets, respectively.

4) WIDTH ω OF HSI CUBE

The width of HSI cube has also a great effect on the classification performance. The larger width represents the more spatial information in HSI cube. But there may be more interfering pixels. Therefore, the HSI cubes with different widths {3, 5, 7, 9, 11, 13, 15} are input to the proposed SpaAG-RAN model to explore the proper widths. As shown in Fig. 12, when the widths of HSI cubes are 11, 5, and 7, the highest OAs are obtained on three data sets. The reason why the optimal widths of HSI cubes of the UP and the BW data sets are smaller than that of the IP data set most likely is that the spatial distributions of land-cover in the UP and the BW data sets are scattered and not as concentrated as the IP data set. On the other hand, the ranges of the undulation of the OAs on three data sets maintain below 1%, which demonstrates the robustness of the proposed model. This is due to the SpaAM can recognize the homogeneous pixels and interfering pixels precisely via the similarities between the center pixel and its neighborhoods. More important, the measurement of the similarities is independent of the width of the HSI cube.

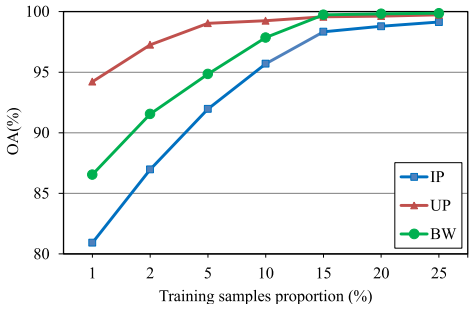

FIGURE 13. OAs (%) of different proportions of training samples on three data sets.

5) TRAINING SAMPLE PROPORTION

In this part, the performance of the proposal with different proportions of training samples is investigated. For each data set, {1%, 2%, 5%, 10%, 15%, 20%, 25%} of samples are randomly selected from each of land-cover categories as the training set. The experimental results are shown in Fig. 13. The OAs increase as the proportions of training samples on three data sets increasing. When the proportions of training samples of three data sets are more than 15%, 5%, and 15%, separately, the OAs will keep in high level.

C. ABLATION STUDY

In this section, two ablation studies are carried out, including the combination of different modules and the weight of the spatial consistency loss function. For each study, the values bring the highest OAs on the test set are adopted.

1) COMBINATION OF DIFFERENT MODULES

The proposed SpaAG-RAN model is composed of the SpaAM, the SpeAM, and the SSFEM. The SpaAM and the SpeAM aim to generate the spatial and spectral attention masks, whereas the SSFEM takes charge of extracting the deep spectral-spatial features. In order to explore the correlations between the three modules and the impacts of them on the classification performance, four schemes with different

TABLE 5. OAs (%) of different schemes on three data sets.

	Scheme_1	Scheme_2	Scheme_3	Scheme_4
IP	96.42	96.94	97.81	**98.34**
UP	96.83	97.26	98.64	**99.04**
BW	97.91	98.34	99.08	**99.74**

combinations of three modules are implemented on three data sets. The four schemes are as follows:

1) Scheme_1: SSFEM.
2) Scheme_2: SpeAM + SSFEM.
3) Scheme_3: SpaAM + SSFEM.
4) Scheme_4: SpaAM + SpeAM + SSFEM.

The OAs of these schemes on the test sets are presented in Table 5. The numbers reported in bold-type denote the best results for each data set. It can be seen that the classification accuracy of Scheme_1 achieve an acceptable level even though they are less than other schemes. Compared to Scheme_1, Scheme_2 brings in the SpeAM to emphasize the contributory bands, which elevates the OAs on three data sets by no less than 0.4%. More inspiring, by installing the SpaAM on the SSFEM, the OAs of Scheme_3 on the IP, UP, and BW data sets grow to 97.81%, 98.64%, and 99.08%, respectively. Scheme_4 is a complete SpaAG-RAN model. It receives the best classification accuracy on three data sets comparing with the other schemes. In the proposed SpaAG-RAN model, the SpaAM can be seen as the guide of the whole network, which highlights the homogeneous pixels and restrains the interfering pixels via the spatial attention masks. By the guidance of the SpaAM, the products of the SpeAM and the SSFEM are ameliorated for better classification.

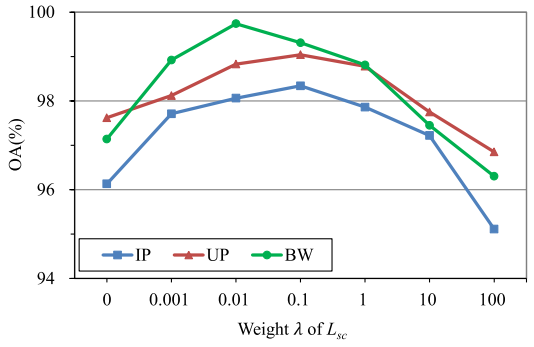

FIGURE 14. OAs (%) of different weights λ of L_{SC} on three data sets, where "0" indicates the spatial consistency loss function is removed and there is only the cross-entropy loss function reserved for optimization.

2) WEIGHT OF THE SPATIAL CONSISTENCY LOSS FUNCTION

Another study is conducted to confirm the availability of the proposed spatial consistency loss function L_{sc} and its contributions for classification by assigning different values to the weight λ. As shown in Fig. 14, when the value of λ is set to 0, i.e. the cross-entropy loss function works alone during the training process, the proposed model receives not less than 96%, 97%, and 97% OAs on three data sets, respectively. However, there are still rooms existed for improvement. After the spatial consistency loss function is installed, for the IP and UP data sets, the OAs reach the highest levels when the value of λ is set to 0.1 whereas the appropriate value of λ for the BW data set is 0.01. The reason might be the spatial information included in the BW data set is not as important as that in the other two data sets as shown in the corresponding ground-truth of them. With the further augment of the value of λ, the OAs start to reduce. Worse still, when the value of λ is set to 100, the OAs are even less than those of without using the function of L_{sc}.

Fig. 15 shows the error and accuracy curves during the training procedures of three data sets when the weight λ of the spatial consistency loss L_{sc} is set to the optimal values and zero. From the upper parts of Fig. 15a-f, the errors of the total loss L and the cross-entropy loss L_{ce} keep the steady levels with minute fluctuation when the number of iteration gets close to 200. For three data sets, the weights λ with the optimal values minimize the value of L_{sc} to the considerably low values. However, the weight λ with zero value results in L_{sc} in the high and gradually increasing levels, which destroys the consistency of the relevant spatial areas and causes the unsatisfactory classification performances.

Therefore, the key to fully apply the advantages of the spatial consistency loss function for classification is to regulate the balance between it and the cross-entropy loss function precisely. The larger weight of L_{sc} may cause the larger deviation of parameters optimization as well as the convergence of network been disturbed. On the contrary, by assigning a smaller weight, the function of L_{sc} will play the auxiliary role during the training procedure, which is more appropriate for the HSI classification missions.

D. COMPARISON WITH OTHER METHODS

To verify the effectiveness, the proposal is compared with two traditional classical methods: SVM with a radial basis function kernel and RF, and ten well discussed DL-based methods: 2-D CNN [41], 3-D CNN [41], spectral spatial residual network (SSRN) [46], spectral spatial attention network (SSAN) [62], center attention network (CAN) [63], double-branch multi-attention mechanism network (DBMA) [59], double-branch dual-attention mechanism network (DBDA) [68], 3-D cascaded spectral-spatial element attention network (CSSEAN) [67], residual spectral spatial attention network (RSSAN) [57], and rotation equivariant feature image pyramid network (REFIPN) [56]. For each method, the network from the original article is adopted. All methods share the same data sets (as illustrated in Section III-A) with the proposed model.

1) QUANTITATIVE COMPARISONS

The quantitative evaluations, including the recalls of each category as well as the means and standard deviations of OA, AA, and κ, are obtained by different methods on the IP, the UP, and the BW test sets. From Table 6 -VIII, several conclusions can be summarized as follows. First, the DL-based

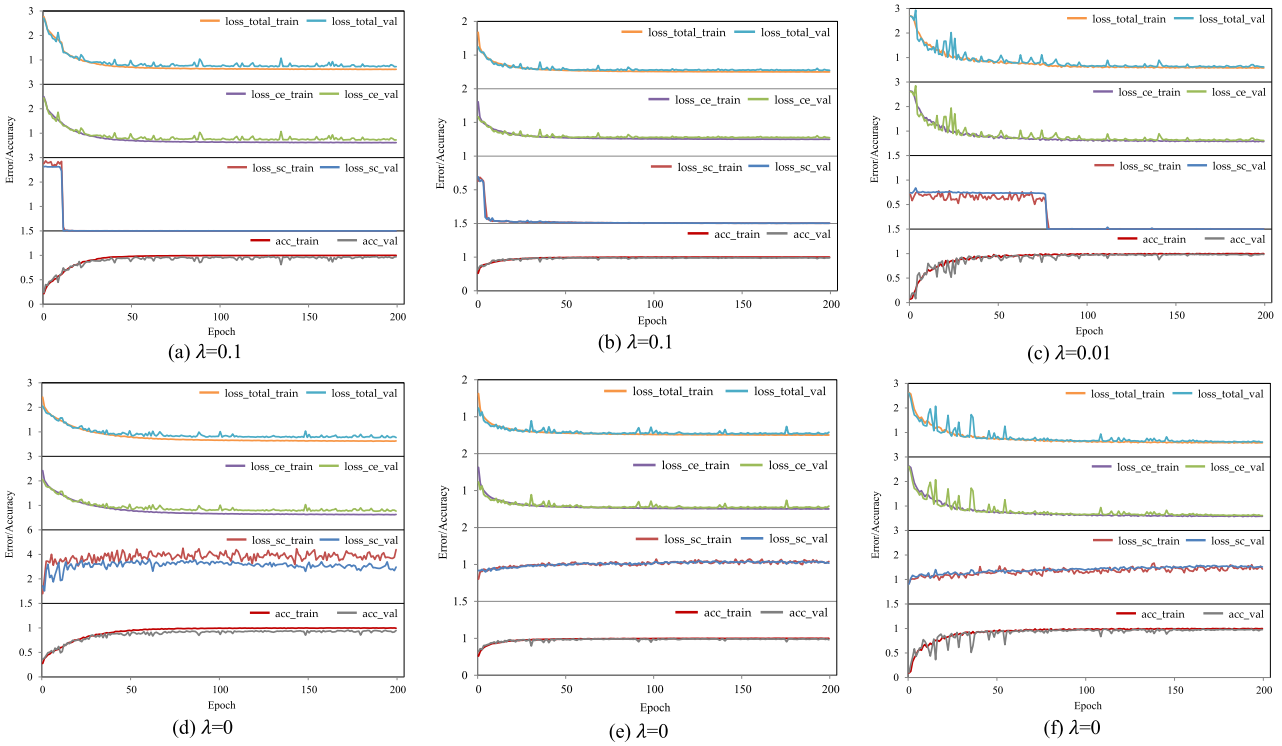

FIGURE 15. Error and accuracy curves of different weights of the spatial consistency loss function L_{SC} on the training and validation sets. (a) and (d) IP, (b) and (e) UP, (c) and (f) BW.

TABLE 6. Classification results (%) for the IP test set using 15% of the available labeled data.

No.	SVM	RF	2-D CNN	3-D CNN	SSRN	SSAN	CAN	DBMA	DBDA	CSSEAN	RSSAN	REFIPN	SpaAG-RAN
1	27.03	37.84	54.05	64.86	86.49	27.03	67.57	64.86	70.27	**97.30**	67.57	70.27	94.59
2	53.19	61.50	78.30	85.74	33.25	93.79	91.43	94.84	90.55	96.59	94.40	**98.08**	92.21
3	28.01	70.33	93.52	88.55	85.99	78.92	98.80	93.37	97.74	81.93	97.89	97.59	**99.25**
4	31.22	59.26	53.97	68.25	67.72	71.43	86.24	52.91	74.60	92.06	91.01	84.13	**97.35**
5	47.03	69.51	88.63	89.15	93.28	91.21	97.42	96.38	96.64	94.06	**100.00**	98.45	98.71
6	98.12	98.46	99.49	99.49	97.26	**100.00**	97.26	97.26	98.97	91.78	98.29	**100.00**	99.66
7	0.00	0.00	52.17	52.17	95.65	82.61	39.13	56.52	34.78	13.04	56.52	65.22	**100.00**
8	99.74	98.95	**100.00**	**100.00**	**100.00**	**100.00**	99.21	99.74	**100.00**	96.07	99.74	**100.00**	**100.00**
9	21.35	31.02	37.50	62.50	31.25	18.75	68.75	**100.00**	50.00	0.00	75.00	75.00	93.75
10	54.31	64.61	76.06	78.76	55.86	89.06	53.80	96.27	**97.43**	86.62	96.78	96.91	93.18
11	96.08	92.62	90.94	94.55	**99.80**	95.01	98.07	93.33	95.98	97.86	98.01	98.07	99.64
12	41.77	35.23	57.17	80.80	61.18	89.87	82.91	93.88	96.20	95.57	91.77	92.62	**97.47**
13	98.17	**100.00**	**100.00**	**100.00**	92.68	**100.00**	**100.00**	99.39	**100.00**	99.39	**100.00**	**100.00**	**100.00**
14	95.85	99.41	99.31	99.01	97.53	96.84	99.60	**99.90**	99.80	99.60	99.80	99.70	99.60
15	66.99	63.11	63.11	88.35	88.35	84.79	81.23	70.87	93.20	79.61	91.91	86.41	**93.53**
16	81.08	90.54	87.84	83.78	59.46	**100.00**	95.95	86.49	93.24	89.19	89.19	95.95	91.89
OA	73.14±1.57	81.96±0.57	91.83±0.81	93.57±0.45	94.25±0.22	94.07±1.33	94.30±0.59	94.97±0.63	96.11±1.06	89.81±0.16	96.85±0.46	97.05±0.13	**98.34±0.18**
AA	57.41±1.95	65.48±1.89	77.00±1.95	83.50±0.46	77.86±0.47	82.46±0.02	84.83±1.40	87.25±0.83	86.84±1.09	81.92±0.38	90.49±1.04	91.15±2.83	**96.93±0.47**
κ	67.13±0.94	75.11±0.85	93.78±0.69	88.90±0.24	77.12±0.72	90.88±0.36	89.72±1.14	92.53±0.91	94.88±0.73	92.46±0.31	96.24±0.53	94.02±0.72	**97.06±0.20**

methods all obtain better performances on three data sets comparing with the traditional methods. For instance, SVM mistakes all the pixels of the categories of ''Grass-pasture-mowed'' (No. 7) of the IP data set and the categories of ''Bitumen'' (No. 7) of the UP data set. Since the spectral information is used only, the SVM cannot fit the distributions of the two categories with limited samples. Similarly, RF, which relies on the elaborate hand-crafted features to finish the classification, also behaves the lower performances on three data sets. Second, comparing with 2-D CNN, the classification performance of 3-D CNN is improved on three data sets to some degree. This is because 2-D CNN only extracts the spatial features from the first principle component using PCA, whereas 3-D CNN exploits the rich spectral-spatial information for classification. Third, the other six being compared methods except RSSAN, which take 3-D CNN as baseline, achieve higher classification performance on three data sets. Specifically, in comparison with 3-D CNN, SSRN

TABLE 7. Classification results (%) for the UP test set using 5% of the available labeled data.

No.	SVM	RF	2-D CNN	3-D CNN	SSRN	SSAN	CAN	DBMA	DBDA	CSSEAN	RSSAN	REFIPN	SpaAG-RAN
1	95.96	94.54	94.96	95.16	98.76	95.26	**99.83**	98.06	99.06	97.07	98.17	98.26	99.40
2	99.83	98.31	98.15	97.84	88.44	99.49	99.21	99.34	99.73	99.90	98.80	99.72	**99.92**
3	8.89	2.96	53.84	58.02	90.47	96.08	94.07	81.95	91.11	**97.72**	95.18	83.85	97.30
4	91.15	94.78	90.10	94.13	98.51	98.62	**98.95**	95.79	96.30	98.19	94.92	97.06	96.19
5	99.17	98.60	98.27	99.75	**100.00**	**100.00**	**100.00**	**100.00**	99.92	**100.00**	**100.00**	**100.00**	99.92
6	21.94	69.60	75.02	87.01	**100.00**	89.42	99.09	96.53	92.89	90.97	99.91	97.95	99.18
7	0.00	18.31	12.54	31.69	94.90	96.15	70.48	**97.24**	79.10	97.58	85.79	96.07	96.07
8	95.47	98.13	96.35	97.77	96.44	77.76	83.34	**99.31**	94.60	90.77	93.90	98.79	97.22
9	83.57	86.60	98.24	99.06	99.88	**100.00**	**100.00**	**100.00**	97.30	99.88	99.88	98.83	99.65
OA	77.41±1.20	89.75±0.75	91.21±0.68	93.88±0.12	95.08±0.21	94.89±1.03	96.49±0.08	97.98±0.83	97.78±1.07	97.20±0.87	97.29±0.27	98.11±0.80	**99.04±0.37**
AA	67.96±1.21	74.84±0.34	79.72±1.54	84.49±1.66	96.38±1.34	94.75±0.39	93.89±1.83	96.47±1.27	94.45±1.14	96.90±0.35	96.29±0.70	96.73±1.91	**98.32±0.19**
κ	74.19±0.81	82.34±0.26	85.68±0.66	89.26±0.52	91.85±0.77	93.98±0.76	95.76±0.40	96.91±0.25	96.03±0.35	96.44±0.40	96.84±0.34	97.31±0.76	**98.67±0.27**

TABLE 8. Classification results (%) for the BW test set using 15% of the available labeled data.

No.	SVM	RF	2-D CNN	3-D CNN	SSRN	SSAN	CAN	DBMA	DBDA	CSSEAN	RSSAN	REFIPN	SpaAG-RAN
1	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
2	90.12	81.32	69.14	97.53	98.77	96.30	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
3	96.50	97.78	**100.00**	**100.00**	**100.00**	96.00	**100.00**	98.00	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
4	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	97.09	**100.00**	**100.00**	**100.00**	**100.00**	95.35	**100.00**	**100.00**
5	77.31	86.01	87.04	90.28	94.91	87.50	93.98	98.61	97.69	99.07	97.69	96.76	**100.00**
6	59.53	88.43	99.07	75.81	80.00	87.91	97.21	98.14	98.60	93.02	**99.07**	96.74	95.52
7	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	99.52	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
8	85.80	**100.00**	**100.00**	**100.00**	**100.00**	99.38	96.30	**100.00**	98.15	**100.00**	**100.00**	**100.00**	**100.00**
9	92.75	94.66	82.07	100.00	**100.00**	94.42	**100.00**	94.02	**100.00**	90.84	98.80	**100.00**	**100.00**
10	97.99	99.55	99.50	95.48	91.96	**100.00**	**100.00**	**100.00**	96.98	**100.00**	**100.00**	97.99	**100.00**
11	97.95	98.91	97.13	**100.00**	**100.00**	97.95	**100.00**	99.18	91.39	97.95	97.95	**100.00**	**100.00**
12	98.62	**100.00**	**100.00**	95.17	**100.00**	**100.00**	**100.00**	99.31	**100.00**	**100.00**	99.31	96.55	**100.00**
13	**100.00**	90.50	90.70	**100.00**	**100.00**	95.35	89.30	**100.00**	98.14	**100.00**	99.53	**100.00**	**100.00**
14	97.37	94.12	92.11	89.47	97.37	88.16	90.79	96.05	96.05	**100.00**	93.42	96.05	**100.00**
OA	94.03±0.28	95.01±1.15	95.79±1.51	96.61±0.39	97.97±0.32	96.15±0.46	98.31±0.21	98.95±0.77	98.53±1.04	98.55±0.26	99.15±0.59	99.00±0.39	**99.74±0.23**
AA	92.66±1.55	95.12±1.85	94.05±0.36	95.98±1.71	97.36±1.64	95.68±0.59	97.68±1.50	98.81±1.00	98.36±0.42	98.63±0.22	98.65±0.29	98.86±0.70	**99.68±0.15**
κ	91.78±0.64	95.18±0.53	94.41±0.60	95.87±0.55	96.95±0.66	95.45±0.37	97.71±0.89	98.67±0.65	98.12±0.65	98.12±0.67	98.96±0.20	98.92±0.65	**99.60±0.36**

introduces the residual block, which improves the OAs of three data sets by 0.68%, 1.20%, and 1.36%, respectively. SSAN and CAN both employ the self-attention block to capture the spectral and spatial attention except that the latter emphasizes the center pixel during the attention acquisition. Therefore, CAN behaves better than SSAN on each data set. However, all of the OAs of SSAN on three data sets are lower than SSRN which does not utilizes the attention mechanism. In Table 8, the OA of SSAN is not even better than that of 3-D CNN. The most likely reason is that SSAN owns a big amount of parameters in feature extraction layers and attention modules, which brings challenge for the convergence of network under the condition of limited samples. The performances of DBMA receive a certain degree of increase comparing with SSAN and CAN. DBDA, which is stated as the improvement of DBMA, designs the spectral and spatial attention modules based on the self-attention mechanism. However, the classification performances of DBDA are better than DBMA on the IP data set only. CSSEAN deploys the lightweight spectral and spatial element attention modules to refine the dimension-reduced spectral and spatial features, which decreases a number of parameters. Nevertheless, this may cause the erratic classification accuracy when the limited samples are employed. For example, on the IP data set, the number of the samples of "Alfalfa", "Grass-pasture-mowed", and "Oats" (No.1, 7, and 9) are 7, 5, and 3, but the recalls of them are 97.30%, 13.04%, and 0.00% (marked with black rectangles in Table 6), respectively. The model of RSSAN, a combination of the 2-D residual block and attention module, simplifies the network and further improves the classification results as well. Among all compared methods, REFIPN reaches higher OAs on three data sets. This is attributed to the spatial attention modules, which refine the spectral-spatial features extracted by its pyramidal network. Last but not least, the proposed SpaAG-RAN model not only receives the best result of OA, AA, and κ but also obtains the overwhelming advantages on the higher recalls of categories comparing with the other methods. Different from the others, the SpaAM captures the salient areas based on the spectral similarity, which weakens the interfering pixels from the neighborhoods of the center pixel. The above superior classification performances demonstrate the effectiveness of the SpaAM as well as the excellent classification ability of the proposal.

2) QUALITATIVE COMPARISONS

The ground-truth (GT) images and visual classification maps of different methods on three data sets are shown in Fig. 16-18. Comparing with the other methods, RSSAN, REFIPN, and the proposal obtain the purer and smoother

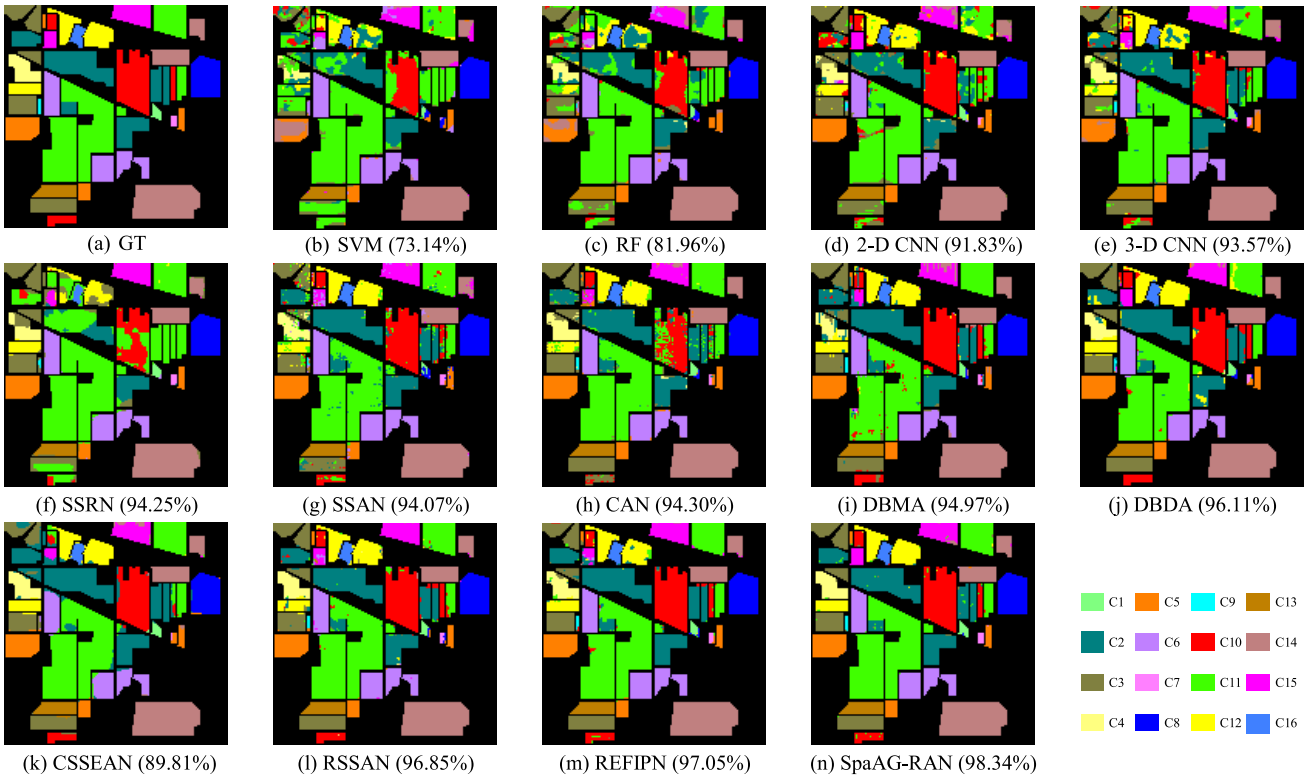

FIGURE 16. Classification maps of different methods on the IP data set (see Table 6). Where "Cn" represents the n-th category.

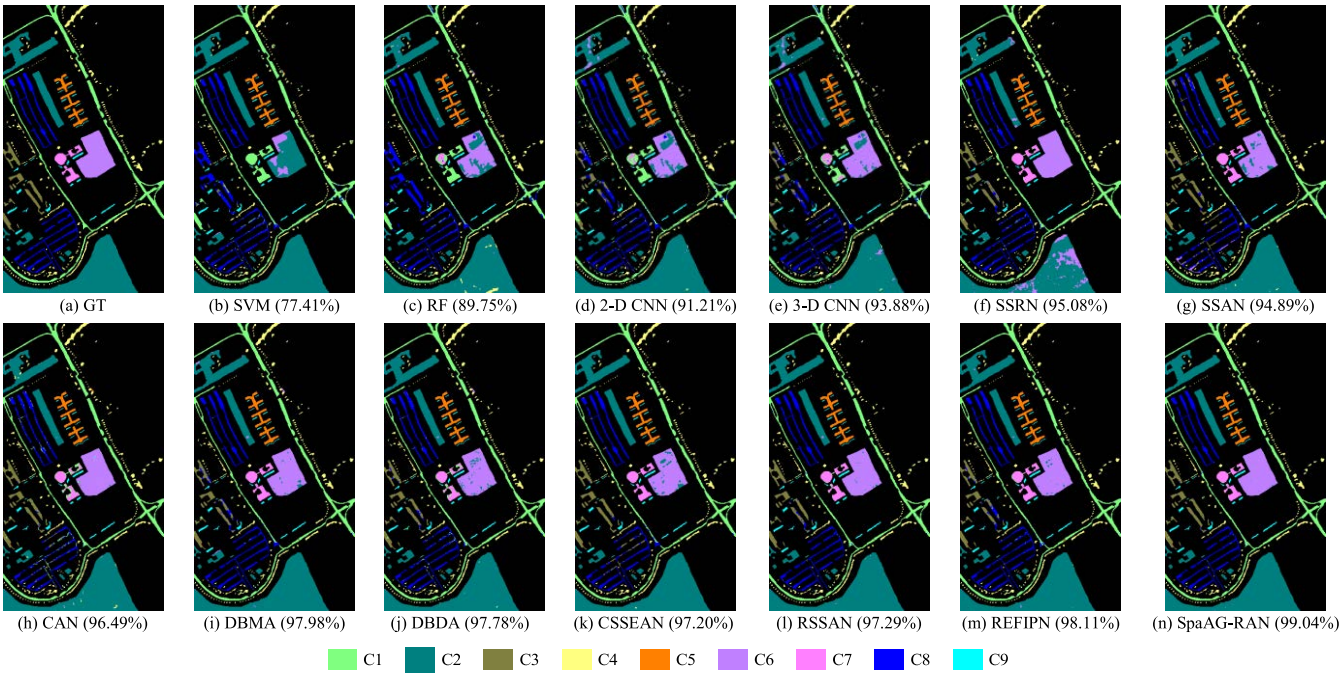

FIGURE 17. Classification maps of different methods on the UP data set (see Table 7). Where "Cn" represents the n-th category.

classification maps. Different from the UP and BW data sets, the distribution of the land-covers on the IP data set tends to be concentrated, which brings a certain of challenge to distinguish the useful pixels from the unwanted pixels. For example, the categories of "Corn" and "Grass-pasture-mowed" (C4 and C7) are always not

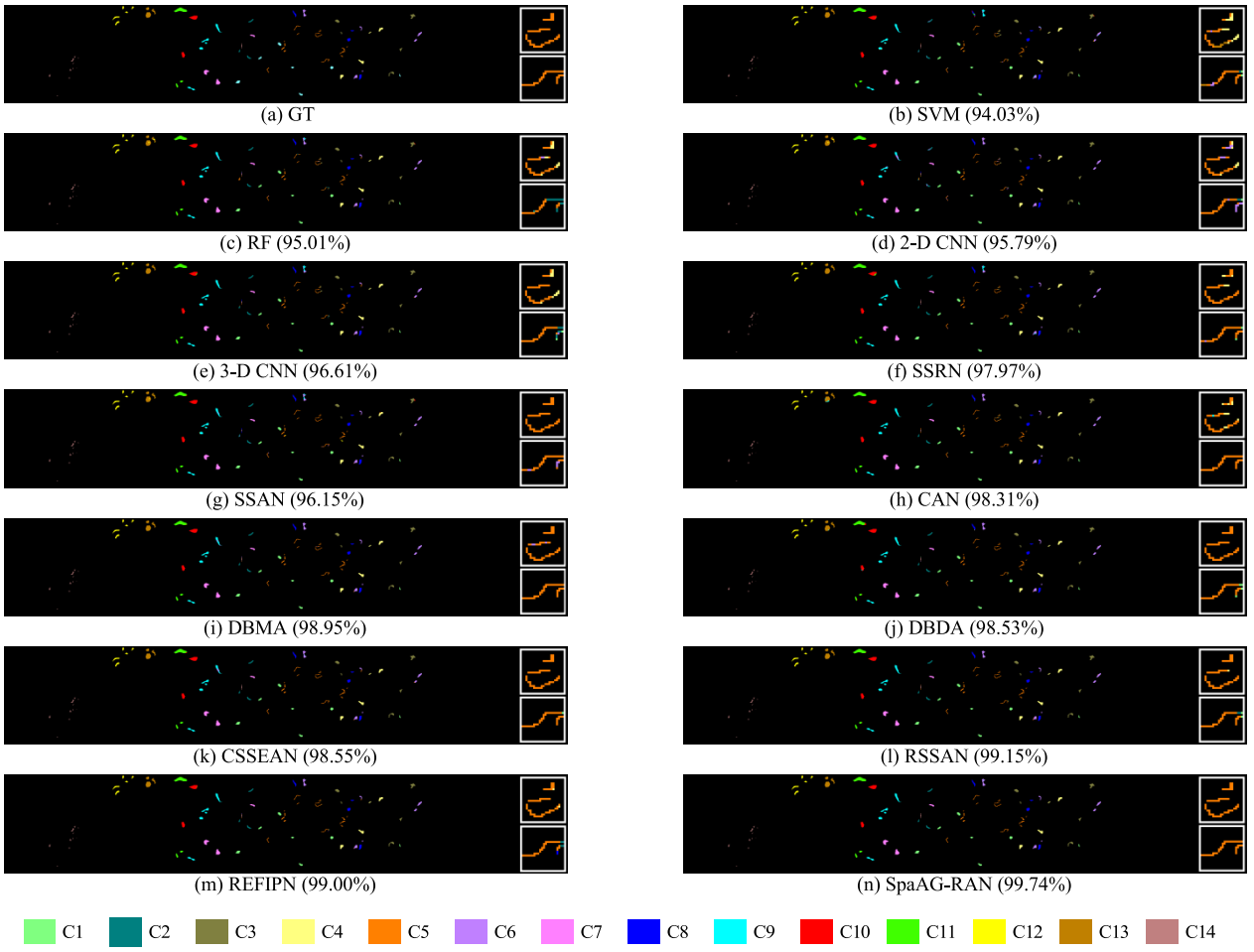

FIGURE 18. Classification maps of different methods on the BW data set (see Table 8). Where "Cn" represents the n-th category.

TABLE 9. Training time and testing time (Seconds) of different methods on each data set.

	Time (s)	SVM	RF	2-D CNN	3-D CNN	SSRN	SSAN	CAN	DBMA	DBDA	CSSEAN	RSSAN	REFIPN	SpaAG-RAN
IP	Training	4.24	5.56	19.49	295.57	1044.59	2753.21	502.12	513.22	654.64	285.66	104.51	613.30	304.51
	Testing	5.94	6.84	0.11	1.56	7.80	19.39	3.65	4.25	5.20	2.02	0.88	2.87	1.86
UP	Training	0.55	0.78	32.51	147.78	424.36	1903.68	165.58	210.98	261.52	108.09	142.78	143.65	162.66
	Testing	3.62	5.31	0.34	1.67	6.88	22.16	2.41	3.06	3.50	1.49	1.50	1.89	2.25
BW	Training	0.17	0.42	3.04	48.87	157.81	1278.24	60.58	71.89	88.44	36.82	30.51	52.86	42.81
	Testing	0.26	0.65	0.03	0.26	0.90	11.17	0.37	0.43	0.61	0.26	0.25	0.36	0.30

classified accurately by most compared methods. Even so, the proposed model still acquires the highest accuracy in these two categories. The similar result can be seen from the category of "Riparian" (C6) on the UP data set and the category of "Reeds1" (C5, the partial results are displayed in the white square boxes of each sub-figure of Fig. 18) on the BW data set. On the whole, in comparison with other methods, the proposed SpaAG- RAN model acquires the excellent classification maps which are almost as same as the corresponding ground-truth images of three data sets. This is because the SpaAG-RAN model is able to extract the discriminating spectral-spatial features from the relevant spatial areas.

3) TIME CONSUMPTION

The training and testing time of the proposed SpaAG-RAN model and the compared methods on three data sets are reported in Table 9. The training time has a closer link to the complexity of the network whereas the testing time intuitively reflects the efficiency of the algorithm in practical application. Among the thirteen methods, the two traditional methods (i.e. SVM and RF), cost less time obviously. In the

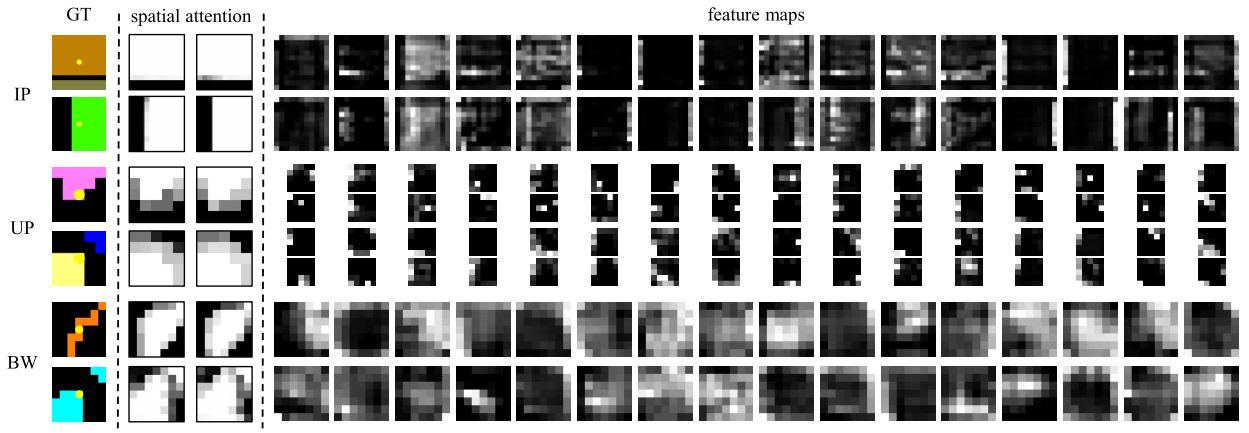

FIGURE 19. Visualization of the spatial attention masks and feature maps of the partial samples from the IP, UP, and BW data sets.

DL-based methods, 2-D CNN owns the fastest speeds as it extracts the spatial features from a single-channel image merely. The RSSAN, which adopts the 2-D CNN as the basic architecture, also reaches the second faster speed on three data sets. The remained methods all utilize the 3-D convolution for feature extraction, so the computation times of them are lengthy. However, the classification performances of them are jagged. It should be noted that the SSAN costs a considerable long time to finish the training procedures as the convolutional layers and the self-attention module both have a great deal of parameters. Among the methods based on 3-D CNN, the CSSEAN consumes the least time to finish training and testing, which has a close association with the pooling layers of it. Nevertheless, the proposed SpaAG-RAN model generates the spatial attention mask via an efficient subtract operation and introduces the few convolutional kernels with a smaller size to extract the spectral-spatial features, which results in a relative fast and efficient performance on three data sets.

E. VISUALIZATION OF THE SPATIAL ATTENTION AND FEATURES

In this part, some visualization studies are conducted to illustrate the power of the SpaAM intuitively to infer the relevant spatial areas. For each sample, its GT map, two spatial attention masks, i.e. $\mathbf{M}_{\mathbf{a}}^{\mathbf{i}}$ (left) and $\mathbf{M}_{\mathbf{a}}^{\mathbf{t}}$ (right), and the features $\mathbf{x}'$, which are acquired from the layer named AP (see Table 1) in the SSFEM, are visualized in Fig. 19. For the convenience of display, the spectral dimension of the features is compressed.

From this figure, the relevant spatial areas described in the two spatial attention masks have the similar spatial structures with the corresponding GT maps. More important, the center pixel of each sample is contained in the relevant spatial areas and assigned the highest spatial weight, which reveals that the production of the spatial attention takes the center pixel into fully account. With the restriction of the spatial consistency loss function, the two spatial attention masks also have the extremely similar distribution. Therefore, as shown in Fig. 19, most feature maps tend to focus on the relevant spatial areas. It is the two spatial attention masks that guide the SSFEM to extract the discriminating spectral-spatial features from the relevant spatial areas only.

IV. CONCLUSION

In this article, a novel SpaAG-RAN model is proposed for HSI classification, which contains a SpaAM, a SpeAM, and an SSFEM. The SpaAM aims to highlight the relevant spatial areas. The SpeAM aims to emphasize the spectral bands which are beneficial to the representation of characteristics. The SSFEM is designed to extract the spectral-spatial features. In the lightweight spectral-similarity-based SpaAM, a novel inverted-shifted-scaled sigmoid activation function is designed to convert each spectral similarity to the appropriate spatial weight. With the guidance of the SpaAM, the SpeAM and the SSFEM can work better. At the same time, to consolidate the capability of the SSFEM to discern the subtle differences between the homogeneous pixels and the interfering pixels, the spatial consistency loss function is exploited to preserve the stability of the spatial attention masks. The experimental results on three public data sets demonstrate the validity of the SpaAM and the outstanding classification performances of the proposal.

However, the scale and the threshold parameters of the inverted-shifted-scaled sigmoid activation function are set manually, which vary in different scenarios. One of the future directions of this work is to realize the adaptive selection of these two parameters. Moreover, a more effective spectral similarity measurement to acquire the more precise spatial attention mask is demanded as well.

REFERENCES

[1] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

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

[3] X. Yang and Y. Yu, "Estimating soil salinity under various moisture conditions: An experimental study," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2525–2533, May 2017.
[4] X. Zhang, Y. Sun, K. Shang, L. Zhang, and S. Wang, "Crop classification based on feature band set construction and object-oriented approach using hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4117–4128, Sep. 2016.
[5] G. Notesco, Y. Ogen, and E. Ben-Dor, "Mineral classification of Makhtesh Ramon in Israel using hyperspectral longwave infrared (LWIR) remote-sensing data," *Remote Sens.*, vol. 7, no. 9, pp. 12282–12296, Sep. 2015.
[6] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 101–117, Jun. 2019.
[7] 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.
[8] B. Waske, S. van der Linden, J. Benediktsson, A. Rabe, and P. Hostert, "Sensitivity of support vector machines to random feature selection in classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2880–2889, Jul. 2010.
[9] L. Ma, M. M. Crawford, and J. Tian, "Local manifold learning-based k-nearest-neighbor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4099–4109, Nov. 2010.
[10] Y. Zhang, G. Cao, X. Li, and B. Wang, "Cascaded random forest for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp. 1082–1094, Apr. 2018.
[11] M. Khodadadzadeh, J. Li, A. Plaza, and J. M. Bioucas-Dias, "A subspace-based multinomial logistic regression for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2105–2109, Dec. 2014.
[12] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.
[13] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, "SVM- and MRF-based method for accurate classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, Oct. 2010.
[14] X. Kang, S. Li, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification with edge-preserving filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2666–2677, May 2014.
[15] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.
[16] P. Quesada-Barriuso, F. Arguello, and D. B. Heras, "Spectral–spatial classification of hyperspectral images using wavelets and extended morphological profiles," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1177–1185, Apr. 2014.
[17] B. Hou, T. Huang, and L. Jiao, "Spectral-spatial classification of hyperspectral data using 3-Dmorphological profile," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2364–2368, Dec. 2015.
[18] L. Fang, S. Li, X. Kang, and J. A. Benediktsson, "Spectral-spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 12, pp. 7738–7749, Dec. 2014.
[19] J. Liu, Z. Wu, Z. Wei, L. Xiao, and L. Sun, "Spatial-spectral kernel sparse representation for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 6, pp. 2462–2471, Dec. 2013.
[20] L. He, J. Li, A. Plaza, and Y. Li, "Discriminative low-rank Gabor filtering for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1381–1395, Mar. 2017.
[21] C. J. D. Porta, A. A. Bekit, B. H. Lampe, and C.-I. Chang, "Hyperspectral image classification via compressive sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8290–8303, Oct. 2019.
[22] Q. Wang, Z. Meng, and X. Li, "Locality adaptive discriminant analysis for spectral–spatial classification of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2077–2081, Nov. 2017.
[23] H. Yuan, Y. Y. Tang, Y. Lu, L. Yang, and H. Luo, "Spectral-spatial classification of hyperspectral image based on discriminant analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2035–2043, Jun. 2014.
[24] D. L. Donoho, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lect.*, vol. 1, p. 32, Aug. 2000.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2015, pp. 1–9.
[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[28] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[29] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, Mar. 2012, pp. 127–135.
[30] E. Merényi, W. H. Farrand, J. V. Taranik, and T. B. Minor, "Classification of hyperspectral imagery with neural networks: Comparison to conventional tools," *EURASIP J. Appl. Signal Process.*, vol. 2014, no. 1, pp. 1–19, Dec. 2014.
[31] 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, Apr. 2017.
[32] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, Aug. 2019.
[33] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, Oct. 2018.
[34] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Hyperspectral image classification using spectral-spatial LSTMs," in *Proc. CCF Chin. Conf. Comput. Vis. (CCCV)*, Nov. 2017, pp. 577–588.
[35] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.
[36] X. Ma, H. Wang, and J. Geng, "Spectral-spatial classification of hyperspectral image based on deep auto-encoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4073–4085, Sep. 2016.
[37] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, Jun. 2017.
[38] Y. Chen, X. Zhao, and X. Jia, "Spectral-spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.
[39] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Learning and transferring deep joint spectral-spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.
[40] N. Li and Z. Wang, "Hyperspectral image ship detection based upon two-channel convolutional neural network and transfer learning," in *Proc. IEEE 5th Int. Conf. Signal Image Process. (ICSIP)*, Oct. 2020, pp. 88–92.
[41] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
[42] C. Wang, N. Ma, Y. Ming, Q. Wang, and J. Xia, "Classification of hyperspectral imagery with a 3D convolutional neural network and J-M distance," *Adv. Space Res.*, vol. 64, no. 4, pp. 886–899, Aug. 2019.
[43] Z. Ge, G. Cao, X. Li, and P. Fu, "Hyperspectral image classification method based on 2D–3D CNN and multibranch feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5776–5788, Sep. 2020.
[44] J. Zheng, Y. Feng, C. Bai, and J. Zhang, "Hyperspectral image classification using mixed convolutions and covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 522–534, Jan. 2021.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[46] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
[47] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral-spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[48] X. Zhang, Y. Wang, N. Zhang, D. Xu, H. Luo, B. Chen, and G. Ben, "Spectral–spatial fractal residual convolutional neural network with data balance augmentation for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10473–10487, Dec. 2021, doi: 10.1109/TGRS.2020.3046840.
[49] Y. Xu, Z. Li, W. Li, Q. Du, C. Liu, Z. Fang, and L. Zhai, "Dual-channel residual network for hyperspectral image classification with noisy labels," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, doi: 10.1109/TGRS.2021.3057689.
[50] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, "Visual attention-driven hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8065–8080, Oct. 2019.
[51] J. Su, J. Zeng, D. Xiong, Y. Liu, M. Wang, and J. Xie, "A hierarchy-to-sequence attentional neural machine translation model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 623–632, Mar. 2018.
[52] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.
[53] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Sep. 2018, p. 17.
[54] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, Sep. 2020.
[55] L. Zhao, J. Yi, X. Li, W. Hu, J. Wu, and G. Zhang, "Compact band weighting module based on attention-driven for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9540–9552, Nov. 2021, doi: 10.1109/TGRS.2021.3053397.
[56] P. Shamsolmoali, M. Zareapoor, J. Chanussot, H. Zhou, and J. Yang, "Rotation equivariant feature image pyramid network for object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2021.3112481.
[57] 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.
[58] Z. Lu, B. Xu, L. Sun, T. Zhan, and S. Tang, "3-D channel and spatial attention based multiscale spatial–spectral residual network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4311–4324, Jul. 2020.
[59] W. Ma, Q. Yang, Y. Wu, W. Zhao, and X. Zhang, "Double-branch multi-attention mechanism network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 11, p. 1307, 2019.
[60] L. Li, J. Yin, X. Jia, S. Li, and B. Han, "Joint spatial–spectral attention network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 10, pp. 1816–1820, Oct. 2021, doi: 10.1109/LGRS.2020.3007811.
[61] N. Li and Z. Wang, "Spectral-spatial fused attention network for hyperspectral image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 3832–3836.
[62] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral-spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.
[63] Z. Zhao, D. Hu, H. Wang, and X. Yu, "Center attention network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3415–3425, 2021.
[64] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, early access, Oct. 8, 2021, doi: 10.1109/TGRS.2021.3115699.
[65] 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.
[66] X. He and Y. Chen, "Optimized input for CNN-based hyperspectral image classification using spatial transformer network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1884–1888, Dec. 2019.
[67] H. Yan, J. Wang, L. Tang, E. Zhang, K. Yan, K. Yu, and J. Peng, "A 3D cascaded spectral–spatial element attention network for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 13, p. 2451, Jun. 2021.
[68] R. Li, S. Zheng, C. Duan, Y. Yang, and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network," *Remote Sens.*, vol. 12, no. 3, p. 582, Feb. 2020.
[69] C. Zhao, M. Tian, and J. Li, "Research progress on spectral similarity metrics," *J. Harbin Eng. Univ.*, vol. 38, no. 8, pp. 1179–1189, Aug. 2017.
[70] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[71] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectified neural networks," in *Proc. AISTATS*, vol. 15, Jan. 2011, pp. 315–323.
[72] *Hyperspectral Remote Sensing Scenes Grupo de Inteligencia Computacional (GIC)*. Accessed: Apr. 24, 2021. [Online]. Available: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes
[73] T. Tieleman and G. Hinton, "Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude," *COURSERA, Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.
[74] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural network," in *Proc. AISTSTS*, 2010, pp. 249–256.

NINGYANG LI received the B.S. degree in remote sensing science and technology from Henan Polytechnic University, Jiaozuo, China, in 2019. He is currently pursuing the M.S. degree in software engineering with the School of Computer Science and Technology, Hainan University, Haikou, China. His research interests include hyperspectral image processing and analysis and deep learning.

ZHAOHUI WANG received the M.S. degree in image processing from the University of Derby, U.K., in 2004, and the Ph.D. degree from the University of Leeds, U.K., in 2008. He then joined the Norwegian Colour and Visual Computing Laboratory, Gjøvik, Norway, to work on visual computing and multispectral color imaging research projects. He joined Hainan University, China, in 2013, where he is currently a Professor of computer science at the Faculty of Computer Science and Technology. His current research interests include hyperspectral image processing and analysis, remote sensing image processing and its applications, computer vision, and deep learning. His professional memberships include IS&T, SPIE, IEEE, and CCF.