S^3AM: A Spectral-Similarity-Based Spatial Attention Module for Hyperspectral Image Classification

Ningyang Li, Zhaohui Wang, Faouzi Alaya Cheikh, *Senior Member, IEEE*, and Mohib Ullah, *Member, IEEE*

***Abstract*— Recently, hyperspectral image (HSI) classification based on deep learning methods has attracted growing attention and made great progress. Convolutional neural networks based models, especially the residual networks (ResNets), have become the architectures of choice for extracting the deep spectral-spatial features. However, there are generally some interfering pixels in the neighborhoods of the center pixel, which are unfavorable for the spectral-spatial feature extraction and will lead to a restraint classification performance. More important, the existing attention modules are weak in highlighting the effect of the center pixel for the spatial attention. To solve this issue, this article proposes a novel spectral-similarity-based spatial attention module (S^3AM) to emphasize the relevant spatial areas in HSI. The S^3AM adopts the weighted Euclidean and cosine distances to measure the spectral similarities between the center pixel and its neighborhoods. To alleviate the negative influence of the spectral variability, the full-band convolutional layers are deployed to reweight the bands for the robust spectral similarities. Both kinds of weighted spectral similarities are then fused adaptively to take their relative importance into full account. Finally, a scalable Gaussian activation function, which can suppress the interfering pixels dynamically, is installed to transform the spectral similarities into the appropriate spatial weights. The S^3AM is integrated with the ResNet to build the S^3AM-Net model, which is able to extract the discriminating spectral-spatial features. Experimental results on four public HSI datasets demonstrate the effectiveness of the proposed attention module and the outstanding classification performance of the S^3AM-Net model.**

***Index Terms*—Center pixel, hyperspectral image classification, residual network, spatial attention, spectral similarity.**

I. Introduction

HYPERSPECTRAL image (HSI) is provided with ample spectral and spatial information and such an incomparable advantage makes the precise recognition of land-covers possible [1]. Therefore, HSI classification, which aims at distributing an appropriate label for each pixel, has drawn a great many researchers from various fields, such as agriculture [2], mineral prospecting [3], and urban planning [4]. However, exploiting the spectral and spatial information of HSI effectively for excellent classification accuracy has always been a challenging problem as the characteristics of high-dimensional and redundancy of HSI.

In the past decade, an enormous amount of methods have been proposed to cope with HSI classification, including traditional methods and deep learning (DL) methods. Traditional models, such as *k*-nearest neighbor [5], support vector machine [6], random forest [7], and so on, are deemed to be the models that can extract the shallow spectral and spatial features from HSI merely, which results in the inferior robustness when handling the complicated scenarios.

As the speedy advances in computing power, DL-based methods have achieved remarkable progress and have been widely used in computer vision (CV) tasks [8], [9] and natural language processing [10]. In the field of HSI classification, DL-based algorithms have seized the dominant status by means of the power of extracting the deep spectral and spatial features [11]. In the early phase, most models finished classification with spectral features only. For instance, stack autoencoder [12] and deep belief network [13] are used to extract the compressed and invariant spectral features from raw spectra, respectively. To take the spectral dependency which is the inherent property between spectral bands into account, recurrent neural networks [14], and long short-term memory networks [15] were introduced to model the correlations hidden in the adjacent bands. In addition, the graph neural networks [16] were coalesced with the multistructure unified embedding [17] and the sparse manifold correlation [18] to integrate diverse characteristics and receive quality features. However, due to the well-known issue of spectral variability, the classification performances of these spectral feature-based models are still unsatisfactory. Meanwhile, supplementing the spatial features for better classification results has been an urgent motivation.

Recently, with the inimitable advantages of local capture and parameters sharing, convolutional neural networks (CNNs) [19], possess the power of feature extraction and have made tremendous progress in the field of HSI classification. On the strength of the variability of convolutional kernels, CNN is capable of extracting the features in different gradations [20]. The early architectures integrated one-dimensional (1-D) CNN and 2-D CNN in parallel to extract the spectral features and spatial features, respectively [21]. But these models fuse the

Manuscript received 5 April 2022; revised 26 May 2022 and 1 July 2022; accepted 9 July 2022. Date of publication 18 July 2022; date of current version 3 August 2022. This work was supported in part by the Framework of the Norwegian Research Council INTPART Project under Grant 309857, in part by the International Network for Image-Based Diagnosis, and in part by the Hainan Key Research and Development Plan for Scientific and Technological Collaboration Projects under Grant GHYF2022015 - Research on Medical Imaging Aided Diagnosis of Infant Brain Development Diseases. *(Corresponding author: Zhaohui Wang.)*

Ningyang Li and Zhaohui Wang are with the Faculty of Computer Science and Technology, Hainan University, Haikou 570228, China (e-mail: fes_map@qq.com; william_hig@163.com).

Faouzi Alaya Cheikh and Mohib Ullah are with the Department of Computer Science, Faculty of Information Technology and Electrical, Norwegian University of Science and Technology, 2815 Gjøvik, Norway (e-mail: faouzi.cheikh@ntnu.no; mohib.ullah@ntnu.no).

Digital Object Identifier 10.1109/JSTARS.2022.3191396

two kinds of features at the stage of decision merely, which fail to take the feature associations into full account. Instead of dealing with both kinds of features separately, 3-D CNN adopts the cubic convolutional kernels to acquire the spectral-spatial features effectively from the HSI cube, which is composed of the center pixel and its neighborhoods [22], [23]. Such architecture can fuse the spectral features and the spatial features at the stage of feature extraction, which is favorable for enhancing the correlations between the spectral and spatial features. Moreover, these benefits will be enlarged as the depth of the network increases [24]. However, the deeper the network is, the more possible the vanishing gradient may happen. To resolve this issue effectively, a novel network architecture, which is named residual network (ResNet) [25], introduced the residual blocks containing the skip connections to deliver the gradient between the layers directly. Such skill of the residual connection is active in various types of architectures for HSI classification, such as convolutional long short-term networks [26], capsule networks [27], graph neural networks [28], and generative adversarial networks [29]. Therefore, a few of related research [30]–[32] reveal that ResNet has become the popular architecture to express the deeper spectral-spatial features.

Though the ResNets have reached acceptable performances, there is an undeniable fact that not all of the pixels and bands in the HSI cube contribute to the feature extraction and classification. The pixels owning the same category as the center pixel play a key role in enhancing the correlations of features. Thus, these pixels together with the center pixel form the relevant spatial area worth focusing well. The same criterion applies to the spectral dimension. Among the hundreds of bands, only partial salient bands are beneficial to distinguish the subtle spectral differences that exist in all types of land covers precisely. To accomplish the purposes, the attention mechanism, an active technique in the fields of neural machine translation [33] and CV [34], was applied to upgrade the capability of a network to describe the discriminating spectral-spatial features for HSI classification. The popular structures of the attention modules in recent literature can be divided into three types mainly as follows.

1) Self-attention (SA) module [35]. It aims at exploring the relationship among different elements (e.g., words of a sentence, pixels of an image) by computing the dot-product similarity to realize the feature recalibration [36]. In HSI classification, the SA modules were applied to take account of the spectral interactions and the spatial correlations to refine the features [37], [38]. Moreover, the integration of several SA modules forms the transformer [39], which is able to model the spectral and spatial relationships in different representation subspaces [40], [41]. However, due to the abundant spectral bands in HSI and the massive matrix operations of SA modules, the transformer often spends lengthy time and excessive memory.
2) Squeeze and excitation (SE) module [42]. On the basis of a global average pooling layer and a symmetrical multilayer perceptron, the SE module can collect the global expressive information and map them to the channel attention. The SE module has made appreciable progress on reweighting the spectral bands adaptively [43], [44]. Besides, it was extended to make full use of spatial contextual information [45], [46].
3) Convolutional block attention module (CBAM) [47]. The CBAM is the comprehensive attention module. The channel attention, which is the extension of the SE module, defines "what" is meaningful. The spatial attention uses the large-scale convolution and the global pooling layers to locate the meaningful content. With the guidance of the CBAM, the distinctive bands as well as the useful pixels in HSI cube were enhanced properly [48]–[50].

It is indisputable that the three kinds of attention modules have gotten the spectral feature extraction better. However, there is a common shortage that they are weak in emphasizing the benefit of the center pixel to the deduction of spatial attention. Specifically, the dot-product matrix of the SA module, the spatial squeezing and excitation operations of the SE module, and the large-scale convolution of the CBAM lack the special treatment for the center pixel. This may cause spatial attention to deviate from the relevant spatial areas, which will impair the extraction of the discriminating features greatly and may result in wrong predictions. Considering the different spectral characteristics between the relevant pixels and the interfering pixels, a spectral-similarity-based spatial attention module (S^{3}AM) is proposed in this article. Due to the fact that the spectral similarity depended on a single metric tends to possess weak representation [44], the S^{3}AM adopts the Euclidean and cosine distances to obtain the spectral similarity. Nevertheless, the immediate similarities of the two original measures are often inexact due to the notorious spectral variability. To solve this deficiency, the S^{3}AM introduces the full-band convolutional (FBC) layer in the calculation of the original similarities to redistribute a group of proper weights for all bands. Hence, the weighted Euclidean and cosine distances (WED and WCD) to obtain robust spectral similarities are constructed. Both kinds of weighted spectral similarities are then fused adaptively to take their relative importance into full account. Finally, a scalable Gaussian (SG) activation function, which weakens the interfering pixels in different scenarios with a flexible threshold, is designed to transform the weighted spectral similarity to the optimal spatial attention mask. Experimental results on four publicly available HSI datasets, including the Indian Pines, Pavia University, Loukia, and XiongAn, show that the S^{3}AM can bring the advantage of the center pixel fully to concentrate on the relevant spatial areas effectively to further improve the classification performance. The main contributions of this article are as follows.

1) A novel S^{3}AM is proposed to capture the relevant spatial areas effectively in the HSI cube. Specifically, the WED and WCD submodules, which adopt the FBC layers to relieve the adverse influence of the spectral variability, are first applied to improve the robustness of the spectral similarities. Both weighted spectral similarities are then integrated adaptively to gain the representative composite spectral similarity. Finally, an SG activation function is designed to convert the spectral similarities to the appropriate spatial weights flexibly in diverse scenes. The S^{3}AM excels at emphasizing the spatial areas relevant intensively

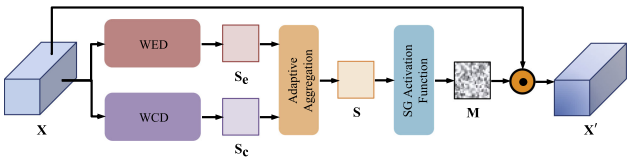

Fig. 1. Architecture of the S^{3}AM. It contains the WED and WCD submodules, the adaptive aggregation, and the SG activation function.

to the center pixel and preserving these crucial areas even in a wider HSI cube.

2) An end-to-end S^{3}AM-Net model, which contains the S^{3}AM and ResNet, is designed to obtain the discriminating features for HSI classification. With the support of the functional S^{3}AM, this model is capable of handling the spatial features as well as the spectral-spatial features efficiently.

The rest of this article is organized as follows. Section II illustrates the proposed S^{3}AM in detail. Section III presents the experiments and analyses on four HSI datasets. Finally, Section IV concludes this article.

II. Spectral-Similarity-Based Spatial Attention Module

A. Overview of the S^{3}AM

In an HSI cube $\mathbf{X} \in \mathbb{R}^{c\times\omega\times\omega\times b}$, the relevant spatial areas are composed of the neighborhoods that belong to the same label as the center pixel, where c, ω, and b denote the number of channels (i.e., filters), width, and the number of bands, separately. The relevant areas deserve emphasis as the features extracted from these areas are beneficial to improve the classification performance. Since there are intense associations between the relevant pixels and the center pixel in terms of spectral similarity, the S^{3}AM is therefore proposed to promote the center pixel to contribute more to the capture of the relevant areas. In particular, to resolve the inferior representational ability of a single similarity led by the abundant spectral bands, the S^{3}AM utilizes the WED and WCD to measure the spectral relevance between the center pixel and its neighborhoods.

The architecture of the S^{3}AM is shown in Fig. 1. It consists of the WED sub-module, the WCD submodule, the adaptive aggregation, and the SG activation function. For an HSI cube $\mathbf{X}$, the WED and WCD submodules are in charge of computing the weighted Euclidean and cosine spectral similarities, i.e., $\mathbf{S_e} \in \mathbb{R}^{1\times\omega\times\omega\times1}$ and $\mathbf{S_c} \in \mathbb{R}^{1\times\omega\times\omega\times1}$, between the center pixel and its neighborhoods at first. Both kinds of similarities behave stronger robustness against the spectral variability due to the operation of band reweighting. Next, to regulate the magnitudes of the two types of spectral similarities for spatial attention rationally, they are merged into the composite spectral similarity $\mathbf{S} \in \mathbb{R}^{1\times\omega\times\omega\times1}$ by adaptive aggregation. Finally, the SG activation function weakens the interfering pixels by considering the specifications of different scenarios and transforms the spectral similarity $\mathbf{S}$ into the spatial attention mask $\mathbf{M} \in \mathbb{R}^{1\times\omega\times\omega\times1}$. With the assistance of $\mathbf{M}$, the HSI cube $\mathbf{X}$ can be refined by

$$\mathbf{X}' = \mathbf{M} \odot \mathbf{X} \tag{1}$$

where $\mathbf{X}' \in \mathbb{R}^{c\times\omega\times\omega\times b}$ is the refined output and "$\odot$" is the element-wise multiplication. During the multiplication, the elements of $\mathbf{X}$ are copied along the axes of channel and band.

B. WED and WCD

1) Weighted Euclidean and Distance: The WED submodule aims to obtain the robust weighted spectral similarity based on the Euclidean distance. The Euclidean distance is the standard metric to measure the difference of amplitude between two spectra in Euclidean space. The smaller the distance is, the higher the similarity is. The original formula of the Euclidean distance S'_{e} is defined as

$$\mathrm{S}'_{\mathrm{e}} = \sqrt{\sum\nolimits_i^N \left(\mathbf{p}_\mathbf{c}^i - \mathbf{X}_\mathbf{1}^i\right)^2} \tag{2}$$

where $\mathbf{p_c} \in \mathbb{R}^{1\times1\times1\times b}$and $\mathbf{X_1} \in \mathbb{R}^{1\times1\times1\times b}$ represent the center pixel and its first neighborhood, respectively. $\mathbf{p}_\mathbf{c}^i$ is the ith band of the center pixel.

However, due to the notorious spectral variability, the partial bands of spectra often fluctuate in the wide range. Thus, the spectral similarity computed by the original formula will represent some deviations, which may amplify the similarity of intraclass and narrow the similarity of interclass.

To resolve this problem, the WED submodule introduces the FBC layer, which contains a convolutional kernel $\mathbf{w} \in \mathbb{R}^{1\times1\times1\times b}$ with the same size as the bands of the HSI cube, to distribute a group of apposite weights for all bands. Specifically, for the relevant pixels, the FBC layer assigns smaller weights for the bands behaving unstable spectral energy to promote the cohesion of the spectral similarity. On the contrary, larger weights are allocated to those bands for the interfering pixels. Consequently, the WED sub-module can obtain the spectral similarity that is stable and reliable in the scene with intense spectral variation.

The weighted Euclidean spectral similarity $\mathrm{S}_{\mathrm{e}}^1$ between the center pixel $\mathbf{p_c}$ and its first neighborhood $\mathbf{X_1}$ can be defined as

$$\mathrm{S}_{\mathrm{e}}^1 = \sqrt{\sum\nolimits_i^b \mathbf{w}^i \cdot \left(\mathbf{p}_\mathbf{c}^i - \mathbf{X}_\mathbf{1}^i\right)^2} \tag{3}$$

where $\mathbf{w}^i$ is the ith weight of the convolutional kernel $\mathbf{w}$ of the FBC layer, $\mathbf{p}_\mathbf{c}^i$ denotes the ith band of the center pixel, and $\mathbf{X}_\mathbf{1}^i$ denotes the ith band of the first neighborhood of the center pixel.

The architecture of the WED submodule is shown in Fig. 2. First, the center pixel extracted from the HSI cube $\mathbf{X}$ is copied ω^2 times to form the center cube $\mathbf{X_c} \in \mathbb{R}^{1\times\omega\times\omega\times b}$. Next, the differences, $\mathbf{D} \in \mathbb{R}^{1\times\omega\times\omega\times b}$, of all bands between $\mathbf{X_c}$ and $\mathbf{X}$ are computed by the element-wise subtraction and multiplication

$$\mathbf{D} = (\mathbf{X_c} \ominus \mathbf{X}) \odot (\mathbf{X_c} \ominus \mathbf{X}). \tag{4}$$

To accomplish the redistribution of the importance of all bands, the FBC layer should be placed after the multiplication. If the FBC layer is installed before the multiplication, the weights

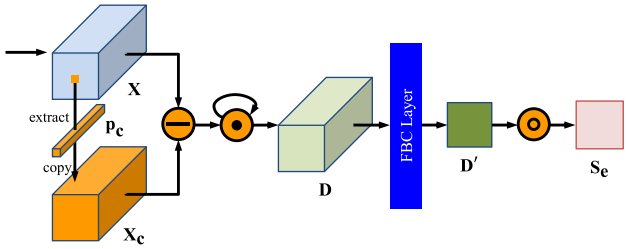

Fig. 2. Architecture of the WED submodule. Where "⊖" and "⊛" denote the elementwise subtraction and the elementwise square root, respectively.

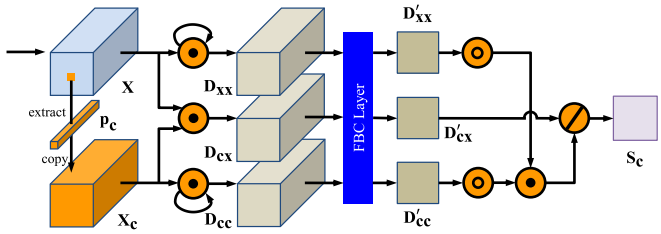

Fig. 3. Architecture of the WCD submodule. Where "⊘" denotes the elementwise division.

of the layer will be limited in the non-negative range due to the subsequent square operation. This may inactivate some weights and lower the efficiency of optimization. More important, another benefit of installing the FBC layer after the multiplication is that it can sum all bands automatically after the weighting operation. This process can be described as follow:

$$\mathbf{D}' = \mathbf{w} * \mathbf{D} = \sum_i^b \mathbf{w}^i \cdot ((\mathbf{X_c}^i \ominus \mathbf{X}^i) \odot (\mathbf{X_c}^i \ominus \mathbf{X}^i)) \tag{5}$$

where $\mathbf{D}' \in \mathbb{R}^{1\times\omega\times\omega\times 1}$ is the summation of the weighted difference along the axis of band and "$*$" denotes the operation of convolution.

Finally, by solving the square root of this summation, the spectral similarity $\mathbf{S_e}$ based on the WED is gained

$$\mathbf{S_e} = \sqrt{\mathbf{D}'} = \sqrt{\sum\nolimits_i^b \mathbf{w}^i \cdot ((\mathbf{X_c}^i \ominus \mathbf{X}^i) \odot (\mathbf{X_c}^i \ominus \mathbf{X}^i))}\,. \tag{6}$$

2) Weighted Cosine Distance: The WCD submodule aims to acquire the robust weighted spectral similarity based on the cosine distance. The cosine distance, as the core theory of spectral angle mapping [51], is also a popular measurement to calculate the directional similarity between two spectra. The closer to one the distance is, the higher the similarity is. The original formula of the cosine distance S'_c between the center pixel $\mathbf{p_c}$ and its first neighborhood $\mathbf{X_1}$ is defined as

$$\mathrm{S}'_\mathrm{c} = 1 - \frac{\sum_i^N \mathbf{p_c}^i \cdot \mathbf{X_1}^i}{\sqrt{\sum_i^N {\mathbf{p_c}^i}^2} \cdot \sqrt{\sum_i^N {\mathbf{X_1}^i}^2}}. \tag{7}$$

Similar to the Euclidean distance, the spectral variability also exerts the negative effects on the cosine distance. Therefore, the FBC layer is also applied in the WCD submodule to obtain the weighted cosine spectral similarity S^1_c

$$\mathrm{S}^1_\mathrm{c} = \frac{\sum_i^b \mathbf{w}^i \cdot (\mathbf{p_c}^i \cdot \mathbf{X_1}^i)}{\sqrt{\sum_i^b \mathbf{w}^i \cdot {\mathbf{p_c}^i}^2} \cdot \sqrt{\sum_i^b \mathbf{w}^i \cdot {\mathbf{X_1}^i}^2}}\,. \tag{8}$$

The architecture of the WCD submodule is shown in Fig. 3. First, the center cube $\mathbf{X_c}$ is constructed with the actions of extracting and copying. Unlike the Euclidean distance, the cosine distance owns three basic summation units, as shown in (7). Thus, the differences are computed in three routes

$$\mathbf{D_{xx}} = \mathbf{X} \odot \mathbf{X} \tag{9a}$$

$$\mathbf{D_{cx}} = \mathbf{X_c} \odot \mathbf{X} \tag{9b}$$

$$\mathbf{D_{cc}} = \mathbf{X_c} \odot \mathbf{X_c} \tag{9c}$$

where the shapes of the prototypes of the three basic summation units $\mathbf{D_{xx}}$, $\mathbf{D_{cx}}$, and $\mathbf{D_{cc}}$ are the same as that of the input $\mathbf{X}$.

Following the analysis in the previous section, the FBC layer is installed after the three elementwise multiplications. In particular, $\mathbf{D_{xx}}$, $\mathbf{D_{cx}}$, and $\mathbf{D_{cc}}$ share the identical FBC layer, which not only remains the consistency of the weight of each band but also decreases the number of parameters. The three weighting processes can be described as

$$\mathbf{D'_{xx}} = \mathbf{w} * \mathbf{D_{xx}} = \sum_i^b \mathbf{w}^i \cdot (\mathbf{X}^i \odot \mathbf{X}^i) \tag{10a}$$

$$\mathbf{D'_{cx}} = \mathbf{w} * \mathbf{D_{cx}} = \sum_i^b \mathbf{w}^i \cdot (\mathbf{X_c}^i \odot \mathbf{X}^i) \tag{10b}$$

$$\mathbf{D'_{cc}} = \mathbf{w} * \mathbf{D_{cc}} = \sum_i^b \mathbf{w}^i \cdot (\mathbf{X_c}^i \odot \mathbf{X_c}^i) \tag{10c}$$

where $\mathbf{D'_{xx}} \in \mathbb{R}^{1\times\omega\times\omega\times 1}$, $\mathbf{D'_{cx}} \in \mathbb{R}^{1\times\omega\times\omega\times 1}$, and $\mathbf{D'_{cc}} \in \mathbb{R}^{1\times\omega\times\omega\times 1}$ are the three basic summation units.

Finally, to get the spectral similarity $\mathbf{S_c}$ based on the WCD, $\mathbf{D'_{cx}}$ is divided by the product of the square roots of $\mathbf{D'_{xx}}$ and $\mathbf{D'_{cc}}$

$$\begin{aligned}\mathbf{S_c} &= \frac{\mathbf{D'_{cx}}}{\sqrt{\mathbf{D'_{cc}}} \odot \sqrt{\mathbf{D'_{xx}}}} \\ &= \frac{\sum_i^b \mathbf{w}^i \cdot (\mathbf{X_c}^i \odot \mathbf{X}^i)}{\sqrt{\sum_i^b \mathbf{w}^i \cdot (\mathbf{X_c}^i \odot \mathbf{X_c}^i)} \odot \sqrt{\sum_i^b \mathbf{w}^i \cdot (\mathbf{X}^i \odot \mathbf{X}^i)}}\,.\end{aligned} \tag{11}$$

C. Adaptive Aggregation

The WED tends to reflect the differences in terms of the amplitudes of both spectra, whereas the WCD mainly reveals the discrepancy of the angles between two spectra. The two kinds of spectral similarities generally make unequal contributions for judging whether two spectra belonged to the same label indeed [52]. To take fully advantage of the two kinds of spectral similarities, the adaptive aggregation, which equips a learnable contribution coefficient α, is built to integrate $\mathbf{S_e}$ and $\mathbf{S_c}$ to the composite spectral similarity $\mathbf{S}$

$$\mathbf{S} = \alpha \cdot \mathbf{S_e} \oplus (1-\alpha) \cdot (1 - \mathbf{S_c}) \tag{12}$$

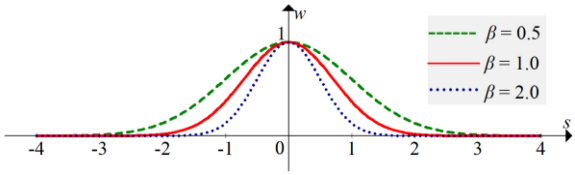

Fig. 4. SG activation function. Where s and w indicate the spectral similarity and spatial weight, respectively.

where "⊕" denotes the elementwise addition. The values of $\mathbf{S_c}$ are subtracted by one to ensure that the monotonicity of similarity with respect to distance between $\mathbf{S_c}$ and $\mathbf{S_e}$ is equal.

During the training procedure, the value of α gets gradually close to the optimal value. The contributions of the spectral similarities based on the WED and the WCD are represented as α and $(1-\alpha)$, separately. With this configuration, the relative importance of $\mathbf{S_e}$ and $\mathbf{S_c}$ is adjusted rationally and the representative composite spectral similarity $\mathbf{S}$ is attained.

D. SG Activation Function

In the composite spectral similarity $\mathbf{S}$, the similarities of the pixels in the relevant areas are little. In particular, the similarity of the center pixel is equal to zero. On the contrary, the similarities of the interfering pixels are great. In addition, the threshold used to distinguish the interfering pixels is usually not fixed in different scenes [52]. Therefore, to transform the spectral similarity into the proper spatial attention mask, an activation function with the following properties is demanded.

1) The smaller the spectral similarity is, the bigger the spatial weight is, and vice versa.
2) The pixels with the zero spectral similarity, e.g., the center pixel, should be assigned the max spatial weight, i.e., one.
3) The capability of this activation function to suppress the interfering pixels should be stable even in various scenes.

To satisfy the three criteria, the SG activation function is designed. Mathematically, it is written as

$$\mathbf{M} = e^{-\beta \cdot \mathbf{S}^2} \tag{13}$$

where β is the learnable scaling parameter and $\mathbf{M}$ is the spatial attention mask. The range of the values in $\mathbf{M}$ is [0, 1].

As shown in Fig. 4, the SG activation function remains the monotonicity and the range of the original Gaussian function. With the help of this specific, the relevant pixels, which own the little similarities are assigned the greater spatial weights whereas the interfering pixels are allocated the smaller spatial weights. The spatial weight of the center pixel is equal to one, which ensures the vital status of it in the relevant spatial areas. More important, the response of the SG activation function is mutable under the influence of the learnable scaling parameter β. The bigger the value of β is, the steeper the curves, the more interfering pixels are weakened. This means that the SG activation function can suppress the interfering pixels in $\mathbf{M}$ dynamically even in different scenes. Therefore, the relevant spatial areas in HSI cube $\mathbf{X}$ can be emphasized effectively.

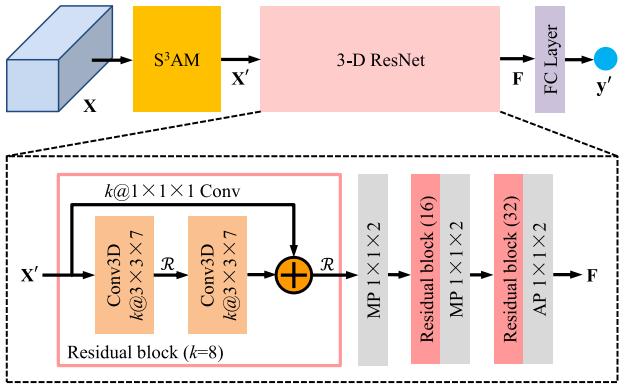

Fig. 5. S^{3}AM-Net model. The S^{3}AM is embedded between the HSI cube and the 3-D ResNet. Where "⊕" denotes the elementwise addition.

E. Instantiation

In this article, 3-D ResNet, which introduce the residual blocks to relieve the vanishing gradients, is adopted as the baseline. By integrating it with the proposed S^{3}AM, the S^{3}AM-Net model is able to enhance the relevant areas to improve the distinction of the spectral-spatial features for classification. The overview of the S^{3}AM-Net model is shown in the upper part of Fig. 5. Considering the characteristic of the S^{3}AM, it is embedded between the input $\mathbf{X}$ and the 3-D ResNet. This ensures the spectral similarities computed from the original HSI cube have higher credibility. Effects of different embedding strategies will be presented in Section III-C-(5).

As shown in the lower part of Fig. 5, the input and output are the refined HSI cube $\mathbf{X}'$ and the discriminating spectral-spatial features $\mathbf{F} \in \mathbb{R}^{32\times\omega'\times\omega'\times b/8}$, respectively. Where ω' denotes the width of the output of the last residual block. The 3-D ResNet contains three residual blocks. In each residual block, two 3-D convolutional layers, which are equipped with the kernels of 3×3×7 and the rectified linear unit [53] activation function, aim to extract the nonlinear spectral-spatial features. Before the output of each residual block, a residual connection is applied to realize the immediate mapping of gradients via a convolutional layer of 1×1×1. The numbers of the kernels k in the three residual blocks are set to {8, 16, 32}, respectively. Following the former two residual blocks, the max-pooling layers are utilized to compress the dominant features, while the average pooling (AP) layer is used after the last residual block to reserve more semantic features [24]. The pooling sizes and strides of all pooling layers are set to 1×1×2. Finally, a fully connected layer predicts the most possible label $\mathbf{y}' \in \mathbb{R}^{1\times C}$, where C is the number of categories.

III. Experiments and Analyses

In this section, four public HSI datasets as well as the implementation details are first described. The ablation studies, classification experiments, parameter analysis, and attention visualization are then presented to demonstrate the structural rationality and effectiveness of the S^{3}AM-Net model.

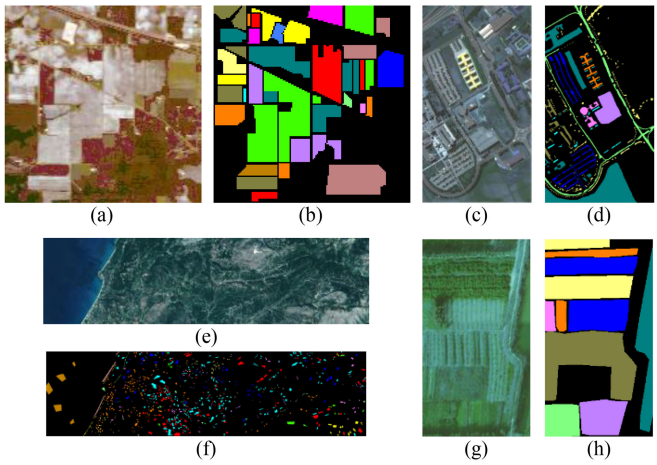

Fig. 6. FC images and GT maps. (a) and (b) Indian Pines. (c) and (d) Pavia University. (e) and (g) Loukia. (f) and (h) XiongAn.

A. Datasets and Evaluation Metrics

Indian Pines: This dataset was collected by the airborne visible/infrared imaging spectrometer sensor over Indian Pines test site in Northwestern Indiana in 1992 [54]. It contains 145×145 pixels and 200 valid spectral bands.

Pavia University: This dataset was collected by the reflective optics spectrometer imaging system sensor over during a flight campaign over Pavia, North Italy in 2002 [54]. It contains 610×340 pixels and 103 valid spectral bands.

Loukia: This dataset was gathered by the Hyperion sensor mounted on the Earth Observing-1 satellite. It is one of the HyRANK benchmark datasets, which have been developed in the framework of the ISPRS Scientific Initiatives in 2018 [55]. It contains 945×249 pixels and 176 valid spectral bands.

XiongAn: This dataset was acquired by the visible and near-infrared imaging spectrometer designed by Shanghai Institute of Technical Physics, Chinese Academy of Sciences in Xiongan New Area in 2017 [56]. It contains 400×220 pixels and 250 valid spectral bands.

The false-color (FC) images and ground-truth (GT) maps of the four datasets are shown in Fig. 6. As shown in Tables I–IV, {5%, 5%, 90%}, {2%, 5%, 93%}, {5%, 5%, 90%}, and {1%, 5%, 94%} of the samples from each category of the four datasets are selected randomly as the training, validation, and test sets, respectively. The training set is used for training the proposed model and the comparison methods. The validation set is used for evaluating the fitting states of models during training procedures merely. The test set is used for verifying the classification performance of the models.

In this article, the overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) are employed to quantify the classification performances of the S^{3}AM-Net model on the three datasets. The higher the scores of the three metrics, the better performance the model receives. To alleviate the influence of the random initialized parameters, all experiments are performed ten times and the average results are reported.

TABLE I
SAMPLES OF THE INDIAN PINES DATASET

No.	Color	Land-cover type	Training	Validation	Test
1		Alfalfa	2	2	42
2		Corn-notill	71	71	1286
3		Corn-mintill	41	42	747
4		Corn	12	12	213
5		Grass-pasture	24	24	435
6		Grass-trees	37	37	656
7		Grass-pasture-mowed	1	1	26
8		Hay-windrowed	24	24	430
9		Oats	1	1	18
10		Soybean-notill	49	49	874
11		Soybean-mintill	123	123	2209
12		Soybean-clean	30	30	533
13		Wheat	10	10	185
14		Woods	63	63	1139
15		Building-Grass-Trees-Drives	19	19	348
16		Stone-Steel-Towers	5	5	83
		Total	512	513	9224

TABLE II
SAMPLES OF THE PAVIA UNIVERSITY DATASET

No.	Color	Land-cover type	Training	Validation	Test
1		Asphalt	132	332	6167
2		Meadows	373	933	17343
3		Gravel	42	105	1952
4		Trees	61	153	2850
5		Painted metal sheets	27	67	1251
6		Bare Soil	100	251	4678
7		Bitumen	27	67	1236
8		Self-Blocking Bricks	74	184	3424
9		Shadows	19	47	881
		Total	855	2139	39782

TABLE III
SAMPLES OF THE LOUKIA DATASET

No.	Color	Land-cover type	Training	Validation	Test
1		Dense urban fabric	14	15	259
2		Mineral extraction sites	3	3	61
3		Non irrigated arable land	27	27	488
4		Fruit trees	4	4	71
5		Olive groves	70	70	1261
6		Broad-leaved forest	11	11	201
7		Coniferous forest	25	25	450
8		Mixed forest	54	54	964
9		Dense sclerophyllous vegetation	190	190	3413
10		Sparse sclerophyllous vegetation	140	140	2523
11		Sparcely vegetated areas	20	20	364
12		Rocks and sand	24	24	439
13		Water	70	70	1253
14		Coastal water	23	23	405
		Total	675	676	12152

TABLE IV
SAMPLES OF THE XIONGAN DATASET

No.	Color	Land-cover type	Training	Validation	Test
1		Chinese scholartree	46	229	1132
2		Water	72	360	3551
3		Corn	163	815	4298
4		Pear	104	523	5681
5		Poplar	38	189	6773
6		Vegetable field	60	302	9823
7		Grassland	12	60	12757
8		Peach	136	679	15320
		Total	631	3157	59335

B. Implementation Details

In the preprocessing and initialization phases, the values of each dataset are normalized into the range from 0 to 1. The normal distribution with unit mean and 0.1 standard deviation are utilized to initialize the FBC layers in the WED and WCD submodules. The contribution coefficient α of the adaptive aggregation is initialized to 0.5 and the scaling parameter β of the SG activation function is set to 1.0 at first. During the back propagation, the ranges of the values of α and β are [0, 1]. Besides, the Xavier normal distribution [57] is used to initialize all layers of baseline. The batch size bs is set to 32 and the total number of training iteration is 200.

During the training procedure, the error $\mathcal{L}$ between the true labels $\mathbf{Y} \in \mathbb{R}^{bs \times C}$ and the predicted labels $\mathbf{Y}' \in \mathbb{R}^{bs \times C}$ in a mini batch are assessed by the famous cross-entropy loss function

$$\mathcal{L} = -\frac{1}{\mathrm{bs}} \sum_{\mathrm{i}=1}^{\mathrm{bs}} \sum_{\mathrm{j}=1}^{\mathrm{C}} \mathbf{Y}_{\mathrm{i,\,j}} \log(\mathbf{Y}'_{\mathrm{i,\,j}}). \tag{14}$$

While training the model, the RMSprop [58] optimizer is adopted to optimize the whole network, where the values of (*learning rate*, *beta1*, *beta2*) are set to (0.001, 0.9, 0.999).

All experiments are performed on a computer with an AMD Ryzen 3600 at 4.07 GHz × 6 with 32-GB RAM and an NVIDIA GeForce GTX 1080Ti graphical processing unit with 12-GB RAM. The operating system is the Ubuntu 16.01. The DL framework is the Keras 2.3.1.

C. Ablation Studies

1) Different Measures of the Spectral Similarities: In this section, to demonstrate the advantage of the combination of the WCD and WED, different measures of the spectral similarities are installed into the S^{3}AM. Besides the adopted WED and WCD, other three measures including the weighted Manhattan distance (WMD), the weighted Chebyshev distance (WCheD), and the weighted Tonimoto coefficient distance (WTD). The spectral similarities, $\mathrm{S}^1_{\mathrm{m}}$, $\mathrm{S}^1_{\mathrm{che}}$, and $\mathrm{S}^1_{\mathrm{t}}$, between the center pixel and its first neighborhood using the three distances are as follow:

$$\mathrm{S}^1_{\mathrm{m}} = \sum_{i}^{b} \mathbf{w}^i \cdot \left| \mathbf{p}^i_{\mathbf{c}} - \mathbf{X}^i_{\mathbf{1}} \right| \tag{15}$$

$$\mathrm{S}^1_{\mathrm{che}} = \max(\mathbf{w}^i \cdot \left| \mathbf{p}^i_{\mathbf{c}} - \mathbf{X}^i_{\mathbf{1}} \right|) \tag{16}$$

$$\mathrm{S}^1_{\mathrm{t}} = \frac{\sum_i^b \mathbf{w}^i \cdot (\mathbf{p}^i_{\mathbf{c}} \cdot \mathbf{X}^i_{\mathbf{1}})}{\sqrt{\sum_i^b \mathbf{w}^i \cdot {\mathbf{p}^i_{\mathbf{c}}}^2} + \sqrt{\sum_i^b \mathbf{w}^i \cdot {\mathbf{X}^i_{\mathbf{1}}}^2} - \sum_i^b \mathbf{w}^i \cdot (\mathbf{p}^i_{\mathbf{c}} \cdot \mathbf{X}^i_{\mathbf{1}})}. \tag{17}$$

The OAs of the proposed model using the five measures on four datasets are reported in Table V. The numbers presented in bold-type denote the best results. From the upper half of the table, the WED and WCD win the top two OAs on four datasets. The WCheD receives the worst performances, which are lower than those of the baseline. It is because the WCheD chooses the max discrepancy of certain band immediately as the spectral similarity, which is likely to lose the important information hid in other bands. The WMD takes the summation of the absolute differences in all bands. The stability of it to handle the bands express large disparity is weaker than WED, so the OAs of it are lower. The WTD, which is the extension of the WCD, obtains the spectral similarity via the intersection and distribution of the spectra. But there are still visible gaps between the performances of the WTD and WCD. It may be the redundant bands that the computation of spectral similarity has a little deviation.

TABLE V
OAs (%) OF THE S^{3}AM-NET MODEL USING DIFFERENT MEASURES (COMBINATIONS) OF THE SPECTRAL SIMILARITY

	Indian Pines	Pavia University	Loukia	XiongAn
baseline	86.54±0.24	95.54±0.26	80.85±0.41	83.99±1.63
WED	91.73±0.77	**97.56±0.11**	83.86±0.53	87.98±0.73
WCD	**92.05±0.26**	97.10±0.06	**84.02±0.39**	**88.42±0.52**
WMD	91.27±0.63	96.85±0.72	83.21±0.64	87.24±0.95
WCheD	89.08±1.78	92.42±0.98	79.26±0.77	83.27±1.28
WTD	91.12±0.99	95.48±0.61	82.74±0.57	87.31±0.69
WED + WCD	**93.31±0.54**	**97.99±0.28**	**84.25±0.30**	**88.95±0.87**
WED + WMD	91.79±0.55	97.63±0.73	83.89±0.23	88.04±0.61
WED + WTD	92.11±1.02	97.59±0.56	83.96±0.38	88.20±0.57
WCD + WMD	92.65±0.88	97.46±0.70	84.13±0.49	88.56±0.77
WCD + WTD	92.43±0.93	97.22±0.39	84.06±0.45	88.48±0.35

TABLE VI
OAs (%) OF THE S^{3}AM-NET MODELS WITH OR WITHOUT (W/O) THE FBC LAYERS ON FOUR DATASETS

	Indian Pines	Pavia University	Loukia	XiongAn
with FBC	**93.31±0.54**	**97.99±0.28**	**84.25±0.30**	**88.95±0.87**
w/o FBC	91.67±0.28	96.92±0.99	83.07±0.24	87.42±0.86

Therefore, the WED and WCD, which achieve the top two ranks on four datasets, are integrated with other two measures, WMD and WTD, to explore the optimal combination of the spectral similarity, separately. As for the WCheD, it is not considered to conduct this study due to its lowest accuracy. From the lower half of Table V, there is an additional promotion for each combination comparing with the cases using single measure. More inspiring, the combination "WED + WCD" reaches the highest OAs on three datasets.

2) Role of the FBC Layer

As the essential parts of the WED and WCD submodules, the FBC layer aims to redistribute a special weight for each band. Thus, the weighted spectral similarities behave good robustness in the scenes with intense spectral variability. To prove this role, the FBC layers in the WED and WCD submodules are removed during the training procedure, which means the weights of all bands are equal.

As shown in Table VI, once the FBC layer is installed, the promotions of the OAs of the S^{3}AM-Net model on the four data sets are no less than 1.6%, 1.0%, 1.1%, and 1.5%, respectively. These discrepancies, which cannot be ignored, show that the band reweighting operation of the FBC layer can weaken the negative effects of the spectral variability to improve the classification performance.

3) Impact of the Adaptive Aggregation: The adaptive aggregation is designed to adjust the relative magnitude of the WED and WCD submodules via the contribution coefficient α.

TABLE VII
OAs (%) OF THE TWO SCHEMES ON FOUR DATA SETS AND THE OPTIMAL CONTRIBUTION COEFFICIENTS OF THE SCHEME "ADAPTIVE"

	Indian Pines	Pavia University	Loukia	XiongAn
equal	92.95±0.42	97.77±0.16	84.10±0.23	88.59±0.74
adaptive	**93.31±0.54**	**97.99±0.28**	**84.25±0.30**	**88.95±0.87**
optimal α	0.4538	0.6514	0.4739	0.4261

TABLE VIII
OAs (%)THE S^{3}AM-NET MODELS WITH THE GAUSSIAN OR SG ACTIVATION FUNCTIONS ON FOUR DATA SETS AND THE OPTIMAL SCALING PARAMETERS

	Indian Pines	Pavia University	Loukia	XiongAn
Gaussian	93.02±0.13	97.65±0.37	83.97±0.42	88.39±1.04
SG	**93.31±0.54**	**97.99±0.28**	**84.25±0.30**	**88.95±0.87**
optimal β	1.4415	1.1599	0.8651	1.3584

For the sake of interpreting the necessity, two kind of schemes using the equal and adaptive aggregations are carried out.

The comparison of the two schemes on four datasets is shown in Table VII. When the equal aggregation is utilized, the value of the contribution coefficient α is fixed as 0.5, which causes the equal contribution between the WED and WCD submodules. Consequently, the OAs of this scheme decline a little on three datasets comparing with the scheme "adaptive". The optimal contribution coefficients of the scheme "adaptive" are shown in the last column of Table VII. From these optimized values, the WCD submodule makes more contributions for the classification on the Indian Pines, Loukia, and XiongAn datasets while the WED submodule is valued on the Pavia University dataset. The similar conclusion can be discovered from comparison of the third and fourth rows in Table V. These results reveal that the adaptive aggregation can regulate the relative importance of the WED and WCD submodules for better performance in different scenarios.

4) Impact of the SG Activation Function: The SG activation function takes charge of the conversion from the composite spectral similarity $\mathbf{S}$ to the spatial attention mask $\mathbf{M}$. With the assistance of its learnable scaling parameter β, the interfering pixels can be well restrained even though in diverse scenes. To exploring the influence of this parameter on the classification performance, besides the SG activation function, the original Gaussian activation function is also introduced for comparison.

The OAs of the two schemes on four datasets are shown in the second and third rows of Table VIII. It can be observed that the OAs of the model using the SG activation function gains the slight increase comparing with the model with the Gaussian activation function. This is because the scaling parameter β enables the SG activation function represents the flexible attitudes to suppress the interfering pixels as precisely as possible in different scenarios. Moreover, from the last row of Table VIII, the optimal values of the scaling parameters of the Indian Pines, Pavia University and XiongAn datasets are larger than that of the Loukia data set, which shows the thresholds used to weaken the interfering pixels of the Indian Pines, Pavia University, and XiongAn datasets are higher. It is the SG activation function that notices these essential specifics for upgrading the classification accuracy.

TABLE IX
OAs (%) OF THE BASELINE AND FOUR NETWORKS WITH DIFFERENT INTEGRATION STRATEGIES ON FOUR DATASETS

	Indian Pines	Pavia University	Loukia	XiongAn
baseline	86.54±0.24	95.54±0.26	80.85±0.41	83.99±1.63
Net_1	92.26±0.40	97.67±0.20	84.03±0.24	88.27±0.47
Net_2	92.98±0.24	97.88±0.14	84.15±0.53	88.76±0.91
Net_2	91.35±0.27	96.42±0.30	82.19±0.78	86.87±0.52
Net_4	**93.31±0.54**	**97.99±0.28**	**84.25±0.30**	**88.95±0.87**

5) Integration Strategies of the S^{3}AM: The purpose of the S^{3}AM is to extract the relevant spatial areas in HSI cube. There are many ways to integrate the S^{3}AM with the baseline, i.e., 3-D ResNet. To seek the best solution, four networks built with various integration strategies are analyzed as follows.

1) Net_1: Embedding the S^{3}AM into the residual blocks of the baseline (Before each residual connection).
2) Net_2: Embedding the S^{3}AM between the residual blocks.
3) Net_3: Embedding the S^{3}AM after the baseline.
4) Net_4: the S^{3}AM-Net model.

As shown in Table IX, Net_4 outperforms other three networks along with the baseline on four datasets slightly. The OAs of the baseline, which does not introduce the S^{3}AM are tolerable. Net_1 aims to adjust the deep features extracted by the convolutional layers inside the residual blocks while Net_2 oversees refining the deep residual features extracted by the residual blocks. However, the OAs of Net_2 are higher than those of Net_1, which may because the residual features create a shortcut path between the HSI cube and the S^{3}AM to preserve the effect of spatial attention masks. Net_3 places the S^{3}AM after the baseline to refine the spectral-spatial features with the spatial attention mask. But OAs of it on four datasets are unsatisfactory comparing with the other networks apart from the baseline. This gap indicates that executing the refinement in the final stage of feature extraction may not be the optimal option. By embedding the S^{3}AM between the input and the baseline, Net_4 can generate the more expressive spatial attention mask from HSI cube directly. Therefore, the network extracts the discriminating spectral-spatial features from the relevant areas for the superior classification accuracy.

D. Comparison With the State-of-the-Arts

To verify the effectiveness of the proposed S^{3}AM-Net model, the 2- and 3-D ResNets, seven attention-based methods, including spectral spatial self-attention network (SSSAN) [38], spectral former (patchwise) (SFP) [40], spectral spatial feature tokenization transformer (SSFTT) [41], compact band weighting module (CBW) [43], densely connected multiscale attention network (DMSAN) [46], double-branch multiattention network (DBMA) [48], and residual spectral spatial attention network (RSSAN) [49], as well as the 2-D version of the proposed model, are reimplemented for comparison. For each method, the architecture from the original article is adopted. During the training, validation, and testing procedures, all methods share the same training, validation, and testing sets listed in Tables I–IV.

TABLE X
CLASSIFICATION RESULTS (%) FOR THE INDIAN PINES TESTING SET USING 5% OF THE AVAILABLE LABELED DATA

No.	ResNet (2-D | 3-D)		SSSAN	SFP	SSFTT	CBW	DMSAN	DBMA	RSSAN	S^3AM-Net (2-D | 3-D)	
1	56.41±0.00	57.95±3.08	**91.11±5.67**	45.41±13.61	73.51±10.73	84.44±10.30	68.57±18.16	90.83±6.82	90.42±20.37	60.51±3.08	72.97±5.13
2	71.63±4.81	75.54±2.44	74.20±5.23	84.78±2.05	88.71±5.57	87.84±2.41	77.26±10.04	83.05±7.24	81.79±6.46	79.47±2.41	**90.64±3.92**
3	75.01±2.08	83.77±1.24	76.81±10.70	86.45±1.52	87.23±4.04	89.49±1.82	87.65±16.81	89.66±4.32	94.64±3.49	87.18±2.55	**95.60±1.00**
4	50.25±0.83	58.31±1.27	90.69±3.20	81.69±3.37	72.28±5.77	87.23±7.16	78.29±23.02	81.66±17.40	92.94±5.88	57.91±2.70	75.66±2.57
5	83.41±2.64	88.42±1.48	**97.03±1.31**	92.66±1.00	94.88±3.07	70.28±27.51	91.22±15.46	95.30±2.77	90.80±5.57	89.49±0.39	92.56±0.38
6	98.03±0.31	97.77±0.06	98.17±1.11	98.12±1.72	99.86±0.13	98.03±1.39	97.53±3.16	97.15±2.77	98.34±1.72	97.55±0.26	**99.88±0.23**
7	32.50±4.08	41.67±1.10	88.89±4.54	61.74±10.43	89.57±8.95	22.22±24.00	73.02±10.04	75.00±17.79	53.47±42.03	50.83±1.67	**91.03±6.96**
8	99.06±0.10	99.26±0.22	98.71±1.82	99.16±1.01	99.63±0.27	99.25±0.84	99.31±1.01	98.87±1.23	**100.00±0.00**	99.85±0.12	99.58±0.21
9	12.94±2.35	25.88±4.71	46.15±27.38	78.75±12.87	68.75±3.95	2.56±3.63	46.15±29.65	65.38±15.86	41.35±22.73	56.47±2.88	**98.75±2.50**
10	75.11±2.75	80.53±1.62	83.86±9.22	**92.25±1.88**	86.67±8.24	85.39±8.02	66.43±18.19	77.81±19.43	77.00±14.46	84.38±1.28	90.06±1.51
11	90.98±1.63	92.42±1.65	95.05±2.24	96.88±0.98	87.33±6.45	89.87±1.44	87.10±13.96	87.03±8.65	91.24±5.74	93.22±2.25	**97.90±2.91**
12	53.29±3.40	63.13±2.55	84.16±14.02	83.50±4.84	82.36±7.90	**89.52±3.00**	72.32±22.85	76.62±13.33	81.75±9.69	65.75±2.19	74.51±4.67
13	99.57±0.28	99.43±0.36	99.25±0.61	99.76±0.30	97.44±0.98	99.50±0.35	98.17±2.16	**99.81±0.33**	95.21±1.72	99.31±0.23	99.63±0.49
14	99.26±0.20	99.29±0.17	98.58±0.50	95.18±0.83	**99.70±0.37**	94.90±5.14	96.13±5.65	99.24±1.04	96.60±1.99	99.01±0.09	99.55±0.37
15	89.15±0.88	89.70±0.83	88.84±7.27	87.44±4.91	88.28±2.73	87.12±7.32	83.89±6.86	84.86±10.03	**98.11±2.26**	92.13±0.71	92.17±3.55
16	76.46±2.48	71.39±1.01	79.44±12.27	**95.95±3.31**	85.95±9.42	83.89±3.42	91.43±10.52	74.58±20.46	90.00±9.90	74.94±1.48	87.03±4.06
OA	83.42±0.83	86.54±0.24	89.38±1.34	91.90±0.41	90.30±1.72	89.42±0.85	85.31±3.81	88.26±2.20	90.02±1.79	88.31±0.33	**93.31±0.54**
AA	72.69±0.70	76.53±0.47	86.93±1.55	86.23±2.29	87.64±2.45	79.47±0.99	82.15±5.80	86.05±2.79	85.85±2.81	80.50±0.58	**90.79±0.62**
κ	80.94±1.01	84.58±0.28	87.82±1.53	90.73±0.47	88.95±1.94	87.92±0.97	83.13±4.50	86.57±2.50	88.61±2.02	86.63±0.36	**92.35±0.61**

TABLE XI
CLASSIFICATION RESULTS (%) FOR THE PAVIA UNIVERSITY TESTING SET USING 2% OF THE AVAILABLE LABELED DATA

No.	ResNet (2-D | 3-D)		SSSAN	SFP	SSFTT	CBW	DMSAN	DBMA	RSSAN	S^3AM-Net (2-D | 3-D)	
1	93.35±2.66	93.93±1.24	97.36±0.52	97.12±0.46	96.29±0.63	95.43±3.09	98.15±0.06	**99.06±0.52**	93.71±2.57	94.47±1.01	98.18±0.71
2	98.82±0.51	98.91±0.79	99.68±0.11	99.72±0.16	99.57±0.30	99.06±0.45	99.57±0.01	99.25±0.33	96.42±2.96	99.79±0.05	**99.83±0.22**
3	77.34±7.84	83.34±4.01	81.23±3.66	85.39±5.38	84.95±7.19	73.01±9.58	**87.89±0.21**	82.66±3.58	80.39±7.84	82.35±5.14	83.79±4.51
4	88.34±1.55	94.23±1.00	95.37±1.09	93.04±0.80	95.94±0.57	95.24±1.86	95.02±0.16	**97.98±0.79**	93.23±1.71	94.02±0.96	97.33±1.24
5	99.67±0.04	99.72±0.22	99.85±0.03	**100.00±0.00**	99.94±0.03	95.63±3.80	99.49±0.04	99.77±0.30	99.94±0.12	99.75±0.04	**100.00±0.00**
6	84.63±2.61	89.75±2.13	93.57±1.30	94.06±2.08	92.74±2.22	96.36±1.71	86.77±0.38	96.34±3.86	93.24±4.26	88.60±2.55	**97.76±2.25**
7	69.21±14.73	92.51±2.19	87.96±6.29	89.26±3.42	92.98±3.20	82.15±12.16	89.08±0.58	85.58±15.57	83.28±6.85	89.72±11.62	**95.40±1.76**
8	93.32±4.23	94.28±2.69	97.30±0.70	94.60±4.04	91.69±2.25	84.72±5.79	95.38±0.29	95.60±0.89	89.90±5.83	**97.74±1.65**	97.42±2.26
9	98.41±0.18	98.98±0.33	98.59±0.36	98.41±0.51	99.05±0.21	97.81±0.82	**100.00±0.00**	99.83±0.13	99.11±0.64	98.48±0.34	99.97±0.05
OA	93.40±0.24	95.54±0.26	96.80±0.25	96.68±0.23	96.40±0.32	94.73±0.37	96.26±0.06	97.26±0.62	93.81±1.30	95.61±0.34	**97.99±0.28**
AA	89.23±1.49	93.96±0.54	94.54±0.60	94.62±0.27	94.79±0.97	91.04±1.41	94.59±0.08	95.12±1.83	92.14±0.85	93.88±1.19	**96.62±0.35**
κ	91.16±0.34	94.08±0.34	95.74±0.33	95.58±0.31	95.21±0.44	93.00±0.49	95.01±0.08	96.37±0.83	91.82±1.66	94.14±0.46	**97.33±0.37**

1) Quantitative Comparisons: The quantitative evaluations, including the recall of each category, OA, AA, and κ, of each method on the Indian Pines, Pavia University, Loukia, and XiongAn test sets are reported in Tables X–XIII. From these results, the following conclusions can be drawn. First, the 2- and 3-D ResNets, as the methods without adopting any attention module, gain the tolerable classification results. After the S^3AM are integrated with them, the classification performances of the 2- and 3-D S^3AM-Net models all gain appreciable promotion. Second, among the attention-based methods, DBMA achieves the highest classification performances on the Pavia University dataset while SFP is the most excellent model for another three datasets. SFP uses the transformers to yield the groupwise spectral embedding for extracting the local spectral sequence features. DBMA combines the CBAM with the densely connected network. Both are helpful to enhance the useful spectral-spatial features. SSSAN employs the spectral and spatial SA modules to refine the bands and pixels, which causes the slight lower performances than DBMA. This indicates the CBAM possesses stronger ability to reweight features than the SA module to some extent. SSFTT utilizes the transformers to model the high-level semantic features and receives the commendable classification performances closed to those of SFP. CBW applies the SE module to generate the spectral attention mask for band selection merely, which causes its OAs lower than those of 3-D ResNet on the last three datasets (marked with rectangles in Tables XI–XIII). Similarly, DMSAN, which extends the SE module to spectral and spatial dimensions, acquires the unsatisfied classification performance on the Indian Pines and Loukia datasets (marked with rectangles in Tables X and XII). Both classification results of CBW and DMSAN reveal that the SE module may behaves inferior capability for capturing the relevant areas comparing with the SA module and CBAM. RSSAN, which integrates the CBAM and 2-D residual blocks, gains the mediocre accuracy. In particular, the OA of RSSAN on the Pavia University dataset is less than that of ResNet (marked with rectangles in Table XI). The most likely reason is that the 2-D residual blocks of RSSAN cause the loss of the spectral information, which is important for HSI classification. Last but not least, though the proposed S^3AM-Net model introduces the spatial attention merely, the number of the highest recall and the OA, AA, and κ of it on four datasets still reaches the best standard. Different from other types of attention modules, the special generation solution of the S^3AM takes the center pixel into fully account, which enables it to weaken the interfering pixels and infer the relevant spatial areas effectively. Therefore,

TABLE XII
CLASSIFICATION RESULTS (%) FOR THE LOUKIA TESTING SET USING 5% OF THE AVAILABLE LABELED DATA

No.	ResNet (2-D | 3-D)		SSSAN	SFP	SSFTT	CBW	DMSAN	DBMA	RSSAN	S^{3}AM-Net (2-D | 3-D)	
1	29.95±12.03	46.74±10.92	48.24±16.59	60.21±12.30	58.93±3.44	42.67±7.12	42.46±8.37	56.04±12.19	44.17±6.29	51.76±13.24	**61.82±9.86**
2	92.73±5.45	92.73±7.24	97.27±2.65	**98.64±1.11**	94.55±3.40	93.18±4.98	93.18±4.07	97.73±3.52	95.91±3.64	94.09±6.36	98.18±2.23
3	79.77±4.76	82.27±4.82	84.60±3.12	85.85±6.07	86.25±5.17	81.19±4.89	83.12±4.25	84.94±4.41	81.88±3.56	81.42±4.98	**88.92±4.13**
4	0.78±1.57	2.35±0.78	3.14±2.66	5.88±3.51	4.31±2.88	1.96±1.24	1.57±0.78	6.27±2.88	1.18±1.57	2.75±2.00	**7.45±4.54**
5	85.77±5.44	88.98±4.19	91.17±2.06	91.33±1.89	91.02±1.70	82.46±6.45	90.43±5.73	92.12±1.57	89.64±1.44	89.73±1.66	**92.14±2.42**
6	8.55±3.34	34.48±9.06	41.24±9.80	39.17±6.46	37.10±4.69	28.28±5.65	25.38±6.31	39.45±9.89	33.93±5.45	32.28±9.34	**45.52±8.33**
7	27.75±8.46	46.46±4.80	49.72±6.10	**59.94±9.30**	47.82±4.65	40.49±5.02	43.02±6.52	57.17±8.83	50.95±7.91	52.18±9.33	59.75±4.37
8	65.71±8.93	63.04±10.27	64.91±9.30	63.30±15.00	64.79±10.99	62.47±10.53	64.28±7.10	68.38±4.63	61.89±6.87	67.06±8.36	**68.44±5.00**
9	79.07±3.94	85.65±3.93	**86.95±1.86**	85.15±4.53	86.36±4.50	84.19±5.31	82.82±5.42	85.40±2.97	83.68±2.45	84.19±3.68	86.14±1.89
10	79.60±2.55	83.21±1.62	83.04±2.09	86.38±2.55	85.06±2.00	83.34±2.25	82.59±3.51	84.41±2.60	**86.42±2.28**	84.16±3.31	85.51±1.16
11	38.33±7.60	39.39±12.76	59.01±4.27	54.07±8.51	50.19±14.95	33.54±15.27	46.92±7.85	54.45±5.37	51.56±8.14	50.57±4.77	**62.36±3.39**
12	92.51±0.95	93.69±1.07	94.11±0.91	**94.89±0.54**	93.88±0.71	93.94±0.81	93.63±0.97	93.11±0.69	93.44±0.46	94.02±0.53	94.26±0.86
13	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**
14	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**	**100.00±0.00**
OA	76.54±0.82	80.85±0.41	82.54±0.43	83.17±0.53	82.65±0.40	79.12±1.14	80.06±0.57	83.07±0.23	81.39±0.62	81.63±0.19	**84.25±0.30**
AA	62.89±1.53	68.50±1.56	71.66±0.89	73.20±1.43	71.45±0.99	66.27±2.60	67.81±1.48	72.82±0.81	69.62±1.61	70.30±1.19	**75.03±1.21**
κ	71.80±1.02	76.92±0.61	79.01±0.57	79.79±0.63	79.13±0.50	74.78±1.45	76.03±0.80	79.71±0.25	77.60±0.84	77.96±0.22	**81.14±0.38**

TABLE XIII
CLASSIFICATION RESULTS (%) FOR THE XIONGAN TESTING SET USING 1% OF THE AVAILABLE LABELED DATA

No.	ResNet (2-D | 3-D)		SSSAN	SFP	SSFTT	CBW	DMSAN	DBMA	RSSAN	S^{3}AM-Net (2-D | 3-D)	
1	53.81±21.21	40.46±17.21	70.62±5.69	74.35±16.27	73.94±13.40	63.66±20.2	67.43±11.18	81.66±20.38	58.87±14.12	60.56±17.19	**82.89±3.41**
2	93.12±3.14	91.92±1.88	95.91±2.07	**96.83±2.49**	95.78±1.34	92.20±3.11	93.55±1.74	96.33±1.40	92.88±1.17	92.83±0.90	96.26±0.80
3	77.01±26.36	94.13±1.21	93.75±2.56	90.48±1.67	93.60±2.70	90.44±2.41	94.67±1.71	93.64±2.80	94.43±2.24	93.13±0.81	**94.97±2.21**
4	88.10±8.27	96.51±2.37	92.80±4.07	96.98±0.98	94.18±6.64	92.88±6.76	95.80±2.77	92.91±8.13	97.23±0.85	90.96±7.65	**97.53±0.87**
5	60.08±30.37	72.95±14.65	75.85±14.13	85.71±7.35	73.97±5.51	67.47±14.8	76.88±9.51	**86.39±4.09**	79.52±5.30	63.12±16.57	83.83±2.81
6	63.38±17.35	61.71±8.51	66.74±6.03	65.93±7.97	65.51±10.07	**69.44±10.4**	62.43±4.92	53.87±12.86	60.04±14.93	57.56±8.75	62.96±6.73
7	40.18±29.33	52.65±29.75	74.40±9.84	72.12±21.19	76.63±4.39	72.88±18.1	71.68±20.63	61.61±20.97	26.93±11.41	**78.46±6.93**	75.41±13.73
8	83.67±6.63	82.76±4.56	89.04±1.71	88.18±3.75	87.95±2.48	82.94±6.61	87.86±5.01	89.84±3.33	83.55±3.10	90.66±2.35	**92.86±1.59**
OA	77.41±11.56	83.99±1.63	87.13±1.60	87.63±1.88	87.12±1.97	83.77±2.56	86.70±2.07	87.29±2.17	84.33±0.16	84.36±3.58	**88.95±0.87**
AA	69.92±12.24	74.26±3.65	82.39±2.63	83.82±3.64	82.69±2.27	78.99±2.50	81.29±3.21	82.03±4.53	74.18±1.08	78.41±5.32	**85.27±2.39**
κ	72.77±13.74	80.56±1.99	84.40±1.97	85.09±2.26	84.38±2.43	80.42±3.02	83.86±2.55	84.61±2.67	80.98±0.22	80.97±4.44	**86.66±1.05**

the S^{3}AM-Net model obtains the preferable classification performances.

2) Qualitative Comparison: The qualitative evaluations of different methods on four datasets are shown in Figs. 7–10. After installing the S^{3}AM, the proposed model corrects many misclassified pixels which are predicted by 2-D ResNet. Comparing with other methods, the proposed 3-D S^{3}AM-Net model gains the more accurate and smoother classification maps. For instance, almost all pixels of the categories of the "Grass-pasture-mowed" and "Oats" (C7 and C9) of the Indian Pines dataset are recognized accurately. Most pixels of the categories of the "Bare Soil" and "Bitumen" (C6 and C7, marked with white ellipses) of the Pavia University dataset are recognized by the proposed 3-D S^{3}AM-Net model correctly, which are close to the corresponding GT map. For the categories "Dense urban fabric" and "Sparsely vegetated area" (C1 and C11) of the Loukia dataset, the proposed model receives the better results. Similarly, the regions of the categories "Chinese scholartree" and "Peach" (C1 and C8) in the right part of Fig. 10(j) are purer than those of most other methods. In summary, benefits from the particular S^{3}AM, the proposed model elevates the classification accuracy on four datasets, especially in the boundaries between different land-covers.

3) Time Consumption: The training time has an inextricable link with the complexity of the network and the size of the training data. The testing time is the intuitive metric of the real-time of algorithm in real application. To evaluate the efficiency of the proposed S^{3}AM-Net model, the training and testing times of it and other compared methods on four datasets are presented in Table XIV. From this table, it is obvious that the 2-D ResNet spends the least time on accomplishing the training and testing procedures. As the 2-D CNN based models, the time consumptions of CBW, RSSAN, and 2-D S^{3}AM-Net models are less than other methods evidently. Among the 3-D CNN based models, three models based on the SA modules, i.e., SSSAN, SFP, and SSFTT, cost considerable much time to finish the optimization and testing since their SA modules which generally possess a great amount of matrix operations for attention generation. By contrast, the proposed S^{3}AM-Net model introduces few parameters to produce the spatial attention mask, which ensures its higher efficiency.

E. Analysis of the Width of HSI Cube

It is well known that the width of the HSI cube has an important effect on the classification performance. As shown in Fig. 11, the larger the width is, the higher is the OA. When the width is set to 3, the OAs on four datasets all keep the lowest level. The OAs reach the highest level and maintain stable when the widths are set to 11, 5, 7, and 5 for the four

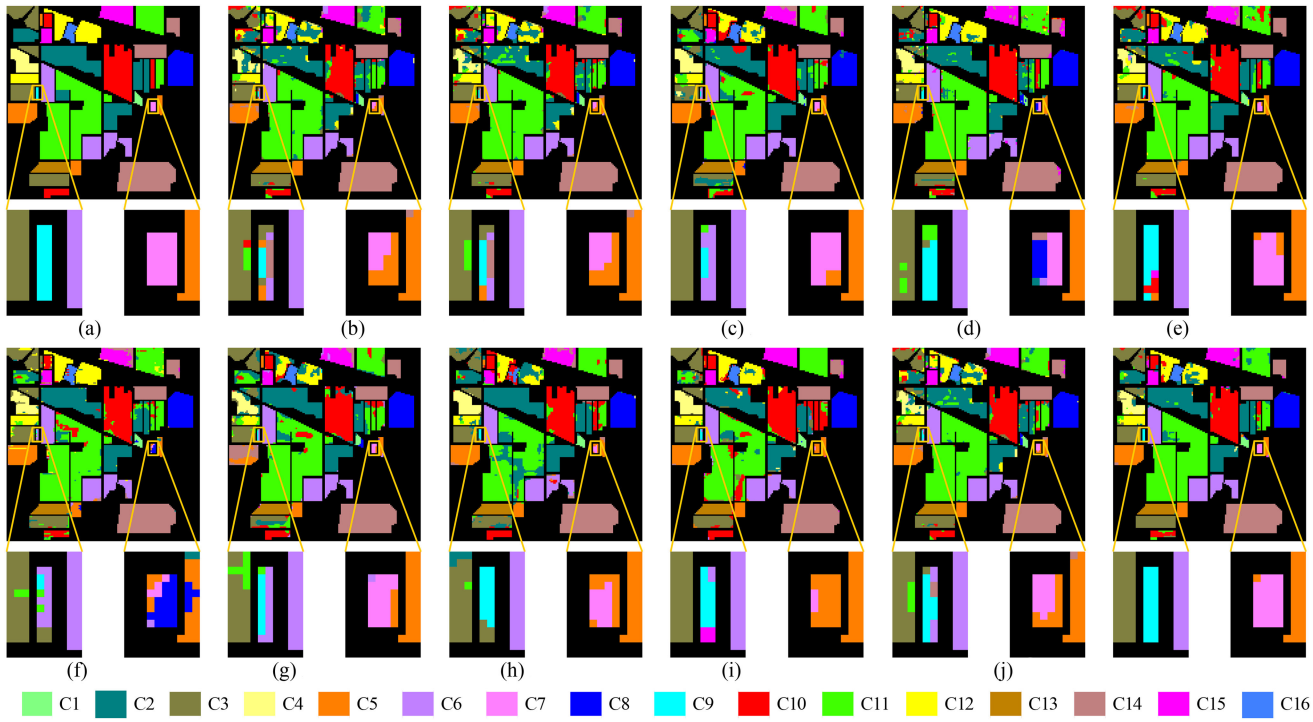

Fig. 7. Classification maps of different methods on the Indian Pines data set. Where "C*n*" represents the *n*th category. (a) GT. (b) ResNet (2-D | 3-D). (c) SSSAN. (d) SFP. (e) SSFTT. (f) CBW. (g) DMSAN. (h) DBMA. (i) RSSAN. (j) S^{3}AM-Net (2-D | 3-D).

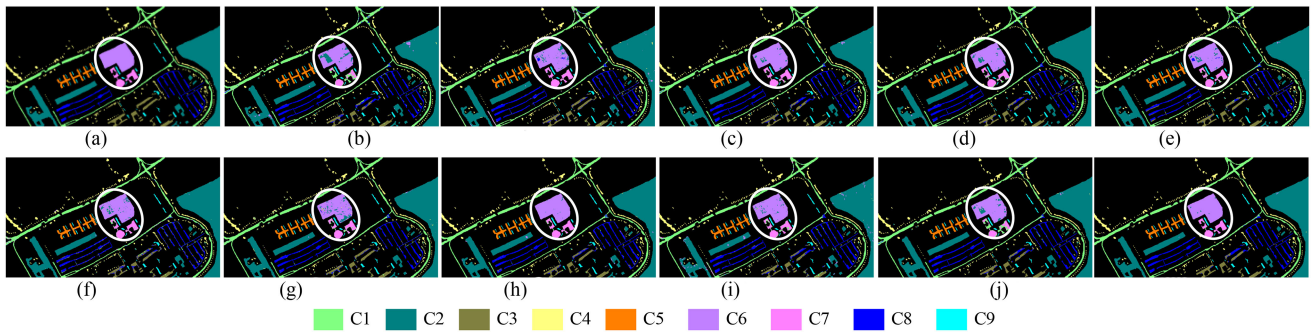

Fig. 8. Classification maps of different methods on the Pavia University dataset. (a) GT. (b) ResNet (2-D | 3-D). (c) SSSAN. (d) SFP. (e) SSFTT. (f) CBW. (g) DMSAN. (h) DBMA. (i) RSSAN. (j) S^{3}AM-Net (2-D | 3-D).

TABLE XIV
TRAINING/TESTING TIMES (SECONDS) OF DIFFERENT METHODS ON FOUR DATASETS

	ResNet (2-D | 3-D)		SSSAN	SFP	SSFTT	CBW	DMSAN	DBMA	RSSAN	S^{3}AM-Net (2-D | 3-D)	
Indian Pines	**19.40/0.65**	188.12/2.04	596.65/6.88	983.84/17.22	1243.28/20.31	73.34/0.83	584.91/6.04	465.33/8.07	92.34/1.42	24.64/0.70	193.34/2.04
Pavia Univeristy	**21.63/0.58**	156.58/3.62	167.43/5.98	430.49/11.35	602.23/13.70	49.18/0.84	359.71/7.75	133.34/3.04	61.15/1.40	25.78/0.68	164.67/3.82
Loukia	**22.07/0.36**	297.91/2.73	336.98/7.73	1107.24/15.61	1405.84/17.46	52.28/0.43	618.34/9.34	221.71/2.15	59.22/0.66	26.35/0.37	303.59/2.82
XiongAn	**15.91/0.96**	250.63/11.78	696.35/15.79	1839.54/34.18	2117.85/43.88	44.70/1.44	1035.37/16.04	178.60/8.52	49.21/2.07	21.11/1.28	258.30/12.09

datasets, respectively. Even though the HSI cube with larger width may introduce more interfering pixels, the proposed S^{3}AM-Net model still obtains the growing OAs. This is because of the characteristic of the S^{3}AM to extract the spatial attention using the spectral similarities between the center pixel and its neighborhoods, which implies that the relevant spatial areas are still well preserved even in wider HSI cubes.

F. Analysis of the Training Samples Proportion

The influence of the training samples proportion on the classification performance also cannot be neglected. As shown in Fig. 12, there are limited samples for some categories of the Indian Pines and Loukia dataset, the OAs are lower when the training proportion is less than 5%. On the contrary, the issue of

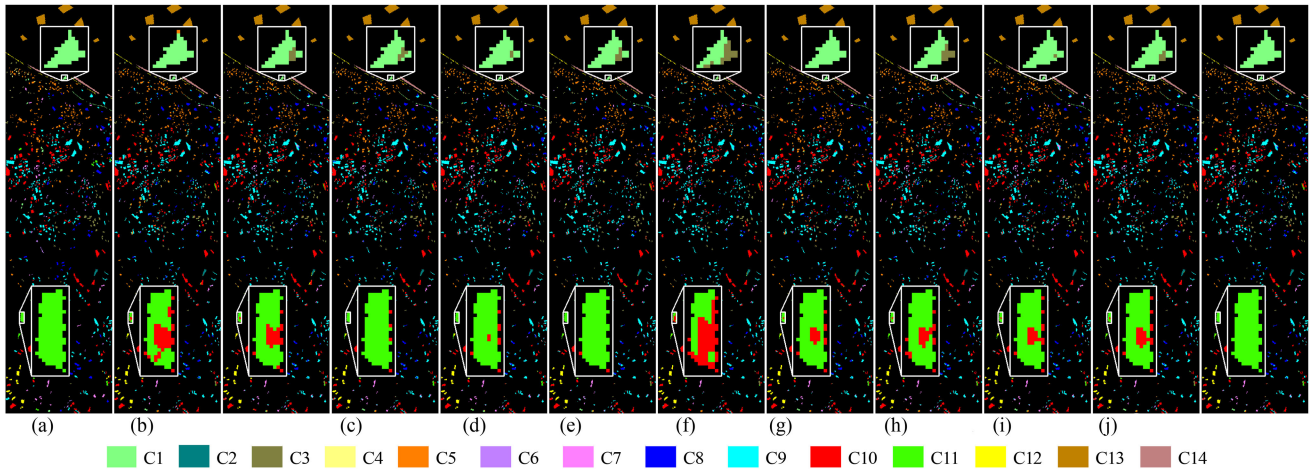

Fig. 9. Classification maps of different methods on the Loukia dataset. (a) GT. (b) ResNet (2-D | 3-D). (c) SSSAN. (d) SFP. (e) SSFTT. (f) CBW. (g) DMSAN. (h) DBMA. (i) RSSAN. (j) S^{3}AM-Net (2-D | 3-D).

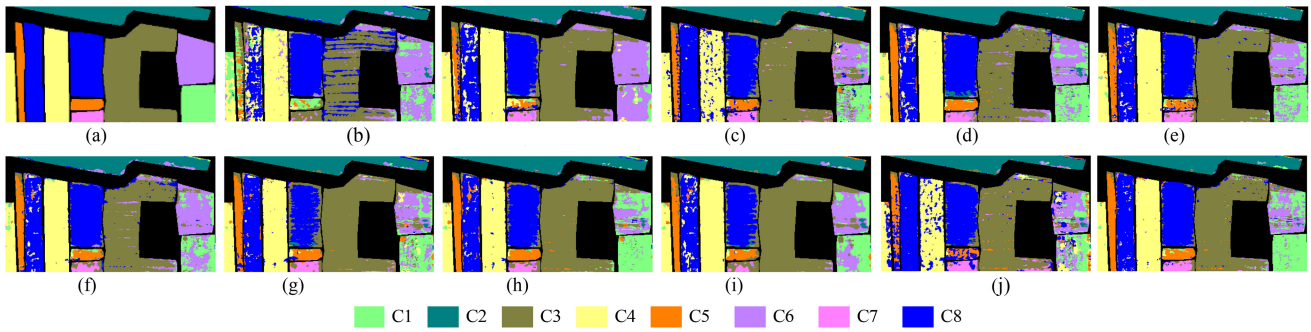

Fig. 10. Classification maps of different methods on the XiongAn dataset. (a) GT. (b) ResNet (2-D | 3-D). (c) SSSAN. (d) SFP. (e) SSFTT. (f) CBW. (g) DMSAN. (h) DBMA. (i) RSSAN. (j) S^{3}AM-Net (2-D | 3-D).

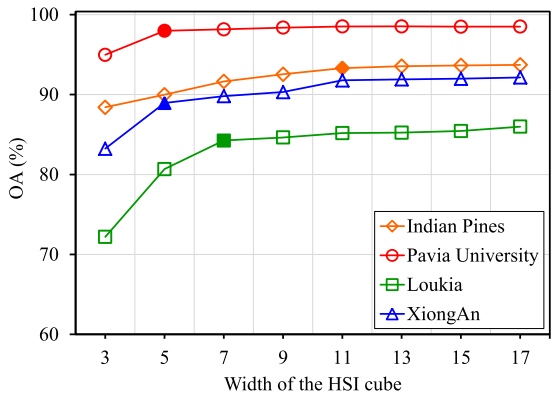

Fig. 11. OAs (%) of the proposed model using different widths on four datasets. Where the solid data markers mean the optimal configurations.

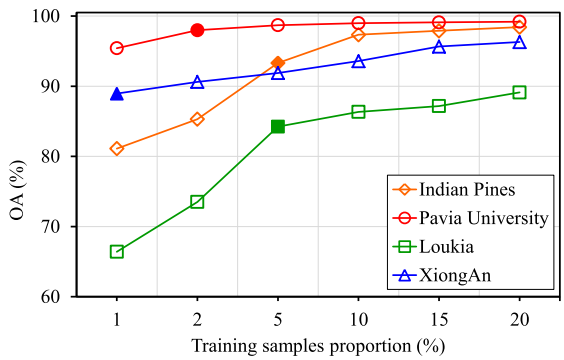

Fig. 12. OAs (%) of the proposed model with different training samples proportions on four datasets.

the limited samples is relieved in other two datasets. Hence, the OAs of them have attained the accuracy of no less than 95% and 85% when there are one percent samples for training merely. To acquire better classification performances and expend less training time, the points where the curves tend to be steady, i.e., 5%, 2%, 5%, and 1%, are selected as the training samples proportions of the Indian Pines, Pavia University, Loukia, and XiongAn datasets, respectively.

G. Comparison Between Different Attention Modules

In this part, so as to interpret the validity of the S^{3}AM, it is replaced with other three kinds of common attention modules, including the SA module [35], the SE module [42], and the CBAM [47], to construct the SA-Net, SE-Net, and CBAM-Net models. From Table XV, it can be observed that the OAs of other three kinds of attention networks on four datasets are lower than those of the proposed S^{3}AM-Net model. The SA module

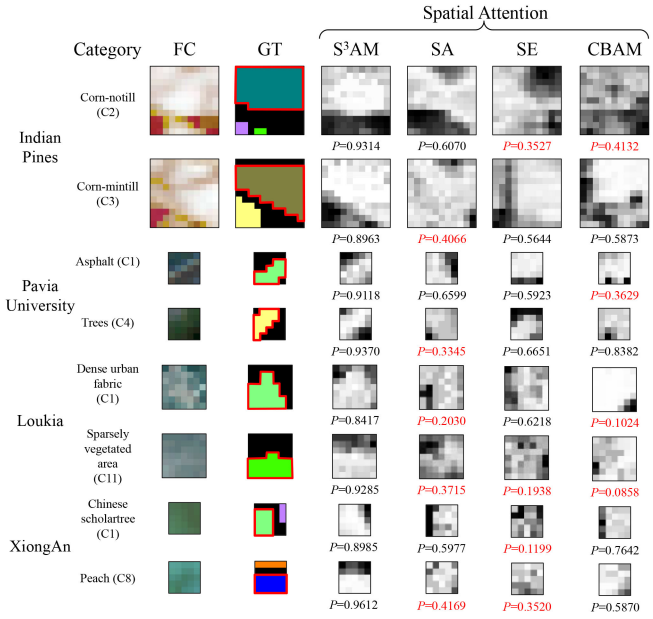

Fig. 13. Visualization of the spatial attention masks generated by the proposed S^{3}AM and other three kinds of attention modules. The relevant spatial areas are marked with red solid polygons. Where P indicates the softmax score of each network for the corresponding category.

TABLE XV
OAs (%) OF THE DIFFERENT ATTENTION NETWORKS ON FOUR DATASETS

	Indian Pines	Pavia University	Loukia	XiongAn
S^{3}AM-Net	**93.31±0.54**	**97.99±0.28**	**84.25±0.30**	**88.95±0.87**
SA-Net	90.43±0.47	96.52±0.34	82.86±0.97	86.94±0.75
SE-Net	88.22±0.91	94.50±0.48	80.23±0.72	86.72±1.19
CBAM-Net	89.76±0.79	96.79±0.65	82.92±1.44	87.15±1.02

explores the correlation between pixels. The SE module adopts the spatial gather and excitation to model the dependency existing in different pixels. The CBAM employs the large-scale convolution to describe the meaningful spatial positions. However, owing to the deficiency of the special operation for the center pixel, the procedures of the three attention modules may be affected by the interfering pixels. Therefore, they are weak in locating the relevant pixels, which are advantageous to classification exactly. Extraordinarily, the S^{3}AM infers the relevant areas via the spectral similarities between the center pixel and its neighborhoods, which facilitates the representation of the pivotal features thereby ensuring better classification performances.

H. Attention Visualization

To clearly illustrate the advantage of the proposed S^{3}AM comparing with other attention modules, including the SA module [38], the SE module [46], and the CBAM [49], their spatial attention masks extracted from several testing samples are visualized in Fig. 13. It shows the FC image, the GT map, and the spatial attention masks of each sample. From this figure, it can be observed that the spatial attention masks generated by the S^{3}AM can assign the highest spatial weights to the center pixels and describe the relevant spatial areas, which are marked in the corresponding GT maps more exactly. This leads to the high softmax scores P predicted by the S^{3}AM-Net model. In contrast, other attention modules, which extract the spatial attention via the dot-product operations, the feature squeezing and excitation, and the large-scale convolution, tend to mistake the interfering pixels for the relevant pixels or even suppress the center pixels. This may cause incorrect predictions with low softmax score P (marked in red). Therefore, the S^{3}AM does better in enhancing the relevant areas and improving the extraction of the discriminating spectral-spatial features.

IV. Conclusion

In this article, a novel S^{3}AM is proposed to emphasize the center pixel and capture the relevant spatial areas, which are beneficial to the classification. The S^{3}AM is composed of the WED submodule, the WCD submodule, the adaptive aggregation, and the SG activation function. First, the WED and WCD are exploited to obtain robust spectral similarities between the center pixel and its neighborhoods. In particular, they all employ the FBC layers to recalibrate each band to weaken the influence of the spectral variability. Next, the adaptive aggregation fuses the two kinds of spectral similarities into a comprehensive spectral similarity by considering their relative importance. Finally, the SG activation function, which can weaken the interfering pixels according to the specificity of different scenes, takes charge of converting the spectral similarity into the proper spatial weights. The S^{3}AM is good at emphasizing the spatial areas relevant to the center pixel and maintaining these areas even in wider HSI cubes. As a flexible component, the S^{3}AM is integrated with the 3-D ResNet to build the S^{3}AM-Net model, which can extract the discriminating spectral-spatial features from the relevant spatial areas of HSI cubes. Experimental results on four public datasets demonstrate the rationality and effectiveness of the S^{3}AM and the superior classification performances of the S^{3}AM-Net model compared with other state-of-the-arts.

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] 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.
[3] G. Notesco, Y. Ogen, and E. Ben-Dor, "Mineral classification of makhtesh roman in Israel using hyperspectral long-wave infrared (LWIR) remote sensing data," *Remote sens*, vol. 7, no. 9, pp. 12282–12296, Sep. 2015.
[4] R. Anand, S. Veni, and J. Aravinth, "Big data challenges in airborne hyperspectral image for urban landuse classification," in *Proc. Int. Conf. Adv. Comp., Com. Inf.*, 2017, pp. 1808–1814.
[5] 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.
[6] 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.
[7] 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.
[8] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2015, pp. 1440–1448.

[10] A. Bordes, X. Glorot, J. Weston, and Y. Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in *Proc. 15th Int. Conf. Art. Intell. Statist.*, 2012, pp. 127–135.

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

[12] 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.

[13] 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.

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

[15] 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.

[16] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neur. Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[17] Y. Duan, H. Huang, and T. Wang, "Semisupervised feature extraction of hyperspectral image using nonlinear geodesic sparse hypergraphs," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5515115.

[18] F. Luo, Z. Zou, J. Liu, and Z. Lin, "Dimensionality reduction and classification of hyperspectral image via multistructure unified discriminative embedding," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517916.

[19] A. Krizhevsky et al., "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neur. Inf. Process. Syst.*, 2012, pp. 1097–1105.

[20] 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.*, 2020, pp. 88–92.

[21] J. Yang, Y. Zhao, and J. C. 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.

[22] 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.

[23] M. Ahmad, A. M. Khan, M. Mazzara, S. Distefano, M. Ali, and M. S. Sarfraz, "A fast and compact 3-D CNN for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art no. 5502205.

[24] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

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

[26] W. -S. Hu, H. -C. Li, Y. -J. Deng, X. Sun, Q. Du, and A. Plaza, "Lightweight tensor attention-driven ConvLSTM neural network for hyperspectral image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 734–745, Apr. 2021.

[27] Z. Mei, Z. Yin, X. Kong, L. Wang, and H. Ren, "Cascade residual capsule network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3089–3106, 2022.

[28] B. Yang, F. Cao, and H. Ye, "A novel method for hyperspectral image classification: Deep network with adaptive graph structure integration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5523512.

[29] T. Jiang, Y. Li, W. Xie, and Q. Du, "Discriminative reconstruction constrained generative adversarial network for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4666–4679, Jul. 2020.

[30] 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.

[31] X. Zhang et al., "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.

[32] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.

[33] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–9.

[34] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6450–6458.

[35] Z. Lin et al., "A structured self-attentive sentence embedding," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–15.

[36] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[37] 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.

[38] X. Zhang et al., "Spectral-spatial self-attention networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5512115.

[39] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neur. Inf. Process. Syst.*, 2017, pp. 5998–6008.

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

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

[42] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[43] 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.

[44] N. Li and Z. Wang, "Spatial attention guided residual attention network for hyperspectral image classification," *IEEE Access*, vol. 10, pp. 9830–9847, 2022.

[45] J. Hu et al., "Gather-excite: Exploiting feature context in convolutional neural network," in *Proc. Adv. Neur. Inf. Process. Syst.*, 2018, pp. 1–11.

[46] H. Gao, Y. Miao, X. Cao, and C. Li, "Densely connected multiscale attention network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2563–2576, 2021.

[47] S. Woo, J. Park, J. -Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 3–19.

[48] 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. 6, pp. 1307–1329, Jun. 2019.

[49] 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.

[50] W. Guo, H. Ye, and F. Cao, "Feature-grouped network with spectral-spatial connected attention for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5500413.

[51] M. A. Cho, P. Debba, R. Mathieu, L. Naidoo, J. van Aardt, and G. P. Asner, "Improving discrimination of savanna tree species through a multiple-endmember spectral angle mapper approach: Canopy-level analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4133–4142, Nov. 2010.

[52] 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.

[53] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectified neural networks," in *Proc. 14th Int. Conf. Art. Intell. Stat.*, Jan. 2011, vol. 15, pp. 315–323.

[54] "Hyperspectral remote sensing scenes - Grupo de inteligencia computacional (GIC)." Accessed: Nov. 28, 2021. [Online]. Available: http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

[55] K. Karantzalos, C. Karakizi, Z. Kandylakis, and G. Antoniou, *HyRANK Hyperspectral Satellite Dataset I. (Version V001)* [data set], Apr. 2018.

[56] Y. Cen et al., "Aerial hyperspectral remote sensing classification dataset of Xiongan new area (Matiwan village)," *J. Remote Sens. (Chin.)*, vol. 24, no. 11, pp. 1299–1306, Nov. 2020.

[57] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural network," in *Proc. 13th Int. Conf. Art. Intell. Statist.*, 2010, pp. 249–256.

[58] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Net. Mach. Learn*, vol. 4, no. 2, pp. 26–31, 2012.

Ningyang Li received the B.S. degree in remote sensing science and technology from Henan Polytechnic University, Jiaozuo, China, in 2019. He is currently working toward 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, Derby, U.K., in 2004, and the Ph.D. degree in color and imaging science from the University of Leeds, Leeds, U.K., in 2008.

He was with the Norwegian Colour and Visual Computing Laboratory, Gjøvik, Norway, to work on visual computing, multispectral color imaging research projects. He was with Hainan University, China, in 2013. He is currently a Professor of computer science with 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.

Dr. Wang professional memberships include IS&T, SPIE, IEEE, and CCF.

Faouzi Alaya Cheikh (Senior Member, IEEE) received a B.Sc. degree in electronics from l'Ecole Nationale d'Ingenieurs de Tunis, Tunis, Tunisia, in 1992, the M.Sc. degree in signal processing, and the Ph.D. degree in information technology from the Tampere University of Technology, Tampere, Finland, in 1997 and April 2004, respectively.

Since 1994, he has been a Researcher with the Signal Processing Algorithm Group, Tampere University of Technology, Tampere, Finland. Since 2006, he has also been with the Department of Computer Science and Media Technology, Gjøvik University College, Gjøvik, Norway, as an Associate Professor. Since January 2016, he has also been with the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. He teaches courses on image and video processing and analysis, and media security. He is currently the cosupervisor of nine Ph.D. students. He has been involved in several European and national projects among them ESPRIT, NOBLESS, COST 211Quat, HyPerCept, IQ-Med, and H2020 ITN HiPerNav. He has authored and coauthored more than 150 peer-reviewed journal and conference articles, and supervised four postdoctoral researchers, nine Ph.D. and a number of M.Sc. thesis projects on the topics of his research interests, which include e-Learning, 3-D imaging, image and video processing and analysis, video-based navigation, biometrics, pattern recognition, embedded systems, and content-based image retrieval.

Prof. Cheikh is a member of NOBIM and Forskerforbundet (The Norwegian Association of Researchers-NAR). He is on the Editorial Board of the *IET Image Processing* Journal and the Editorial Board of the *Journal of Advanced Robotics and Automation*, and the technical committees of several international conferences. He is an expert reviewer to a number of scientific journals and conferences related to the field of his research.

Mohib Ullah (Member, IEEE) received the bachelor's degree in electronic and computer engineering from the Politecnico di Torino, Torino, Italy, in 2012, the master's degree in telecommunication engineering from the University of Trento, Italy, in 2015, and the Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 2019.

He is currently a Postdoctoral Research Fellow with NTNU, where he is involved in research, management, teaching, and industrial projects. He has authored and coauthored several high-impact peer-reviewed journals, conferences, and workshop articles on the topics of his research interests, which include medical imaging, crowd analysis, object segmentation, behavior classification, and tracking.

Dr. Ullah was a Program Committee Member for the International Workshop on Computer Vision in Sports. He was a Chair for the Technical Program at the European Workshop on Visual Information Processing. He is the Guest Editor of the applied sciences journal. He is the Reviewer of many conferences and journals (*Neurocomputing* (Elsevier), *Neural Computing and Applications* (Elsevier), *Multimedia Tools and Applications* (Springer), IEEE ACCESS, the *Journal of Imaging*, IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, IEEE International Conference on Image Processing, and IEEE International Conference on Advanced Video and Signal-Based Surveillance).