Siamese Spectral Attention With Channel Consistency for Hyperspectral Image Classification

Leiquan Wang, *Member, IEEE*, Yao Lin, Jinyun Liu, Zhongwei Li, and Chunlei Wu, *Member, IEEE*

***Abstract*—Abundant spectral features are the precious wealth of hyperspectral images (HSI). Nevertheless, well-designed spectral feature is still a challenge that affects the performance of the classifier, especially with insufficient number of training samples. To make up the poor discriminability of spectral feature, double-branch methods are proposed by fusing parallel spectral and spatial branches. However, this structure does nothing to improve the quality of spectral feature, which is regarded as the most valuable information for HSI information. In this article, we propose a siamese spectral attention network with channel consistency (SSACC) to focus on obtaining discriminative spectral features, thus improving the generalization ability of the classifier. Two kinds of HSI cubes with different patch sizes are generated as the input of SSACC. The two cubes are divided into top and bottom branches and then be fed into the siamese network to obtain the refined spectral features. Then, self-attention is conducted to interacting with each channel for the spectral features enhancement. Meanwhile, two attention maps are obtained to display the spectral structures of each branch. A channel consistency regularization is performed on the two attention maps by enforcing the two branches to possess similar spectral patterns when identifying the same centric pixel. Extensive experiments conducted on the three HSI datasets verify the superiority of the obtained spectral feature. Furthermore, the proposed method applying convolution only on the spectral domain outperforms the state-of-the-art double-branch methods which integrate the spectral and spatial features simultaneously.**

***Index Terms*—Channel consistency, double-branch, hyperspectral image (HSI) classification, spectral siamese.**

I. Introduction

As a special remote sensing, hyperspectral image (HSI) integrates the unique advantages of spectrograph and optical cameras, owning high-resolution spectral signature and large-scale spatial information. The task of HSI classification is to assign a specific land-cover label to each hyperspectral pixel. By analyzing the spectral signature and spatial information, HSI classification possesses powerful discriminability on observation objects with wide application in earth observation tasks, such as agriculture estimation [1], atmospheric environment forecast [2], water quality monitoring [3], ocean species identification [4], and urban development [5]. However, due to the well-known Hughes phenomenon and the curse of dimensionality [6], the exploitation of redundantly continuous spectral and spatial information remains a hot yet challenging topic in remote sensing field [7], [8].

Conventionally, HSI is described as a 3-D cube data, involving 1-D spectral signatures and 2-D spatial information [9]. According to the type of information utilized, HSI classification can be roughly categorized into spectral-based methods and spectral–spatial-based methods. HSI contains abundant spectral signatures owing to its hundreds of narrow contiguous wavelength bands. Each pixel in HSI is represented as a signature, encoding plentiful physical properties. Spectral-based approaches primarily focus on the spectral signatures, taking an original pixel as the input, such as logistic regression [10], linear discriminant analysis [11], and support vector machine (SVM) [12]. The aforementioned methods are of shallow-layer, with limited representation capacity to handle complex and new situations. With the breakthrough of deep learning (DL) [13]–[16], deep models have also been exploited to incorporate plentiful spectral signatures. Hu *et al.* [17] treated the spectral signature as a 1-D signal and conducted the 1-D convolutional neural network on spectral domain for HSI classification. Mou *et al.* [18] handled the hyperspectral pixel as sequential data and performed the recurrent neural network (RNN) to infer the semantic label. However, they suffered from the lack of adequate training samples to fit the data distribution. Therefore, the trained models often lead to poor generalization and, thus, are sensitive to the disturbance on spectral signatures. It has been common to increase the amount of training data in a disguised form to address the issue. With the aid of pixel-pair, Li *et al.* [19] augmented the training data by a wide margin so as to maintain the advantage of CNN. Similarly, generative models have also been explored to handle the sore point of insufficient labeled HSI pixels [20], [21], while only an individual pixel was used during the testing phase. The nature of spectral variability that is susceptible to atmospheric effects, instrument noises, and incident illumination, has not been resolved effectively. Different from common RGB image classification, HSI classification is worthy of a deep plowing on spectral signatures.

Spectral–spatial-based methods incorporate spatial information to complement the spectral signatures for HSI

Manuscript received June 16, 2021; revised July 29, 2021; accepted September 20, 2021. Date of publication September 24, 2021; date of current version October 18, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 62071491, in part by the Joint Funds of the National Natural Science Foundation of China under Grant U1906217, and in part by the Fundamental Research Funds for the Central Universities, under Grant 19CX05003A-11. *(Corresponding author: Zhongwei Li.)*

Leiquan Wang and Chunlei Wu are with the College of Computer Science and Technology, China University of Petroleum, Qingdao 266555, China (e-mail: richiewlq@gmail.com; wuchunlei@upc.edu.cn).

Yao Lin and Zhongwei Li are with the College of Oceanography and Space Informatics, China University of Petroleum, Qingdao 266555, China (e-mail: 2795561928@qq.com; lizhongwei@upc.edu.cn).

Jinyun Liu is with the SINO-Pipeline International Company Limited, Beijing 100028, China (e-mail: jinyun.liu@cnpc.com.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3115129

classification. By holding the label coherence of adjacent pixels, spectral–spatial-based methods use a pixel centric 3-D cube as the input, whose label is determined by its centric pixel [22]. Up to now, CNN has been seen as one of the most effective ways to extract spectral–spatial features by modeling the relationship of adjacent pixels. Chen *et al.* [23] proposed a 3-D CNN-based model with combined regularization to extract effective spectral–spatial features. Lee *et al.* [24] adopted a multiscale convolutional filter bank to explore local contextual interactions by jointly exploiting local spectral–spatial relationships of neighboring pixels. With the CNN architecture evolution, ResNet [25], CapsuleNet [26], and DenseNet [27] have been introduced into HSI classification to obtain discriminative spectral–spatial features. However, these state-of-the-art CNN-based methods adopted a single 3-D cube input style, subject to the fixed patch size. Zhang *et al.* [28] exploited diverse region-based inputs to investigate the contextual interactional spectral–spatial features to alleviate these restricts, while it still adheres to the adjacent pixels coherence assumption, where the adjacent pixels are assumed to share the same labels [29]. Besides, LSTMs have also been incorporated to manage the dependencies among the dense spectral–spatial sequences [30]. Zhou *et al.* [31] regarded each hyperspectral data as data sequences and use LSTM to model the dependency in the spectral and spatial domains, respectively. Hu *et al.* [32] proposed a spatial-spectral ConvLSTM 3-D neural network by extending LSTM to the 3-D version to preserve the intrinsic structure information in the hyperspectral data. However, the fundamental problem of exploiting spatial information, the negative influence of interfering pixels in the 3-D cube, whose label is different from that of the centric pixel, remains untouched. Consequently, most of spectral–spatial-based methods perform significantly better in homogenous regions than in heterogeneous regions. Totally different land-cover labels may be obtained due to the variations on the patch size of the same centric pixel. Scholars all deem the spatial information as the complementary to the spectral signatures, as proved by the classification accuracy. However, there are not explicit explanations on the exact role of spatial information and whether it is irreplaceable to obtain high-quality classifiers.

The superpixel is also adopted for spectral–spatial-based HSI classification by taking segmented superpixel as input to alleviate the interference of interfering pixels [33]–[35]. Superpixel-based methods are subject to the preselected superpixel segmentation algorithm. Moreover, different neighborhoods should make differentiated contributions to the centric pixel recognition. How to emphasize informative pixels and suppress interfering pixels in spatial region is a challenging yet hot topic in HSI field [36]–[38]. Inspired by the human visual perception, attention mechanism (AM) has also been encoded into HSI classification, which selectively attend to the most task-relevant parts of the input signal [39]. By highlighting discriminative features, AM aided CNN model shows the superiority of HSI classification on both spectral and spatial domains. Given that not all bands are equally informative and predictive for HSI classification, Mou *et al.* [40] designed spectral attention module to adaptively recalibrate spectral bands by selectively emphasizing informative bands and suppressing less useful ones. By merging spatial information, a spectra-wise AM with 3-D patches was introduced to enhance the distinguishability of spectral features by Fang [41]. In this study, although spatial information was considered, in influence of interfering pixels was not mentioned [41]. Zhu *et al.* [22] proposed spectral–spatial attention network by cascading spectral AM and spatial AM in sequential, which emphasizes useful bands and pixels simultaneously. Two-branch spectral–spatial attention networks, such as SSAtt [37], DBMA [42], and DBDA [43], were also proposed for HSIC by exploiting spectral attention subnetwork and spatial attention subnetwork separately. However, the two subnetworks have no necessary interaction until the eventual combination. In addition, most of AM-based methods intermix spatial pixels by performing 2-D or 3-D CNN. The complementarity between the spectral and spatial branches is well exploited. However, the exact relationship between spatial and spectral information has not been well investigated yet.

Through the review of the abovementioned literature, some problems arise as follows.

1) Is it feasible and reliable to identify the land-cover labels using only spectral signature without dealing with complex spatial information? If so, how to extract discriminative and robust spectral features with fewer training samples?
2) Is it necessary to perform a subtle identification of every pixel located in the 3-D input cube to obtain prominent label separability. If not, how to alleviate the influence of harmful neighborhood pixels?

An end-to-end siamese spectral attention network with channel consistency (SSACC) for HSIC is proposed to address the abovementioned problems. The goal of this article is to improve spectral representation ability by exploring the correlations within a continuous spectrum aided by the adjacent spatial information. Different from previous spectral and spatial attention-based two-branch methods, the proposed SSACC applies only spectral attention to the multipatches with different scales. The multiscale patches are of the same centric pixel and with the same semantic labels. The SSACC is proposed based on the assumption that different patches with the same centric pixel may pay close attention to the spectral channels (see Fig. 1). From the perspective of AM, the emphasized spectral channels should be as similar as possible. The spatial information is used to reduce the interference of spectral representability. Specifically, a pixel can be represented as two pixel centric 3-D cubes with different patch sizes that can act as different branches to capture spectral features, respectively, by performing siamese spectral networks. Spectral AMs are then performed on the two branches to adaptively learn the weights of each spectral channels. The interactions between the branches are established by conducting the channel consistency assumption on the attention maps. The channel consistency term is exploited on the two branches to promote the robustness of learned spectral features for HSI classification. Different from traditional spectral–spatial-based methods, all convolutions are conducted on the spectral domain with the kernel size of $1 * 1 * d$ to keep the original spatial relationship without any transformation. As

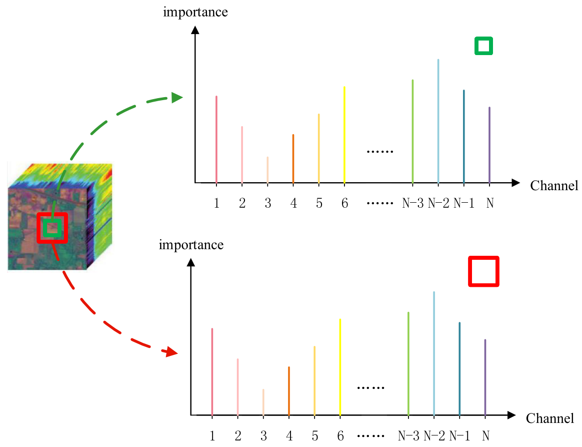
importance
Channel
1 2 3 4 5 6 ...... N−3 N−2 N−1 N
importance
Channel
1 2 3 4 5 6 ...... N−3 N−2 N−1 N

Fig. 1. Schematic diagram of spectral structure for the proposed methods. The input cubes for the same centric pixel with similar spatial size should share the similar spectral structure when identifying the same land-cover. The spectral consistency may improve the generalization ability of network, alleviating the dependency on the number of labeled samples.

an auxiliary information, neighborhood pixels are explored to enhance the discriminative spectral features.

The main contributions of this article are as follows.

1) A siamese spectral attention network is proposed to establish the implicit interaction between two branches. The siamese strategy is exploited to allow the two branches to possess the identical network structure and share the same parameters. The convolution operations are conducted only on the spectral dimension, and the subtle identification of every pixel is not required to alleviate the influence of interfering pixels.
2) The channel consistency term is proposed to establish the explicit interactions between two different spectral branches. The channel consistency is performed to enforce the two branches to possess similar spectral patterns when identifying the same centric pixel.
3) Extensive experiments are conducted on the three public HSI datasets. Experimental results demonstrate that the proposed SSACC achieves the best performance with fewer parameters compared with the state-of-the-art two-branch networks.

The rest of this article is organized as follows. Section II introduces the related work and Section III describes the proposed SSACC in detail. Next, the experimental results and comprehensive analysis are given in Section IV. Finally, Section V concludes this article.

II. Related Work

In this section, the background information of AM is reviewed and a summary of the double-branch methods in HSI classification is presented.

A. Attention Mechanism

The sophisticated data processing capability of humans enables them to perceive the information efficiently and achieves precise consciousnesses and awareness. Inspired by the human perceptions, the AM is exploited to selectively focus on the informative elements and ignore the irrelevant contents, which has been a popular concept in the DL community in recent years. Various AMs (e.g., self-attention [44], CBAM [45], nonlocal [46], GCA [47], ECA [48]) are proposed to handle different tasks (e.g., image caption [49], object localization [50], and image classification [51]). No matter what AMs is, the crucial issue lies in the identification of the parts worthy more attention of the task. In the HSIC community, the AM has also been widely used to selectively focus on important spectral of spatial information. Mou *et al.* [40] proposed a learnable spectral attention module that explicitly allows the spectral manipulation of hyperspectral data within a CNN. However, the model did not take spatial information into consideration. Pan *et al.* [52] proposed to combine bi-RNN-based spectral attention and CNN-based spatial attention. For spectral domain, each pixel was represented as a continuous spectral curve. Different attention weights were assigned by modeling relationships of inner channels. For spatial domain, spatial features are regarded as a complementary to spectral ones, where the inner-spatial dependency were exploited to support spatial attention. Gao *et al.* [53] added the AM into the preactivation residual block. Sun *et al.* [54] make a attention module that can be embedded anywhere in the spectral module and spatial module for HSIC. Based on this model, Lu *et al.* [55] proposed a 3-D attention module that consists of a channel attention module and a spatial attention module. Swalpa *et al.* [56] proposed an attention-based adaptive spectral–spatial kernel module that was introduced for the first time to learn selective 3-D convolutional kernels for HSIC. The abovementioned methods have achieved satisfactory results by applying AMs to spectral or spatial domain. However, this increased demands for training samples meet the need of complex networks, which is a challenge for HSIC at this stage.

B. Double-Branch Methods

In the HSIC community, the double-branch is a representative structure by taking advantage of the complementarity of spectral and spatial information. Since HSIs have not only certain spatial information but also rich spectral information, Xu *et al.* [57] proposed a band grouping-based LSTM and a multiscale CNN as the spectral and spatial feature extractors, respectively. Zhong *et al.* [25] developed two consecutive residual blocks to learn spectral and spatial representations separately, through which discriminative features can be extracted. Wang *et al.* [58] proposed an end-to-end fast and dense spectral–spatial convolution network framework for HSI classification. Ma *et al.* [42] proposed a double-branch multiattention mechanism network (DBMA) for HSI classification. With two branches to extract spectral and spatial feature, respectively, this network can reduce the interference between the two types of features. Deng *et al.* [59] incorporated active learning into double-branch network, where the learned deep joint spectral–spatial features are more generic and robust. Hao also proposed a double-branch methods for HSIC, with one branch employing a stacked denoising autoencoder to encode the spectral signatures and the

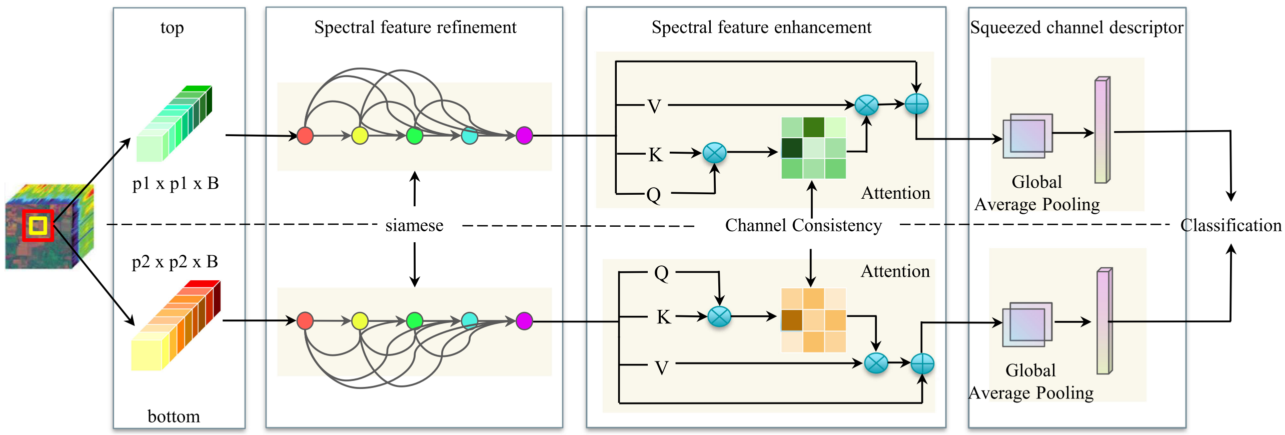

Fig. 2. Overall framework of SSACC. The siamese strategy is used in the top and bottom branches of spectral feature refinement module, aiming to capture effective spectral features with $1 * 1 * d$ convolutions. The spatial information is incorporated in the spectral feature enhancement module, where the channel consistency is used to promote the robustness of spectral features, meanwhile, suppress the influence of interfering pixels. The channel consistency can also be regarded as the constraint strategy of the siamese structure. Finally, the refined features are squeezed after global average pooling for classification.

other branch exploiting a CNN to deal with the corresponding HSI cubes. However, most of the methods mentioned above incorporated the spectral branch and spatial branch separately, which simply aims to improve the accuracy of classification, and they failed to improve the quality of spectral or spatial features. The aim of this article is to obtain a highly discriminative spectral feature by establishing the interactions between the separated branches.

III. Methodology

A. Problem Formulation

Given an HSI dataset, it is denoted as $\mathcal{X} = \{x_1, x_2, \ldots, x_{HW}\} \in \mathbb{R}^{H\times W\times B}$, where H and W are the spatial height and weight, respectively, and B is the band number of spectral signature. Consequently, the total number of pixels in $\mathcal{X}$ is represented as $N = HW$. Without loss of generality, the first N_l pixels are randomly sampled from each land-cover category, where $N_l \ll N$. Let $\mathcal{X}_\mathcal{L} = \{x_1, x_2, \ldots, x_{N_l}\}$ be the available labeled pixel set, and the land-cover label set that corresponds to $\mathcal{X}_\mathcal{L}$ be denoted as $\mathcal{Y}_\mathcal{L} = \{y_1, y_2, \ldots, y_{N_l}\} \in \mathbb{R}^{1\times 1\times L}$, where L is the number of land-cover categories with $y_i \in \{1, 2, \ldots, l\}$. The unlabeled pixel set is then represented as $\mathcal{X}_\mathcal{U} = \{x_{N_l+1}, x_{N_l+2}, \ldots, x_N\}$. The task of HSI classification on X is to assign a proper land-cover label to each pixel $x_i \in \mathcal{X}_\mathcal{U}$ by building HSIC models on $(\mathcal{X}_\mathcal{L}, \mathcal{Y}_\mathcal{L})$.

B. Overview of SSACC

The motivation of this article is to extract discriminative spectral feature to alleviate the spectral mixing effect by introducing channel consistency hypothesis, which refers to that the same material may appear with different spectral or different materials may have the same spectral signatures [60]. Therefore, the double-branch spectral attention network is exploit based on the channel consistency hypothesis.

The introduced channel consistency hypothesis is:

When ϵ is infinitely small, a centric pixel x_i with different window length p and $p + \epsilon$ may show similar spectral patterns and obtain the same results on identifying and classifying the land-cover categories.

In order to clearly illustrate the proposed SSACC, the overall framework is shown in Fig. 2. In general, SSACC combines the double-branch strategy with channel consistency to avoid spectral redundancy and achieve better class separability. As shown in Fig. 2, the proposed SSACC contains two parallel branches, and both branches have the same modules. Specifically, the framework is split into the following five modules.

1) *Dataset generation*. Preparations are made to generate two different sized patch sets for the inputs of the model.
2) *Spectral feature refinement*. The highly correlated spectral signatures are generally refined with the dense block [61].
3) *Spectral feature enhancement*. The spectral feature is enhanced by highlighting the key channel with the application of AM.
4) *Channel consistency regularization*. A special bond is established to link the top and bottom branches by modeling the channel consistency hypothesis.

The classification losses from the two branches and channel consistency loss are integrated into a unified network for the end-to-end training. In the inference stage, the final prediction results are obtained by fusing predictions from the top and down branches, without considering channel consistency. In the following sections, each module is be presented in detail.

C. Dataset Generation

In the CNN-based HSIC models, each pixel $x_i \in \mathcal{X}$ is routinely cropped into a square box with fixed length p_1 to generate a 3-D cube set $Z = \{z_1, z_2, \ldots, z_N\} \in \mathbb{R}^{p_1\times p_1\times B}$ with x_i as the central pixel vector. The label of HSI cube z_i is considered as the same as that of x_i, i.e., y_i. The labels of other neighboring pixels in the $p_1 \times p_1$ centric window are unknown. Different from traditional CNN-based HSIC methods, another 3-D cube set $\widetilde{Z}$ is established by a new square box during the dataset generation phase, whose length is a tiny increment ϵ on p_1, denoted as $p_2 = p_1 + \epsilon$. Based on the original split principle on $\mathcal{X} = \{\mathcal{X}_{\mathcal{L}}, \mathcal{X}_{\mathcal{U}}\}$, the training set is generated as $Z_{\text{train}} = \{Z_{\mathcal{L}}, \widetilde{Z}_{\mathcal{L}}, Y_{\mathcal{L}}\}$ and the validation set is generated as $Z_{\text{val}} = \{Z_{\mathcal{V}}, \widetilde{Z}_{\mathcal{V}}, Y_{\mathcal{V}}\}$ and the testing set is represented as $Z_{\text{test}} = \{Z_{\mathcal{U}}, \widetilde{Z}_{\mathcal{U}}\}$, where $Z_{\mathcal{L}} = \{z_1, z_2, \ldots, z_{N_l}\} \in \mathbb{R}^{p_1 \times p_1 \times B}$, $\widetilde{Z}_{\mathcal{L}} = \{\widetilde{z}_1, \widetilde{z}_2, \ldots, \widetilde{z}_{N_l}\} \in \mathbb{R}^{p_2 \times p_2 \times B}$, $Z_{\mathcal{V}} = \{z_{N_l+1}, z_{N_l+2}, \ldots, z_{N_l+l}\} \in \mathbb{R}^{p_1 \times p_1 \times B}$, $\widetilde{Z}_{\mathcal{V}} = \{\widetilde{z}_{N_l+1}, \widetilde{z}_{N_l+2}, \ldots, \widetilde{z}_{N_l+l}\} \in \mathbb{R}^{p_2 \times p_2 \times B}$, $Z_{\mathcal{U}} = \{z_{N_2\,l+1}, z_{N_2\,l+2}, \ldots, z_N\} \in \mathbb{R}^{p_1 \times p_1 \times B}$, and $\widetilde{Z}_{\mathcal{U}} = \{\widetilde{z}_{N_2\,l+1}, \widetilde{z}_{N_2\,l+2}, \ldots, \widetilde{z}_N\} \in \mathbb{R}^{p_2 \times p_2 \times B}$. After that, $(Z_{\mathcal{L}}, Y_{\mathcal{L}})$ and $(\widetilde{Z}_{\mathcal{L}}, Y_{\mathcal{L}})$ are then fed into the top and bottom branches, respectively, to train the model. The input of each branch has the same channel number, but different in the spatial size.

D. Spectral Feature Refinement

The high spectral resolution is the most prominent characteristic of HSI, which provides hundreds of spectral bands. However, the highly correlated spectral signatures result in high intraclass variation and low interclass difference. Therefore, direct explorations of the original spectral signatures may yield a poor class separability caused by the spectral redundancy. Inspired by the dense connections, the spectral dense block serves to handle the complex spectral property preliminarily. The role of the spectral feature refinement module is similar to that of principal component analysis. However, the spectral feature refinement module benefits the whole framework from joint training in an end-to-end manner with other modules.

Siamese strategy is adopted in the spectral feature refinement module, i.e., the top and bottom branches have the same configuration with the same parameters and weight. Siamese focuses on learning discriminative embeddings that aggregate the same classes. Parameter updating is mirrored across both subnetworks, promoting its feasibility under small-samples conditions.

To protect the original spatial relationship from being tampered with by convolution operations, the convolutional kernels of $1 * 1 * d$ (height * width * channel) are used throughout the entire spectral feature refinement module, without spatial information aggregations. d is a predefined convolutional parameter in the channel dimension. Considering the large number of input cubes, a $1 * 1 * d, k_0$ convolutional layer (CON1) with the down sampling stride (1,1,2) is first applied to reduce the number of bands for both branches, where k_0 is the kernel number of the 3-D convolution. Consequently, feature maps in the shape of $(p_1 * p_1 * c, k_0)$ and $(p_2 * p_2 * c, k_0)$ are obtained, respectively, where $c = (B - d + 1)/2$.

Then, the obtained feature maps are fed into m spectral dense blocks sequentially, in which the kernel size, padding and stride of convolution (CON2) are $(1 * 1 * d, k_1)$, "same" and $(1, 1, 1)$. The spectral dense block layer is designed to ensure maximum information transmit during automatic feature learning, including several convolution operations with direct connections. The direct connections allow the previous features to be passed to the all subsequent layers, which is regarded as a kind of feature reuse. As a result, the output feature maps of the mth layer for the top branch can be represented as

$$f_m = F_m([f_0, f_1, \ldots, f_{m-1}]) \tag{1}$$

where $[f_0, f_1, \ldots, f_{m-1}]$ denotes the concatenation of feature maps from layers $0, \ldots, m-1$ in the channel dimension. F_m is a module containing operations, such as convolution, activation and batch normalization (BN). Similarly, the output feature maps of the mth layer for the bottom branch can be found as

$$\widetilde{f}_m = \widetilde{F}_m([\widetilde{f}_0, \widetilde{f}_1, \ldots, \widetilde{f}_{m-1}]). \tag{2}$$

Therefore, two feature maps of $(p_1 * p_1 * c, k_0 + mk_1)$ and $(p_2 * p_2 * c, k_0 + mk_1)$ are generated.

At last, another convolution (CON3) of $(1 * 1 * c, C)$ are adopted to get the refined spectral representations $(p_1 * p_1 * 1, C)$ and $(p_2 * p_2 * 1, C)$, where C is the kernel number of CON3. During the whole of this session, the convolutions are conducted only on the spectral dimension of HSI cubes to keep the original spatial information without neighbor aggregations. Fig. 3 shows the detailed architecture of the spectral feature refinement module.

E. Spectral Feature Enhancement

The representation of input HSI cubes is refined through the spectral feature refinement module. However, the obtained C channels make different contributions to HSIC. Inspired by the AM, correlations among the C channels are captured to adjust bands weights adaptively. Informative channels are highlighted to enhance the spectral feature representation. Various AMs can be used in this session, such as self-attention, context attention. To illustrate the effectiveness of channel consistency, the classical self-attention is applied in SSACC. The procedure is presented in Fig. 4.

By squeezing the output of the spectral feature refinement module, the inputs of this module for the top branch are represented as $A \in \mathbb{R}^{p_1 \times p_1 \times C}$. Without additional parameters, the query, key, and value matrices are obtained by reshaping the input A, denoted as Q, K, and V. Specifically, the shape of Q, K and V are $\mathbb{R}^{C \times n}$, $\mathbb{R}^{n \times C}$, and $\mathbb{R}^{C \times n}$, respectively, where $n = p_1 \times p_1$ is the spatial area of input cubes. Then, a matrix multiplication operation is conducted on Q and K, following a softmax layer to obtain the normalized attention map $D \in \mathbb{R}^{C \times C}$

$$D_{ji} = \frac{\exp(Q_i * K_j)}{\sum_{i=1}^{n} \exp(Q_i * K_j)} \tag{3}$$

where D describes the similarity between query and key and D_j indicates the correlation of other channels with the jth channel of HSI cubes. After that, a matrix multiplication operation is executed on V and D with a reshape operation $r(.)$ to generate the attention feature. Finally, a summation operation is performed on the attention feature and the input A with the skip connection. As a result, the final output of spectral feature enhancement for the top branch $E \in \mathbb{R}^{p_1 \times p_1 \times C}$ could be obtained

$$E = r(DV) + A \tag{4}$$

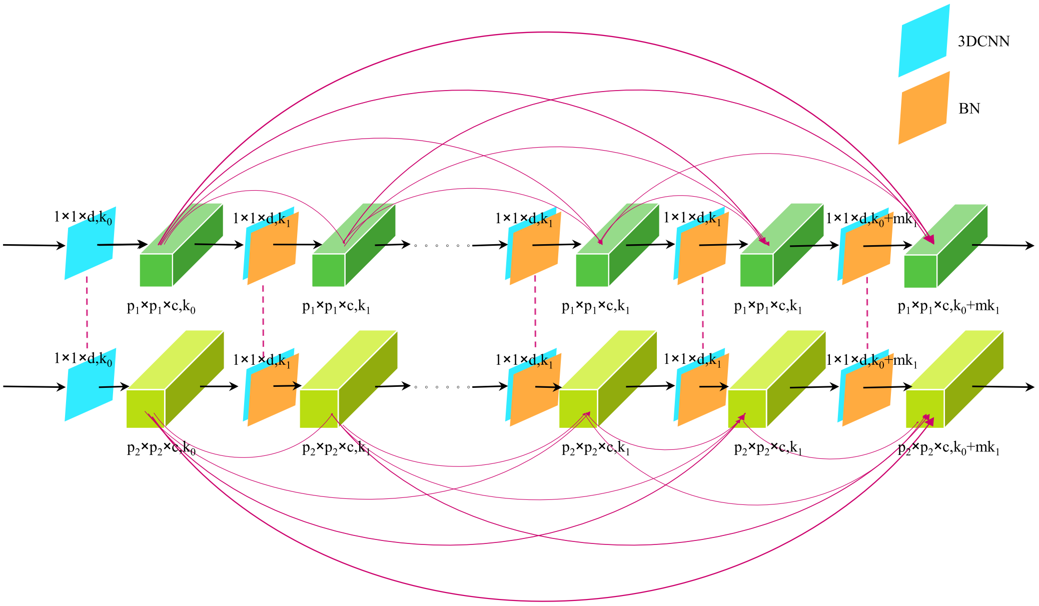

Fig. 3. Detailed architecture of the spectral feature refinement module.

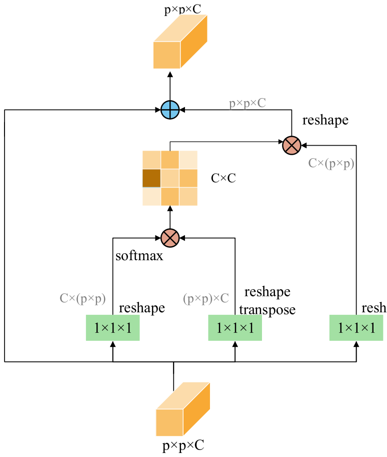

Fig. 4. Architecture of self-attention in the spectral feature enhancement module.

where reshape is utilized to convert attention feature $\mathbb{R}^{C\times N}$ to $\mathbb{R}^{p_1\times p_1\times C}$. The channel AM aims to enhance the extracted spectral features to benefit the feature representations.

Likewise, the attention map $\widetilde{D}\in\mathbb{R}^{C\times C}$ and enhanced feature $\widetilde{E}\in\mathbb{R}^{p_2\times p_2\times C}$ for the bottom branch can also be obtained through the spectral feature enhancement module.

In the process, spatial information is used in the spectral AM to learn the enhanced spectral representation. The neighborhood pixels could improve the generalization ability of learned spectral feature. Meanwhile, the influence of harmful neighborhood pixels is alleviated. Compared to the traditional spatial branch, there is no need to perform a subtle identification of every pixel located in the 3-D input cube.

F. Channel Consistency Regularization

In the classification stage, the enhanced features E and $\widetilde{E}$ are fed into batch normalization and nonlinear layer in order to stress the nonlinear internal structures hidden in the data. Then, the features are squeezed through the global average pooling layer to generate a channel descriptor for each channel. The squeezed descriptors $S\in\mathbb{R}^{C}$ and $\widetilde{S}\in\mathbb{R}^{C}$, which are served as the final representation of the input cubes, are sent to the fully connection layer with a softmax activation function to determine the final categories, respectively. Cross entropy is employed as the loss function of SSACC for classification. The classification loss functions for top (L_t) and bottom (L_b) branches are given as follows:

$$L_t=-\frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{L}\mathbb{1}\{y_k=l\}\log\frac{e^{\theta_l^T S_k}}{\sum_{l=1}^{L}e^{\theta_i^T S_k}}\tag{5}$$

$$L_b=-\frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{L}\mathbb{1}\{y_k=l\}\log\frac{e^{\theta_l^T \widetilde{S}_k}}{\sum_{l=1}^{L}e^{\theta_i^T \widetilde{S}_k}}\tag{6}$$

where S_k and $\widetilde{S}_k$ represent the final extracted features of the original HSI cube x_k for top and bottom branches, respectively, y_k is the truth land-cover label of HSI cube x_k, K is the number of samples in a minibatch, L is the number of land-cover labels, $\mathbb{1}\{y_k=l\}$ is the indicator function

$$\mathbb{1}\{y_k=l\}=\begin{cases}1 & y_k \text{ is the same as } l\\ 0 & \text{otherwise}\end{cases}.\tag{7}$$

3-D cube-based HSIC methods hold the rule that the HSI cube can be identified as the label of centric pixel. Therefore, different views (path size) of the same central pixel have the identical land-cover labels. Furthermore, different views within a certain scale should also show similar spectral patterns. To address the issue, a channel consistency regularization is introduced to extract the discriminative spectral features. It serves to compare the difference between the attention maps D and $\widetilde{D}$ produced in the spectral feature enhancement module. The channel consistency regularization is formulated as

$$L_c = \frac{1}{K}\sum_{k=1}^{K} \| D_k - \widetilde{D}_k \| \tag{8}$$

where $\| \, . \, \|$ calculates the distance of D_k and $\widetilde{D_k}$. The channel consistency regularization enables the network to learn discriminative spectral features by forcing the top and bottom branches to receive homologous channel correlations explicitly.

The final loss function pays attention to the loss of the correct category, and considers the loss of channel structure consistency. Therefore, the final loss function is defined as follows:

$$L = L_b + L_t + \lambda L_c \tag{9}$$

where λ is a hyperparameter to balance the classification loss and channel consistency regularization. The parameters in the SSACC are learned by back propagation and stochastic gradient descent.

For testing, the final land-cover labels are determined by averaging the results of the two branches. However, the channel consistency regularization is not carried out in the testing phase.

IV. Experiments and Analysis

A. Datasets Description

Indian Pine (IP): IP was captured by the AVIRIS sensor over the IPs test site in North-Western Indiana. It consisted of 145 × 145 pixels with a spatial resolution of 17 m/pixel. There were 224 spectral reflectance bands in IP, where the wavelength ranges from 400 to 2500 nm. By removing the bands covering the region of water absorption, the final number of bands was reduced to 200. The ground truth covered 16 classes of interest, which were mostly various crops in different growth phases. The numbers of training, validation, and testing samples for each class are detailed in Table I.

Pavia University (PU): PU data were gathered by the ROSIS sensor during a flight campaign over Pavia, northern Italy. It consisted of 610 × 340 pixels with a spatial resolution of 1.3 m/pixel. There were 115 spectral reflectance bands in PU with a spectral resolution of 4 nm, where the wavelength ranges from 430 to 860 nm. By removing 12 noisy channels affected by the absorption of water vapor, the final 103 bands were used in the experiments. The samples contained 9 classes of interest, which were mostly various crops in different growth phases. The numbers of training, validation and testing samples for each class are detailed in Table II.

Salinas Valley (SV): The SV dataset was acquired over Salinas Valley, California by the Airborne Visible/Infrared Imaging Spectrometer sensor. It consisted of 512 × 217 samples with a spatial resolution of 3.7 m/pixel. There were 224 spectral reflectance bands in SV ranging from 400 to 2500 nm. After removing 20 water absorption bands, the final 204 bands were retained in the experiments. The available land-cover category covered 16 classes of interest. Table III shows the detail information on the numbers of training, validation, and testing samples for each class of interests.

TABLE I
Number of Training, Validation, and Testing Samples for the IP Dataset

Order	Class	Total Number	Train	Val	Test
1	Alfalfa	46	3	3	40
2	Corn-notill	1428	14	14	1400
3	Corn-mintill	830	8	8	814
4	Corn	237	3	3	231
5	Grass-pasture	483	4	4	475
6	Grass-trees	730	7	7	716
7	Grass-pasture-mowed	28	3	3	22
8	Hay-windrowed	478	4	4	470
9	Oats	20	3	3	14
10	Soybean-notill	972	9	9	954
11	Soybean-mintill	2455	24	24	2407
12	Soybean-clean	593	5	5	583
13	Wheat	205	3	3	199
14	Woods	1265	12	12	1241
15	Buildings-Grass-Trees-Drives	386	3	3	380
16	Stone-Steel-Towers	93	3	3	87
	Total	10249	108	108	10033

TABLE II
Number of Training, Validation, and Testing Samples for the PU Dataset

Order	Class	Total Number	Train	Val	Test
1	Asphalt	6631	33	33	6565
2	Meadows	18649	93	93	18463
3	Gravel	2099	10	10	2079
4	Trees	3064	15	15	3034
5	Painted metal sheets	1345	6	6	1333
6	Bare Soil	5029	25	25	4979
7	Bitumen	1330	6	6	1318
8	Self-Blocking Bricks	3682	18	18	3646
9	Shadows	947	4	4	939
	Total	42776	210	210	42356

B. Experiments Setup

In the experiments, overall accuracy (OA), average accuracy (AA), and Kappa coefficient (κ) was used to evaluate the proposed method quantitatively. OA indicates the ratio of the number of correctly predicted pixels to the total number of pixels. AA refers to the mean of accuracies in different categories. κ measures the consistency between the ground truth and classification results. The higher of the three metric values, the better the classification results are.

All experiments were carried out on a system with NVIDIA GeForce RTX-2070 GPU and 16 GB main memory. The Adam optimizer with the learning rate of 0.0005 was used for model training. The convolutional kernels used in the spectral feature refinement module were all $1 \times 1 \times 7$. The balance parameter λ was set to 0.1. The channel of refined spectral feature C and the

TABLE III
NUMBER OF TRAINING, VALIDATION, AND TESTING SAMPLES FOR THE SV DATASET

Order	Class	Total Number	Train	Val	Test
1	Brocoli-green-weeds-1	2009	10	10	1989
2	Brocoli-green-weeds-2	3726	18	18	3690
3	Fallow	1976	9	9	1958
4	Fallow-rough-plow	1394	6	6	1382
5	Fallow-smooth	2678	13	13	2652
6	Stubble	3959	19	19	3921
7	Celery	3579	17	17	3545
8	Grapes-untrained	11271	56	56	11159
9	Soil-vinyard-develop	6203	31	31	6141
10	Corn-senesced-green-weeds	3278	16	16	3246
11	Lettuce-romaine-4wk	1068	5	5	1058
12	Lettuce-romaine-5wk	1927	9	94	1824
13	Lettuce-romaine-6wk	916	4	4	908
14	Lettuce-romaine-7wk	1070	5	5	1060
15	Vinyard-untrained	7268	36	36	7196
16	Vinyard-vertical-trellis	1807	9	9	1789
	Total	54129	263	348	53603

length of squeezed channel descriptor were both set to 60. The optimal patch sizes combination were $p_1 = 7$, $p_2 = 11$ for IP, $p_1 = 5$, $p_2 = 13$ for PU and $p_1 = 9$, $p_2 = 11$ for SV. Table IV provides the implementation details for IP. The implementation details for PU and SV are displayed in the same way.

C. Comparing With Other Methods

To demonstrate the effectiveness of the proposed method, we compared the proposed SSACC method with several widely used methods such as SVM, SSRN [25], FDSSC [58], DBMA [42], MAFN [62] and the state-of-the-art double-branch dual-AM network DBDA [43] methods. All parameters of each classifier were set according to the original papers.

As shown in Tables I–III, 1%, 0.5%, and 0.5% of the pixels were randomly chosen as training samples for IP, PU, and SV, respectively. Tables V–VII reports the average values and standard deviations for the metrics: OA, AA, and κ over 10 runs on the three datasets. Generally, the proposed SSACC achieves the best results. In all but one case, it shows great advantages over other methods for all the metrics, which verify the effectiveness of the proposed SSACC.

Specifically, SVM performed the worst among all the methods, which confirmed that the deep models have their advantages on HSIC than the conventional methods. Among the deep models, the classification accuracy of SSRN was lower than other methods on the three datasets. This was due to the fact that the SSRN adopts Resnet as backbone, while other methods employed DenseNet. It demonstrated the superiority of dense connection structure for HSIC, where multilevel features were reused to improve the generalization ability of network. Furthermore, the results of FDSSC were lower than most of attention-based methods, such as DBDA, MAFN, and SSACC. The main reason was that the AM can suppress interfering information for feature learning, which was useful and beneficial for HSIC. However, DBMA performs worse than FDSSC on the IP dataset, where attention was also used for HSIC. The results demonstrated that the AM could not fully resolve the problem of discriminative HSI feature extraction, especially when the number of training samples was insufficient. Moreover, DBMA, DBDA, and SSACC are all double-branch methods. DBMA and DBDA employed the two branches to extract spectral and spatial features separately, aiming to utilize the complementarity of spectral and spatial features to improve the performance of HSIC. Nevertheless, the purpose of the proposed SSACC is to find effective spectral structure hidden in the HSI. It is believed that HSI provides a wealth of spectral information, which is the most important information for HSIC. As shown in Tables V–VII, SSACC achieves the best performance by only spectral features, demonstrating the effectiveness and discriminability of the extracted spectral features. SSACC built the explicit and implicit interactions between the two branches, which was the major competitive advantage on obtaining discriminative spectral features. The implicit interaction derived from the siamese structure, while the explicit one stems from the channel consistency regularization. The channel consistency assumption could also alleviate the influence of interfering pixels with similar effects as that of spatial attention. Finally, SSACC was noted to have the minimum standard deviations, which was an important characteristic related to the generalization capability of the HSIC methods. Variance reduction of the network guarantees a reduction in generalization error [60]. The special effect could also be attributed to the effects of interactions between the top and bottom branches.

Figs. 5–7 visualize the classification maps of different methods on the three datasets. The visual classification maps are consistent with the results listed in Tables V–VII. It can also be found that SSACC obtains smooth classification maps, which alleviates the influences of spectral variability effectively.

D. Comparisons With a Varied Number of Training Samples

In this part, the generalizability of the compared methods toward different numbers of labeled training samples was investigated. 0.5%, 1%, 3%, and 5% labeled samples were randomly selected as training data for the IP dataset while. 0.25%, 0.5%, 1%, and 2% labeled samples for the PU and SV datasets. Fig. 8 shows the overall accuracies of different methods on the three datasets with varied training samples proportions. Generally, more training samples could improve the performance of all the methods. It could be found that SSACC produces the highest OA values in all cases. The merit was superior particularly when the percentage of labeled samples for training was small. It demonstrated that SSACC captures more discriminative

TABLE IV
IMPLEMENTATION DETAILS OF THE MODULE

Module	Input size	Layer operation	Output size
spectral refinement	Top:7×7×200 Bottom:11×11×200	Conv3D:1×1×7,24	7×7×97,24 11×11×97,24
	Top:7×7×97,24 Bottom:11×11×97,24	BN-Conv3D:1×1×7,12	7×7×97,12 11×11×97,12
	-	Concatenate	7×7×97,36 11×11×97,36
	Top:7×7×97,36 Bottom:11×11×97,36	Conv3D:1×1×7,12	7×7×97,12 11×11×97,12
	-	Concatenate	7×7×97,48 11×11×97,48
	Top:7×7×97,48 Bottom:11×11×97,48	Conv3D:1×1×7,12	7×7×97,12 11×11×97,12
	-	Concatenate	7×7×97,60 11×11×97,60
spectral enhancement	Top:7×7×1,60 Bottom:11×11×1,60	Attention	7×7×1,60 D_k: 60×60 11×11×1,60 $\widetilde{D_k}$: 60×60
squeezed channel	Top:7×7×1,60 Bottom:11×11×1,60	BN-Dropout-GAP	1×60 1×60

TABLE V
CATEGORIZED RESULTS FOR THE IP DATASET WITH 1% TRAINING SAMPLES

Order	Class	SVM	SSRN	FDSSC	DBMA	DBDA	MAFN	Proposed
1	Alfalfa	06.45	47.05	90.21	58.61	88.71	75.00	**93.28**
2	Corn-notill	45.56	63.64	73.47	72.63	77.73	76.40	**86.69**
3	Corn-mintill	38.98	70.30	83.98	86.31	**87.93**	76.05	86.72
4	Corn	20.17	**83.03**	52.75	65.41	73.42	66.02	77.55
5	Grass-pasture	57.91	98.54	99.83	95.49	**100.0**	91.61	**100.0**
6	Grass-trees	76.42	95.02	95.93	97.08	**98.01**	93.61	96.69
7	Grass-pasture-mowed	16.67	86.15	49.26	33.93	**93.23**	46.34	92.25
8	Hay-windrowed	84.84	96.55	99.47	98.71	98.31	90.00	**99.83**
9	Oats	27.27	49.51	49.22	11.86	55.65	**64.29**	48.01
10	Soybean-notill	51.67	70.18	78.38	63.83	78.42	53.04	**79.04**
11	Soybean-mintill	55.58	81.39	93.98	87.21	91.61	78.39	**95.96**
12	Soybean-clean	24.29	68.13	90.54	59.79	87.69	67.49	**95.42**
13	Wheat	78.54	95.25	91.51	96.08	94.54	**100.0**	92.47
14	Woods	74.84	96.34	96.19	**98.82**	98.48	98.33	98.05
15	Buildings-Grass-Trees-Drives	32.75	89.87	**95.24**	87.58	93.68	66.10	90.58
16	Stone-Steel-Towers	**98.59**	89.77	83.13	79.33	87.69	78.43	83.87
OA		55.96±1.11	80.14±1.41	86.74±2.11	81.06±0.49	89.34±0.81	78.12±0.35	**91.10±0.05**
AA		49.41±1.11	80.05±1.56	84.49±1.81	73.99±0.96	**84.94±0.88**	77.88±0.43	84.59±0.18
Kappa		49.17±5.55	77.28±1.64	84.94±2.34	78.41±0.58	87.87±0.91	74.86±0.86	**89.89±0.06**

The bold numbers indicate the best performance.

TABLE VI
CATEGORIZED RESULTS FOR THE PU DATASET WITH 0.5% TRAINING SAMPLES

Order	Class	SVM	SSRN	FDSSC	DBMA	DBDA	MAFN	Proposed
1	Asphalt	81.98	98.45	99.05	94.91	93.41	**97.84**	97.59
2	Meadows	90.92	97.90	97.63	98.23	**98.44**	98.19	97.68
3	Gravel	52.64	77.84	83.64	85.71	90.55	83.91	**95.63**
4	Trees	94.00	99.44	**99.92**	97.93	97.81	99.31	99.61
5	Painted metal sheets	92.69	99.44	96.96	93.80	99.38	**99.62**	95.29
6	Bare Soil	82.78	98.24	99.82	98.81	97.73	98.40	**100.0**
7	Bitumen	57.20	86.72	98.75	98.40	97.80	97.49	**99.02**
8	Self-Blocking Bricks	80.93	82.56	75.70	83.41	86.22	91.71	**94.44**
9	Shadows	**99.78**	97.36	92.88	94.08	98.00	99.36	95.30
OA		84.86±1.11	95.26±0.64	95.07±0.56	95.62±0.91	95.80±1.08	96.98±0.35	**97.57±0.20**
AA		81.44±0.22	93.11±1.22	93.93±0.52	94.18±1.15	95.49±0.83	96.21±0.39	**97.17±0.28**
Kappa		79.75±0.23	93.70±0.85	93.44±0.75	94.17±1.21	94.41±1.45	95.99±0.42	**96.77±0.27**

The bold numbers indicate the best performance.

TABLE VII
CATEGORIZED RESULTS FOR THE SV DATASET WITH 0.5% TRAINING SAMPLES

Order	Class	SVM	SSRN	FDSSC	DBMA	DBDA	MAFN	Proposed
1	Brocoli-green-weeds-1	99.69	93.26	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
2	Brocoli-green-weeds-2	99.06	99.29	98.43	99.76	99.98	99.52	**99.99**
3	Fallow	88.22	94.58	99.67	96.63	99.45	99.84	**100.0**
4	Fallow-rough-plow	97.68	92.94	94.94	91.88	94.06	94.85	**95.70**
5	Fallow-smooth	97.86	98.65	98.00	97.62	97.13	98.75	**99.97**
6	Stubble	**100.0**	99.96	99.99	99.97	**100.0**	99.48	**100.0**
7	Celery	99.29	99.96	99.97	98.95	99.90	**100.0**	**100.0**
8	Grapes-untrained	69.63	86.62	94.41	83.67	93.53	96.54	**95.90**
9	Soil-vinyard-develop	96.79	99.14	**99.58**	98.91	99.47	99.42	99.52
10	Corn-senesced-green-weeds	83.00	93.04	97.87	95.10	92.68	98.05	**99.32**
11	Lettuce-romaine-4wk	92.14	96.67	99.43	88.54	98.98	99.34	**99.68**
12	Lettuce-romaine-5wk	78.46	93.82	99.21	96.78	98.43	99.57	**99.83**
13	Lettuce-romaine-6wk	91.45	85.48	99.40	98.13	99.69	99.13	**99.90**
14	Lettuce-romaine-7wk	93.77	93.79	95.92	**97.98**	96.49	95.49	86.79
15	Vinyard-untrained	62.73	81.70	85.18	80.47	78.41	87.21	**94.20**
16	Vinyard-vertical-trellis	98.66	99.52	99.16	97.27	99.97	99.95	**100.0**
OA		85.16±0.51	91.58±3.02	95.52±1.79	91.77±2.45	94.13±2.88	96.42±0.49	**97.83±0.13**
AA		90.53±0.16	94.28±2.48	97.57±0.98	95.10±0.93	96.76±1.44	97.09±0.56	**98.17±0.16**
Kappa		83.41±0.36	90.62±3.36	95.01±1.96	90.83±2.72	93.46±3.24	96.47±0.43	**97.58±0.14**

The bold numbers indicate the best performance.

features than other methods. Meanwhile, SSACC improved the generalization ability of network due to the implicit and explicit interactions between the double branches.

E. Comparisons on the Parameters

Fig. 9 reports the model complexity and OA of different methods in terms of the number of trainable weight parameters updated during backpropagation. It could be seen that the proposed SSACC achieves the best classification accuracy with the least number of trainable parameters for all three datasets. Noteworthily, FDSSC, DBMA, DBDA, and SSACC were all DenseNet based methods. FDSSC obtained the spectral–spatial by performing spectral and spatial networks sequentially. In contrast, DBMA and DBDA employed spectral and spatial networks parallelly, which reduced the number of parameters. Moreover, the proposed SSACC had a similar architecture as DBDA and DBMA. The difference was that SSACC performs two parallel spectral networks where only $1 \times 1 \times d$ convolution kernels were used, without spatial convolutional operations. In addition, channel consistency regularization would not introduce any additional trainable parameters. Therefore, SSACC achieved the best classification accuracy with the fewest parameters among the compared methods.

F. Influence of ϵ

The patch size of cropped HSI cubes has a great influence on HSIC results. For SSACC, it was essential to set patch sizes for the top and bottom branches, respectively. To evaluate the influence of the patch size combination, the performance of

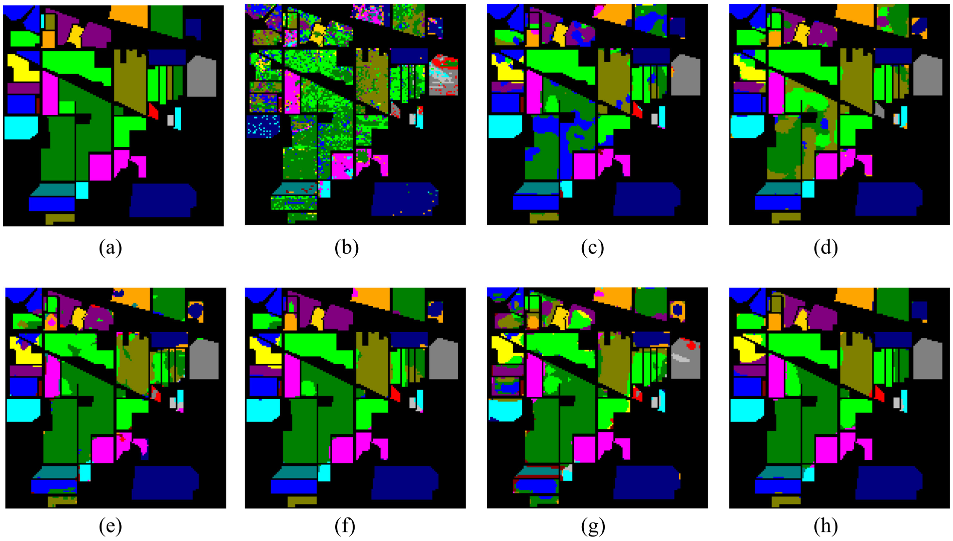

Fig. 5. Classification maps for the IP dataset. (a) Ground-truth.(b)–(h) Classification map of SVM, SSRN, FDSSC, DBMA, DBDA, MAFN, and SSACC.

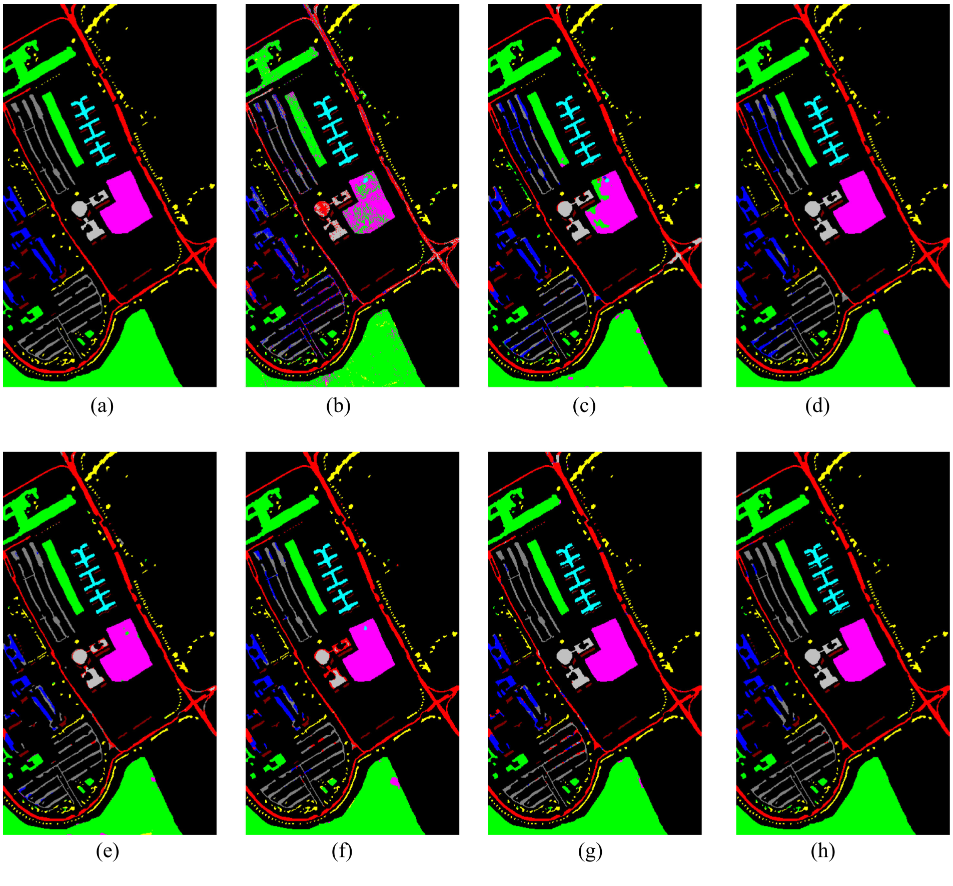

Fig. 6. Classification maps for the PU dataset. (a) Ground-truth. (b)–(h) Classification map of SVM, SSRN, FDSSC, DBMA, DBDA, MAFN, and SSACC.

SSACC was investigated by fixing the patch size of the bottom branch (p_2) and vary the patch size of the top branch (p_1) to investigate the performance of SSACC. The fixed p_2 is set to 11, 13, and 11 for IP, PU, and SV, which was the optimal value through experimental verification for a single branch. p_1 was chosen from $\{1, 3, 5, 7, 9, 11\}$. As shown in Fig. 10, a proper spatial size deviation $\epsilon = p_2 - p_1$ can improve the OA of classification. A small ϵ did not take spatial differentiation into account, which was not conducive to obtaining discriminative features. In contrast, a large ϵ might bring unanticipated interfering pixels beyond the scope of the channel consistency hypothesis.

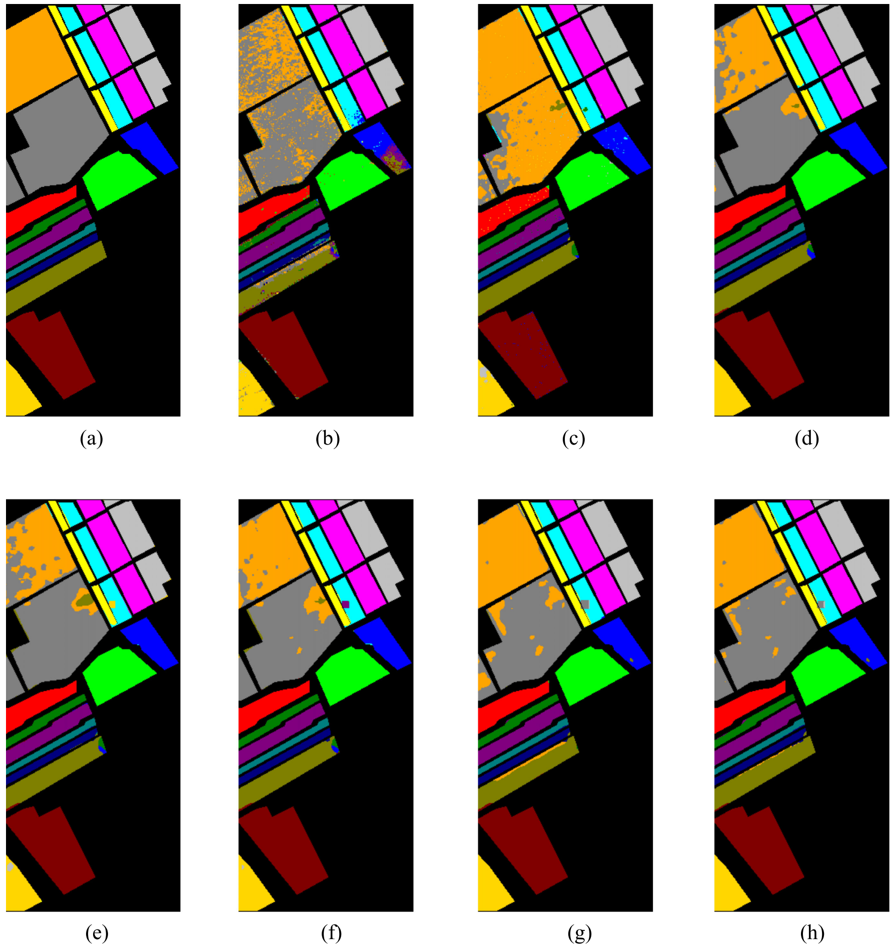

Fig. 7. Classification maps for the SV dataset. (a) Ground-truth. (b)–(h) Classification map of SVM, SSRN, FDSSC, DBMA, DBDA, MAFN, and SSACC.

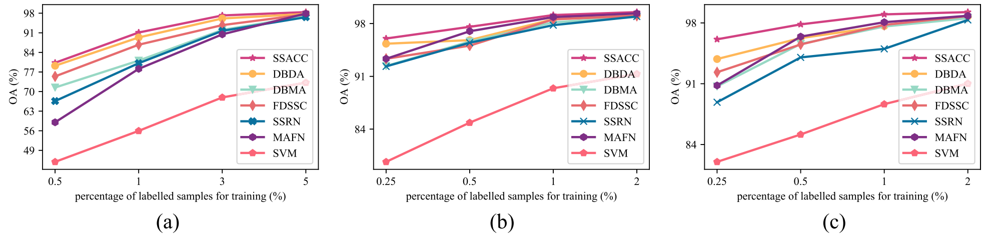

Fig. 8. OA (%) with varied training samples proportions for different methods in the three datasets.

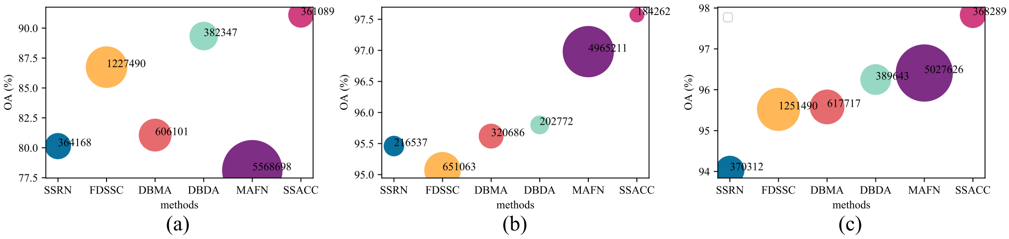

Fig. 9. Comparisons on the number of parameters and OA (%) for different methods in the three datasets.

TABLE VIII
Classification Accuracies of SSACC With Varied λ in the Three Datasets

λ	IP			PU			SV		
	OA	AA	Kappa	OA	AA	Kappa	OA	AA	Kappa
0	90.63±0.46	84.03±0.34	89.60±0.30	97.26±0.28	96.68±0.48	96.42±0.38	97.55±0.35	97.89±0.31	97.01±0.26
0.01	91.01±0.11	84.39±0.30	89.79±0.12	97.27±0.35	96.60±0.64	96.42±0.47	97.60±0.15	98.01±0.19	97.09±0.10
0.1	91.10±0.05	84.59±0.18	89.89±0.06	97.57±0.20	97.17±0.28	96.77±0.27	97.83±0.13	98.17±0.16	97.58±0.14
1	91.07±0.11	84.89±0.31	89.86±0.12	97.37±0.15	96.59±0.38	96.56±0.19	97.67±0.11	98.10±0.21	97.26±0.06
10	90.45±0.37	84.29±0.89	89.14±0.43	95.07±0.79	90.34±0.87	92.94±0.63	90.35±0.21	91.02±0.51	90.01±0.33
100	69.32±3.07	61.49±8.79	63.88±3.83	88.33±2.68	80.45±4.47	86.76±3.01	79.61±1.01	80.21±2.59	78.96±2.03

TABLE IX
Ablation Study in Terms of OA(%) for the Proposed SSACC

	siamese	consistency	IP	PU	SV
SSACC	√	√	91.10±0.05	97.57±0.20	97.83±0.13
SSACC-P		√	90.93±0.26	97.44±0.42	97.65±0.39
SSACC-O	√		90.63±0.46	97.26±0.28	97.55±0.35
SSACC-PO			88.37±0.89	95.03±1.21	94.11±1.52

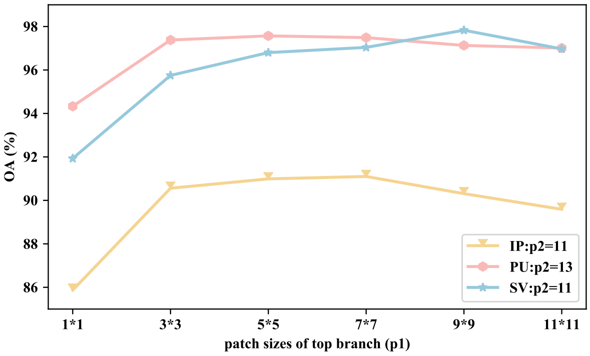

Fig. 10. OA (%) of SSACC with different input patch sizes in the three datasets.

G. Influence of λ

Parameter λ was employed to balance the classification loss and channel consistency regularization in SSACC. To verify the contribution of channel consistency regularization, experimental results were reported with varied λ on the three datasets in Table VIII. Parameter λ was chosen from $\{0, 0.01, 0.1, 1, 10, 100\}$. It was observed that SSACC achieves the best performance when λ was around 0.1 for all the three datasets, indicating that the channel consistency regularization played an important role in capturing discriminative spectral structure. Meanwhile, λ was insensitive when the value varies from 0.01 to 1. It demonstrated that channel consistency regularization was an effective auxiliary term for HSIC. When $\lambda \geq 10$, the classification accuracies dropped rapidly. If we overemphasize on spectral structure consistency, the network could pay more attention to assign the same weights to spectral structure, regardless of what the semantic label should be. Therefore, an overemphasis of spectral structure consistency would cause the performance degradation.

H. Ablation Study

In this part, experiments were conducted to verify the effectiveness of siamese structure and channel consistency regularization. The siamese structure was employed in the spectral refinement module and channel consistency regularization was conducted on the attention maps. Both of the two measures were exploited to obtain discriminative spectral features and high generalization HSIC models. To shed light on the contributions of the two components, Table IX reports the classification results of SSACC without siamese structure or channel consistency on the three datasets. SSACC-P represents that pseudosiamese structures are adopted to replace the siamese structure in the spectral refinement module, where the parameters are not shared between the double branches. SSACC-O denotes the reduced SSACC by removing channel consistency regularization. SSACC-PO represents both the components are replaced. It can be observed that lacking any one of the components will inevitably hurt the OA and standard deviation. Therefore, the siamese structure and channel consistency regularization turned out to be contributive for discriminative spectral feature extraction. The two components work collaboratively to render satisfactory classification performance for HSIC.

I. Consistency Visualization

To validate that the spectral consistency are captured by the proposed SSACC, we visualize the difference values of attention maps of top branch D_1 and bottom branch D_2. Three kinds of input combinations for top and bottom are chosen for visualization. The visualization results are shown in Fig. 11. Comparing the difference values of attention maps, the proposed SSACC obtains cleaner maps by performing siamese strategy and channel attention consistency, which validates SSACC can capture the spectral consistency.

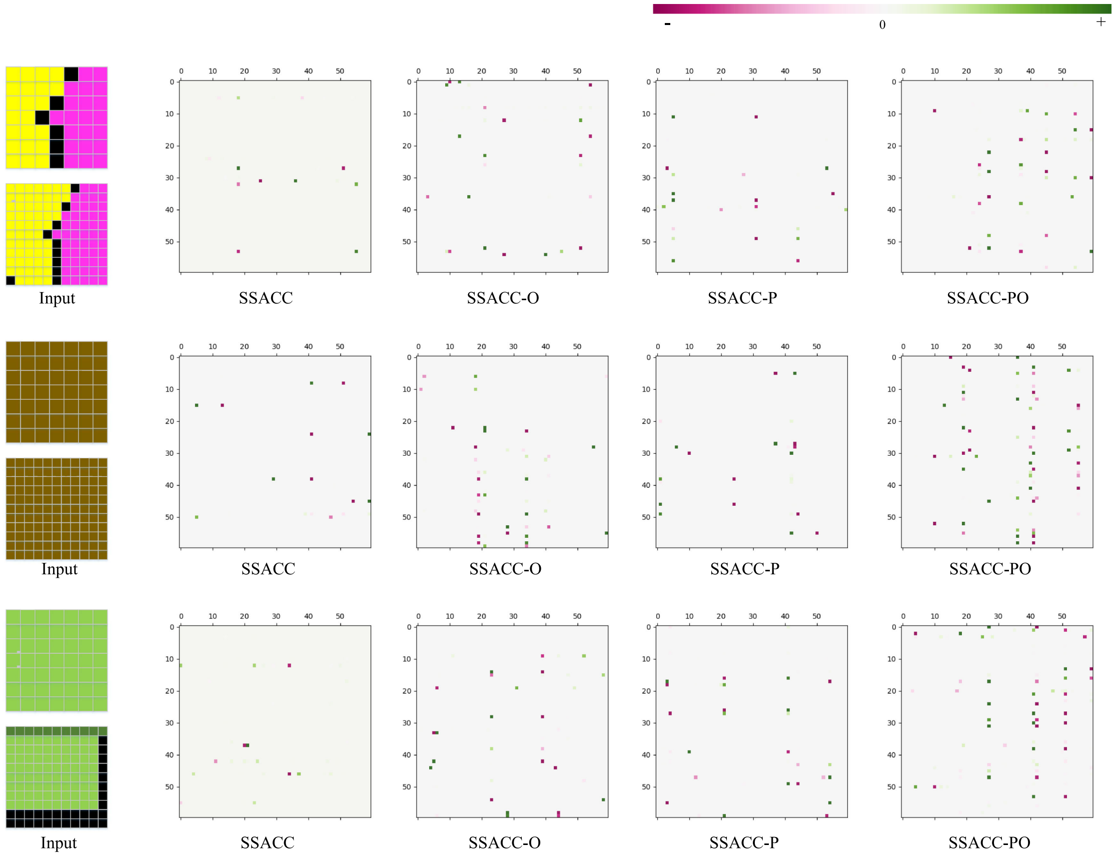

Fig. 11. Spectral consistency visualization.

V. Conclusion

In this article, an SSACC was developed for discriminative spectral feature learning and HSI classification. The proposed SSACC consists of two branches with implicit and explicit interactions. The siamese structure is the implicit interaction, which reduces the demands for the number of training samples to a certain extent. The channel consistency regularization is the explicit interaction, which is a key term to capture discriminative spectral features. These measures improve the classification accuracy and generalization ability of the proposed SSACC. The experimental results on three public HSI datasets indicate that the proposed SSACC can yield better performance than the state-of-the-art HSI classification methods with fewer number of parameters.

References

[1] L. Liang, L. Di, L. Zhang, M. Deng, and H. Lin, "Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method," *Remote Sens. Environ.*, vol. 165, pp. 123–134, 2015.

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

[3] B. Arabi, M. S. Salama, J. Pitarch, and W. Verhoef, "Integration of in-situ and multi-sensor satellite observations for long-term water quality monitoring in coastal areas," *Remote Sens. Environ.*, vol. 239, 2020, Art no. 111632.

[4] P. Mohanty, S. Panditrao, R. Mahendra, H. S. Kumar, and T. S. Kumar, "Identification of coral reef feature using hyperspectral remote sensing," *Proc. SPIE - Int. Soc. Opt. Eng.*, vol. 9880, 2016, Art no. 98801B.

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

[6] 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.

[7] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[8] Y. Zhang, W. Li, R. Tao, J. Peng, Q. Du, and Z. Cai, "Cross-scene hyperspectral image classification with discriminative cooperative alignment," *IEEE Trans. Geosci. Remote Sens.*, to be published. doi: 10.1109/TGRS.2020.3046756.

[9] 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.

[10] 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.

[11] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, Mar. 2009.

[12] 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.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 1097–1105, 2012.

[14] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[16] A. R. Sharma and P. Kaushik, "Literature survey of statistical, deep and reinforcement learning in natural language processing," in *Proc. Int. Conf. Comput., Commun. Autom.*, 2017, pp. 350–354.

[17] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, 2015, Art. no. 258619.

[18] 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.

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

[20] Y. Zhan, D. Hu, Y. Wang, and X. Yu, "Semisupervised hyperspectral image classification based on generative adversarial networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 212–216, Feb. 2018.

[21] X. Wang, K. Tan, Q. Du, Y. Chen, and P. Du, "Caps-tripleGAN: Gan-assisted capsnet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7232–7245, Sep. 2019.

[22] 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.

[23] 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.

[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] 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.

[26] P. V. Arun, K. M. Buddhiraju, and A. Porwal, "Capsulenet-based spatial-spectral classifier for hyperspectral images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1849–1865, Jun. 2019.

[27] C. Zhang, G. Li, and S. Du, "Multi-scale dense networks for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9201–9222, Nov. 2019.

[28] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.

[29] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral-spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[30] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, 2017, Art no. 1330.

[31] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Hyperspectral image classification using spectral-spatial LSTMs," *Neurocomputing*, vol. 328, pp. 39–47, 2019.

[32] W.-S. Hu, H.-C. Li, L. Pan, W. Li, R. Tao, and Q. Du, "Spatial-spectral feature extraction via deep convLSTM neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4237–4250, Jun. 2020.

[33] C. Shi and C.-M. Pun, "Superpixel-based 3 D deep neural networks for hyperspectral image classification," *Pattern Recognit.*, vol. 74, pp. 600–616, 2018.

[34] S. Jia, X. Deng, M. Xu, J. Zhou, and X. Jia, "Superpixel-level weighted label propagation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 5077–5091, Jul. 2020.

[35] P. Sellars, A. I. Aviles-Rivero, and C.-B. Schönlieb, "Superpixel contracted graph-based learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4180–4193, Jun. 2020.

[36] 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.

[37] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-aided CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2281–2293, Mar. 2021.

[38] S. Pande and B. Banerjee, "Adaptive hybrid attention network for hyperspectral image classification," *Pattern Recognit. Lett.*, vol. 144, pp. 6–12, 2021.

[39] L. A. I. Qiuxia, S. Khan, Y. Nie, S. Hanqiu, J. Shen, and L. Shao, "Understanding more about human and machine attention in deep neural networks," *IEEE Trans. Multimedia*, vol. 23, no. 7, pp. 2086–2099, 2021.

[40] 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, Jan. 2020.

[41] B. Fang, Y. Li, H. Zhang, and J. Cheung-Wai Chan, "Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism," *Remote Sens.*, vol. 11, no. 2, pp. 159–177, 2019.

[42] 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, 2019, Art no. 1307.

[43] 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. 582–607, 2020.

[44] A. Vaswani *et al.*, "Attention is all you need," 2017, *arXiv:1706.03762*.

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

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

[47] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNET: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, pp. 1971–1980, 2019.

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

[49] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[50] J. Choe, S. Lee, and H. Shim, "Attention-based dropout layer for weakly supervised single object localization and semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published. doi: 10.1109/TPAMI.2020.2999099.

[51] D. Zoran, M. Chrzanowski, P.-S. Huang, S. Gowal, A. Mott, and P. Kohli, "Towards robust image classification using sequential attention models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9483–9492.

[52] E. Pan, Y. Ma, X. Mei, X. Dai, and J. Ma, "Spectral-spatial classification of hyperspectral image based on a joint attention network," in *Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 413–416.

[53] H. Gao, Y. Yang, D. Yao, and C. Li, "Hyperspectral image classification with pre-activation residual attention network," *IEEE Access*, vol. 7, pp. 176587–176599, 2019.

[54] 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.

[55] Z. Lu, B. Xu, L. Sun, T. Zhan, and S. Tang, "3D channel and spatial attention based multi-scale spatial spectral residual network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 7, pp. 4311–4324, 2020.

[56] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral-spatial kernel resnet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021.

[57] 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.

[58] W. Wenju, D. Shuguang, J. Zhongmin, and S. Liujie, "A fast dense spectral-spatial convolution network framework for hyperspectral images classification," *Remote Sens.*, vol. 10, no. 7, 2018, Art no. 1068.

[59] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, "Active transfer learning network: A unified deep joint spectral-spatial feature learning model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1741–1754, Mar. 2019.

[60] S. Shabbir and M. Ahmad, "Hyperspectral image classification-traditional to deep models: A survey for future prospects," 2021, *arXiv:2101.06116*.

[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[62] Z. Li *et al.*, "Hyperspectral image classification with multiattention fusion network," *IEEE Geosci. Remote Sens. Lett.*, to be published. doi: 10.1109/LGRS.2021.3052346.

Leiquan Wang (Member, IEEE) received the Ph.D. degree in communication and electrical systems from Beijing University of Posts and Telecommunications, Beijing, China, in 2016.

He is currently a Lecturer with College of Computer and Communication Engineering, China University of Petroleum (East China), Dongying, China. His current research interests include multimodal fusion, cross modal retrieval, and image/video caption.

Yao Lin is currently working toward the Postgraduate degree with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Dongying, China.

Her current research interests include hyperspectral image classification.

Jinyun Liu received the bachelor degree from the China University of Petroleum (East China), Dongying, China, in 2003.

He is currently an Engineer with the SINO-Pipeline International Company Limited, Beijing, China. His research interests include machine learning, remote sensing image processing, and object detection.

Zhongwei Li received the Ph.D. degree from the China University of Petroleum, Dongying, China, in 2011.

He is currently a Professor with the College of Oceanography and Space Informatics, China University of Petroleum. His current research interests include remote sensing image processing and ocean numerical forecasting and cloud computing.

Chunlei Wu (Member, IEEE) received the Ph.D. degree majoring in computer application technology from Ocean University of China, Qingdao, China, in 2014.

He is currently an Associate Professor with the College of Petroleum (East China), Dongying, China. He is currently with the University of Victoria, Victoria, BC, Canada, as a Visiting Scholar. He has authored and coauthored more than 30 journal and conference papers and textbooks. His current interests include image and video processing, and machine learning.