Attention-Based Second-Order Pooling Network for Hyperspectral Image Classification

Zhaohui Xue, *Member, IEEE*, Mengxue Zhang, Yifeng Liu, and Peijun Du, *Senior Member, IEEE*

***Abstract*— Deep learning (DL) has exhibited huge potentials for hyperspectral image (HSI) classification due to its powerful nonlinear modeling and end-to-end optimization characteristics. Although the superior performance of DL-based methods has been witnessed, some limitations can still be found. On the one hand, existing DL frameworks usually resorted to first-order statistical features, whereas they rarely considered second-order or higher order statistical features. On the other hand, the optimization of complex hyperparameters (e.g., the layer number and convolutional kernel size) is time-consuming and a very tough task, making the designed DL framework unexplainable. To overcome these challenges, we propose a novel attention-based second-order pooling network (A-SPN). First, a first-order feature operator is designed to model the spectral–spatial information of HSI. Second, an attention-based second-order pooling (A-SOP) operator is designed to model discriminative and representative features. Finally, a fully connected layer with softmax loss is used for classification. The proposed framework can obtain second-order statistical features in an end-to-end manner. In addition, A-SPN is free of complex hyperparameters tuning, making it more explainable and easily equipped for classification tasks. Experimental results based on three common hyperspectral data sets demonstrate that A-SPN outperforms other traditional and state-of-the-art DL-based HSI classification methods in terms of generalization performance with limited training samples, classification accuracy, convergence rate, and computational complexity.**

***Index Terms*— Attention mechanism, classification, deep learning (DL), hyperspectral image (HSI), second-order pooling.**

I. Introduction

HYPERSPECTRAL image (HSI) consists of hundreds of narrow continuous wavelength bands throughout the electromagnetic spectrum with a very high spectral resolution, which can provide abundant information to accurately discriminate similar materials of interest [1]. As a fundamental procedure of HSI analysis, HSI classification has been widely used in precision agriculture, forestry monitoring, mineral survey, environmental assessment, and so on.

Advanced machine learning and pattern recognition methods have greatly promoted the developments of HSI classification [2]–[5]. Deep learning (DL) can automatically learn data-adaptive high-level features in a hierarchical manner, which has been intensively used for HSI classification scenarios due to its revolutionary power in modeling spectral–spatial features [6]–[10]. Typical DL frameworks include stacked autoencoder (SAE) [11], convolutional neural network (CNN) [12], deep belief network (DBN) [13], recurrent neural network (RNN) [14], and long short-term memory (LSTM) [15].

Among those DL methods, CNN has become the leading architecture for most image recognition, classification, and detection tasks [16]. Extensive explorations of using CNN for HSI classification were made in recent years, which can be categorized as 1-D CNN, 2-D CNN, 3-D CNN, and some hybrids of them. 1-D CNN is used to process vectorized information, which can extract the spectral features along the radiometric dimension by using 1-D convolutional and pooling operations for HSI [12], [17]. However, 1-D CNN overlooks the spatial distribution patterns of HSI by flattening the spatial image into 1-D vector. In this context, 2-D CNN is more advocated since it applies some 2-D convolutions with different shapes on each 2-D image, which is capable of modeling the spatial information of HSI [18]. However, the spectral–spatial integral properties of hyperspectral imaging that produces a cube data might be wasted when dealing with 2-D CNN. To overcome this issue, 3-D CNN is adapted to HSI classification, which works simultaneously on the spectral and spatial domains using 3-D convolutions and serves as is a very promising approach directly handling the hyperspectral data cube [19]–[23]. Substantial studies based on these prototypes of CNN have been exploited recently, including data augmentation [24], [25], semisupervised and active learning [26]–[28], transfer learning [29]–[33], and multiscale scheme [34]–[39].

Although those existing CNN-based HSI classification methods have achieved remarkable performances, some drawbacks can still be observed. On the one hand, they mainly used the first-order pooling operation (e.g., max pooling, average pooling), which ignores the spectral correlations. On the other hand, the optimization of structural parameters (e.g., the layer

Manuscript received December 3, 2020; accepted December 27, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 41971279 and in part by the Fundamental Research Funds for the Central Universities under Grant B200202012. *(Corresponding author: Zhaohui Xue.)*

Zhaohui Xue and Mengxue Zhang are with the School of Earth Sciences and Engineering, Hohai University, Nanjing 211100, China (e-mail: zhaohui.xue@hhu.edu.cn).

Yifeng Liu is with the China Academy of Electronics and Information Technology, National Engineering Laboratory for Risk Perception and Prevention (NEL-RPP), Beijing 100041, China.

Peijun Du is with the Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, Nanjing University, Nanjing 210023, China, also with the School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China, and also with the Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing University, Nanjing 210023, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TGRS.2020.3048128.

Digital Object Identifier 10.1109/TGRS.2020.3048128

number, the convolutional kernel size, and the filter number) in most of the existing architectures is extremely laborious.

Recently, CNN with high-order statistics has exhibited big potentials in computer vision and pattern recognition [40]. Particularly, second-order pooling [41] can aggregate second-order statistics within or between features, and it can effectively exploit the correlation information between different channels. Moreover, second-order pooling has attracted some interest in HSI classification. In [42], the multiscale covariance maps were used in 2-D CNN to fully exploit the spatial–spectral information presented in HSI. In [43], the authors proposed a two-branch CNN network, where a covariance matrix between different spectral bands was constructed by clustering local neighboring pixels. In [44], a mixed CNN with covariance pooling was proposed to aggregate spectral–spatial feature maps for HSI.

However, there are two defects of the existing second-order-based CNNs for HSI classification. On the one hand, they still need arduous optimization of structural parameters since they are based on the traditional CNN architecture. On the other hand, the second-order pooling neglects the inherent spatial heterogeneity within a patch feature. In this circumstance, directly applying second-order pooling to HSI classification may lead to poor performance. Specifically, the class-oriented second-order statistics may be dominated by irrelevant pixels belonging to different classes when facing heterogeneous regions.

In this article, we propose a novel A-SPN. First, a first-order feature operator is designed to extract the spectral–spatial information of HSI. Then, an attention-based second-order pooling (A-SOP) operator is designed to model discriminative and representative features. Finally, a fully connected (FC) layer with softmax loss is used for classification. Instead of using a hierarchical structure composed of some interleaved convolution and pooling layers in traditional 2-D CNN or 3-D CNN, the proposed framework has a plain structure with much fewer structural parameters, thus leading to higher computational efficiency. In addition, SOP is employed to produce the representative features by considering both the spatial and spectral domains, whereas the attention mechanism introduces the diversity of neighboring pixels in the SOP.

To sum up, the main innovative contributions of our work can be summarized as follows.

1) We propose an A-SOP operator which can improve the second-order statistics of SOP by introducing spatial attention weighting. This spatial attention scheme can assign different magnitudes for disparate pixels by using a correlation matrix and a learnable cosine distance function. Consequently, A-SOP can produce more representative and discriminative features.
2) We propose an end-to-end network (A-SPN) that is unique in the literature. A-SPN avoids complex structural parameter optimization since we design a plain architecture composed of a first-order feature operator, an A-SOP, and a softmax classifier. Moreover, A-SPN can achieve superior performance, and it is more efficient compared with the other methods.

II. Related Work

As mentioned earlier, second-order pooling has been investigated for HSI classification. For example, covariance pooling combined with CNN was exploited in this community [42]–[44]. Notice that covariance pooling belongs to second-order pooling since it can extract the covariance matrix of different channels. The very limited studies have validated the representative capability of second-order pooling. However, second-order pooling is usually incorporated with a classical CNN architecture, which still needs complex structural parameter optimization.

Furthermore, CNN with attention mechanism has been intensively investigated in HSI classification community, e.g., the spatial attention [45], spectral attention [46]–[48], and spectral–spatial attention [49]–[51] mechanisms. Previous studies have demonstrated that attention mechanism can promote deep models to emphasize more salient features but suppress less useful ones. As for computer vision tasks, a global second-order pooling block was proposed to introduce higher order representation for large-scale natural image classification [52], and a second-order attention network was proposed to extract more powerful features for single image super-resolution [53].

Accordingly, a reasonable hypothesis is that the attention mechanism may improve second-order pooling for HSI classification since a spatial attention operation can enhance SOP in modeling class-oriented second-order statistics. Inspired by the idea, we intend to exploit A-SOP for HSI classification. Furthermore, there are two essential differences between related works and our proposed method.

1) We, for the first time, explore a second-order pooling with attention mechanism for HSI classification, which has significant differences in terms of motivation and application filed compared with the existing works.
2) We propose a plain network composed of three components with fewer structural parameters and higher computational efficiency compared with other CNN-based networks that focused on more complicated architectures or components.

III. Proposed Method

Let $\mathbf{X} \in \mathbb{R}^{N \times B}$ be an HSI with B bands and C classes for a total of N pixels. The proposed framework includes a first-order feature operator used to generate the shallow feature maps, an attention-based second-order pooling operator designed to obtain the deep feature maps, and a softmax classifier used to guide the training process and produce the classification results. Note that all the equipped units are differentiable during training, which promotes an end-to-end learning manner of the proposed network. A graphical illustration of the proposed framework is shown in Fig. 1.

A. First-Order Feature Operator

A first-order feature operator is designed to model the spectral–spatial information of HSIs. First, we adopt principal component analysis (PCA) to obtain a decorrelated HSI,

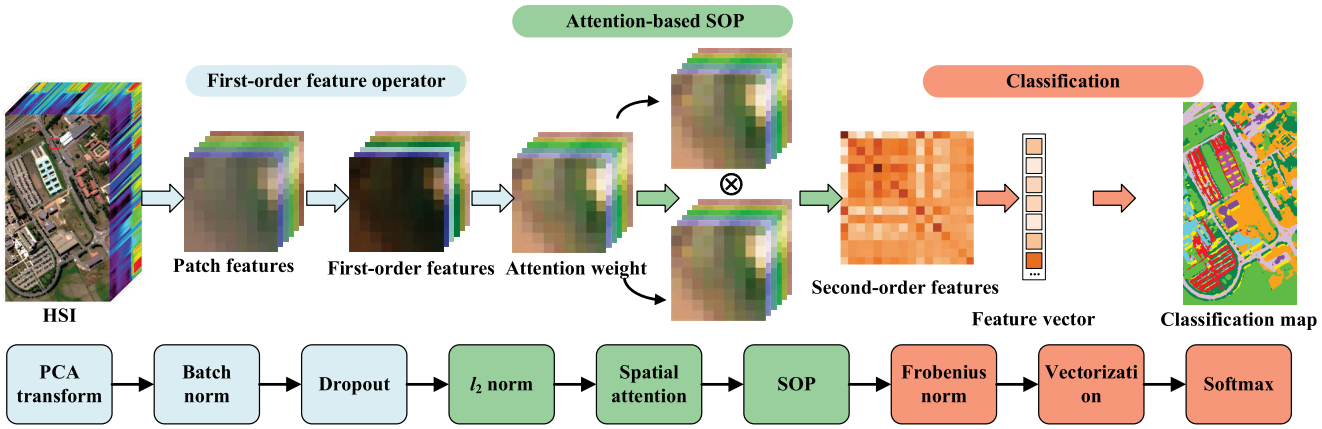

Fig. 1. Graphical illustration of A-SPN.

$\widehat{\mathbf{X}} \in \mathbb{R}^{N\times K}$ (where $K = B$ is the dimension of the output). Then, we use a squared moving window ($p \times p$) to generate patch features representing spatial information. Finally, batch normalization is used to adjust different patch features, and a dropout layer is added to reduce the risk of overfitting. We denote by $\mathbf{F}_{1\text{st}} \in \mathbb{R}^{M\times K}$ (where $M \equiv p \times p$ is the number of neighboring pixels) the first-order features learned from HSI. Instead of employing a CNN as a first-order feature operator, we use a first-order feature operator of a plain structure that contains fewer parameters in order to avoid the occurrence of gradient vanishing and reduce the risk of overfitting.

B. Second-Order Pooling Operator

To calculate the second-order statistical representations of the obtained first-order features, we propose a second-order pooling (SOP) operator taking the form

$$\mathbf{F}_{\text{SOP}} = \mathbf{F}_{1\text{st}}^{\top}\mathbf{F}_{1\text{st}} \tag{1}$$

where $\mathbf{F}_{\text{SOP}} \in \mathbb{R}^{K\times K}$ is a real symmetric matrix.

It is obvious that (1) calculates an inner product of each column–row vector pair in $\mathbf{F}_{1\text{st}}$. SOP can yield more discriminative and compact features due to the fact that: 1) by focusing on the correlations between spectra over identical spatial locations, SOP can make full use of the hyperspectral information; 2) by using inner production, SOP aggregates second-order statistics of patch features, which is beneficial to capture global spatial–spectral information; and 3) different from other handcraft second-order feature extractors, SOP is embedded into a network, and it is learnable and data-adaptive. However, SOP did not consider the diversity of different pixels in the neighborhood since the contribution of neighboring pixels to the learned features may be different.

C. Attention-Based SOP

An A-SOP is proposed to introduce the diversity of neighboring pixels in SOP. First, an ℓ_2 normalization is used to weaken the scale-dependence of the input features [54]. Then, an attention mask is learned based on the correlation matrix of neighboring pixels

$$\mathbf{S} = \mathbf{F}_{1\text{st}}\mathbf{F}_{1\text{st}}^{\top} \tag{2}$$

where $\mathbf{S} \in \mathbb{R}^{M\times M}$ is a correlation matrix of pairwise neighboring pixels. Apparently, $\mathbf{S}$ is also a real symmetric matrix. With the correlation matrix at hand, we design a learnable cosine distance function to measure the similarity between each neighboring pixel and the central pixel

$$\rho_i = \frac{\mathbf{S}_i \Lambda \mathbf{S}_0^{\top}}{\|\mathbf{S}_i\| \left\|\mathbf{S}_0^{\top}\right\|} \tag{3}$$

where $i \in \{1, 2, \ldots, M\}$, $\mathbf{S}_i$ is an arbitrary neighboring pixel vector, $\mathbf{S}_0$ is the central pixel vector within the neighborhood, and $\Lambda \in \mathbb{R}^{M\times M}$ is an additional diagonal matrix making ρ to be learnable. Note that $\mathbf{S}$ but not $\mathbf{F}_{1st}$ is used as the input in (3) is because $\mathbf{S}$ can model more representative information compared with $\mathbf{F}_{1st}$. Since $\mathbf{S}$ contains the spectral similarity information between two pixels at various positions, it might be more effective to leverage pixel vectors of $\mathbf{S}$ to calculate ρ.

The attention weights are then normalized into the one with unit sum by using a softmax function, which is found to give better convergence

$$w_i = \frac{e^{\rho_i + b_i}}{\sum_{j=1}^{M} e^{\rho_j + b_j}} \tag{4}$$

where $\sum w_i = 1$ for $i \in \{1, 2, \ldots, M\}$, b is a bias, and a diagonal weight matrix $\mathbf{W} \in \mathbb{R}^{M\times M}$ can then be constructed based on w_i.

Finally, the proposed A-SOP operator can be formulated as

$$\begin{aligned}\mathbf{F}_{\text{A-SOP}} &= (\mathbf{W}\mathbf{F}_{1\text{st}})^{\top}(\mathbf{W}\mathbf{F}_{1\text{st}}) \\ &= \mathbf{F}_{1\text{st}}^{\top}\mathbf{W}^2\mathbf{F}_{1\text{st}}.\end{aligned} \tag{5}$$

A-SOP inherits the main characteristics of SOP. In addition, A-SOP exclusively introduces the diversity of neighboring pixels when representing second-order features by using a data-adaptive and learnable weighting scheme. In this process, the pixels with larger weights play a decisive role, while the influence of pixels with smaller weights is suppressed.

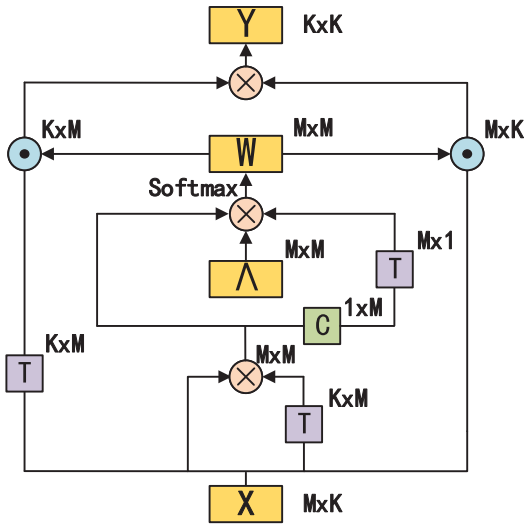

Fig. 2. Graphical illustration of the proposed A-SOP. Binary operations include "⊗" and "⊙," where "⊗" denotes matrix multiplication and "⊙" denotes elementwise multiplication. Unary operations include "**T**" and "**C**," where "**T**" denotes transpose and "**C**" denotes pick out the labeled pixel. The yellow boxes represent features or parameters.

TABLE I
ARCHITECTURE OF A-SPN

Component	Layer	Operation	Dimension
First-order feature operator	1	PCA transform	$M \times K$
	2	Batch normalization	$M \times K$
	3	Dropout	$M \times K$
A-SOP	4	ℓ_2 normalization	$M \times K$
	5	Correlation matrix	$M \times M$
		Cosine distance measure	$M \times M$
		Attention weight	$M \times M$
		SOP	$K \times K$
Classification	6	Frobenius normalization	$K \times K$
	7	Vectorization	$1 \times K^2$
	8	FC + Softmax	$1 \times C$

Consequently, A-SOP can discard redundant features and intensify second-order statistics. A graphical illustration of the proposed A-SOP is shown in Fig. 2.

D. Classification

Finally, a Frobenius normalization layer and a vectorization layer are used to vectorize the output features. Then, the last FC layer with softmax loss is used for classification. In this context, a novel A-SPN is built for HSI classification. The detailed architecture of A-SPN is listed in Table I, and an algorithmic description of A-SPN is stated in Algorithm 1. Note that, since each component of A-SPN is differentiable, the whole network can be trained in an end-to-end manner.

IV. Experimental Results

A. Hyperspectral Data Sets

1) The first data set was acquired by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over the Indian Pines test site in Northwestern Indiana. After the removal of the several noisy and water absorbed bands, the number of wavelength bands in the acquired image is 200 ranged from 0.4 to 2.5 μm, and the image size in the pixel is 145 × 145, with a moderate spatial resolution of 20 m. The false-color composite of the Indian Pines image and the corresponding ground-truth map are shown in Fig. 3. The ground-truth map totally includes 16 classes.
2) The second data set was acquired by the Reflective Optics Spectrographic Imaging System (ROSIS) sensor over the urban area of the University of Pavia, Italy. The number of wavelength bands in the acquired image is 103 ranged from 0.43 to 0.86 μm, and the image size is 610 × 340, with a very high spatial resolution of 1.3 m. The false-color composite of the University of Pavia image and the corresponding ground-truth map are shown in Fig. 4. The ground-truth map totally includes nine classes.
3) The third data set was acquired over the University of Houston campus and was released in the 2013 Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest. The number of wavelength bands in the acquired image is 144 ranged from 0.38 to 1.05 μm, and the image size in the pixel is 349 × 1905, with a high spatial resolution of 2.5 m. The false-color composite of the University of Houston image and the corresponding ground-truth map are shown in Fig. 5. The ground-truth map totally includes 15 classes.

Algorithm 1 A-SPN

Input: HSI $\mathbf{X}$, the component dimension K, and the window size p.
Step 1: Split all the labeled data into training and test sets.
Step 2: Initialize the weights for the network.
Step 3: Apply first-order feature operator on each pixel in $\mathbf{X}$ and obtain first-order features $\mathbf{F}_{1\text{st}} \in \mathbb{R}^{M \times K}$.
Step 4: Apply A-SOP on $\mathbf{F}_{1\text{st}}$ and obtain attention-based second features $\mathbf{F}_{\text{A-SOP}} \in \mathbb{R}^{K \times K}$.
Step 5: Adopt Softmax classifier to obtain the posterior probability.
Step 6: Optimize the loss function and update weights by using a mini-batch stochastic gradient descent algorithm.
Step 7: Repeat Step 4–6 until meet the stop criterion.
Output: Predict labels for the whole image with the trained model.

B. Experimental Settings

1) The hyperparameters for the proposed methods are listed in Table II, where Root Mean Square Propagation (RMSprop) is chosen as the optimizer [55]. For the nonstructural parameters, the training epoch is set as 15, and the batch size is set as 64 referred to [15]. On the other hand, the learning rate is analyzed in Section IV-C, and the exponential decay factor is empirically set to the same value as the learning rate. As for the structural parameters, the dropout rate is set as a common value of 0.5. We only optimize the patch size in Section IV-D. In this context, A-SPN avoids complex structural parameters optimization because it does not include convolutional layers and FC layers, both of which inherently have some structural parameters.

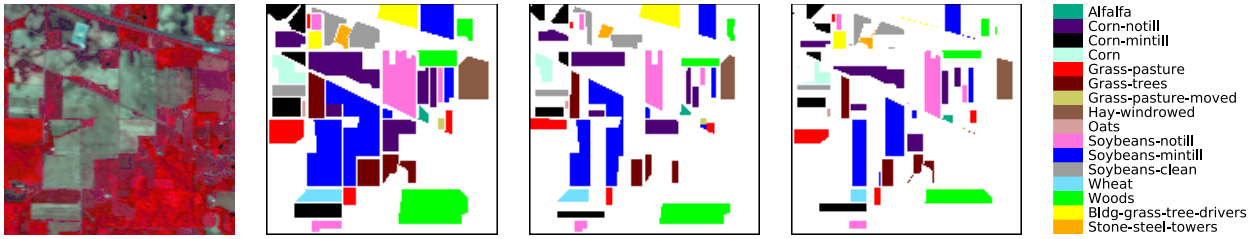

Fig. 3. Indian Pines data sets. (a) False-color composite image. (b) Ground-truth map. (c) Disjoint training samples. (d) Disjoint test samples.

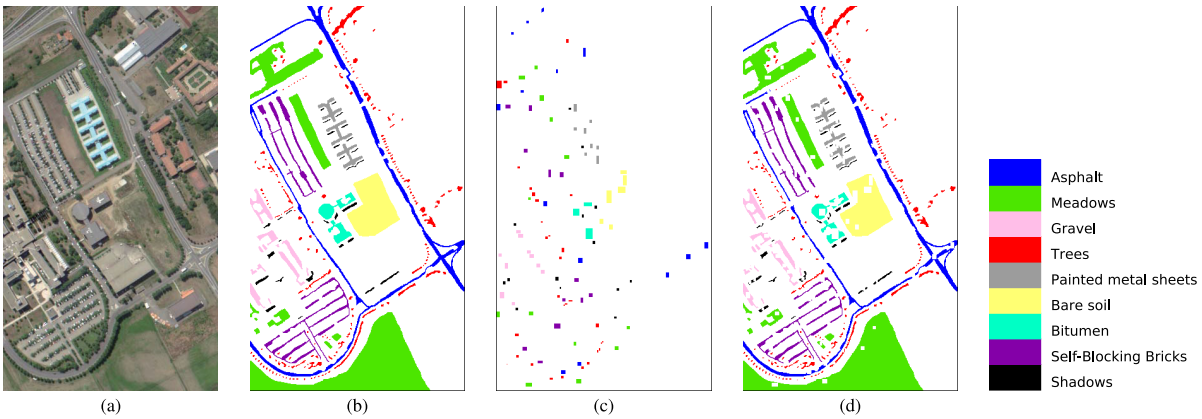

Fig. 4. University of Pavia data sets. (a) False-color composite image. (b) Ground-truth map. (c) Disjoint training samples. (d) Disjoint test samples.

TABLE II
HYPERPARAMETERS OF A-SPN

Type	Name	Value
Non-Structrual	Training epoch	15
	Batch size	64
	Initial learning rate	0.1
	Exponential decay	0.1
Structrual	Dropout rate	0.5
	Patch size	9×9

Precisely, there are only two structural parameters in A-SPN should be determined before training the network. However, traditional CNN-based methods usually have more structural parameters, such as the kernel size, the layer number, and the number of filters, to name a few.

2) The weight initialization is varied for different components in the proposed method. In the first-order feature operator, the PCA transform matrix is fixed during training, and batch normalization is initialized with $\gamma = 1$ and $\beta = 0$. For A-SOP, the diagonal elements of matrix Λ are initialized as one, and the bias b is initialized as zero. For classification, the weights of the FC layer are initialized symmetrically obeying truncated normal distribution with a mean of 0 and a standard deviation of 1e-4.
3) For performance comparison, we conduct support vector machine with a radial basis function (RBF-SVM) [56], SAE [11], 2-D convolutional neural network (2-D CNN) [19], 3-D convolutional neural network (3-D CNN) [19], spectral–spatial residual network (SSRN) [57], LSTM [15], and spectral–spatial unified network (SSUN) [15], including four classic methods and two state-of-the-art DL methods for HSI classification. For RBF-SVM, we use a grid search with cross-validation to optimize the parameters. For SAE, LSTM, and SSUN, the hyperparameters are in accordance with the corresponding articles. For 2-D CNN, 3-D CNN, and SSRN, the patch size is set as 9×9 for a fair comparison, and the other parameters are set following the original article. When using spatially disjoint training and test samples, the hyperparameters of SAE, LSTM, 2-D CNN, 3-D CNN, and the proposed A-SPN are kept the same, whereas the epochs of SSRN and SSUN are, respectively, set to 50 and 200 to obtain optimal results.
4) For sample selection, there are two sampling strategies considered in this article. In the first situation, we randomly select labeled samples from the ground-truth map as the training set in the experiments, i.e., 10% per class for the Indian Pines data set, 5% per class for the University of Pavia data set, and 50 samples per class for the University of Houston data set. In the second situation, we adopt a sampling strategy based on spatially disjoint samples, these spatially disjoint training and test samples are available from the GRSS DASE website.[1] More details of the training and test samples separation are listed in Tables III–V and visually inspected in Figs. 3–5. Note that we do not use data augmentation in both these situations for a fair comparison.
5) All implementations are carried out by using TensorFlow 1.12.0 and Keras 2.2.4 on a Ubuntu 16.04 long term support (LTS) desktop equipped with GeForce

[1] http://dase.grss-ieee.org

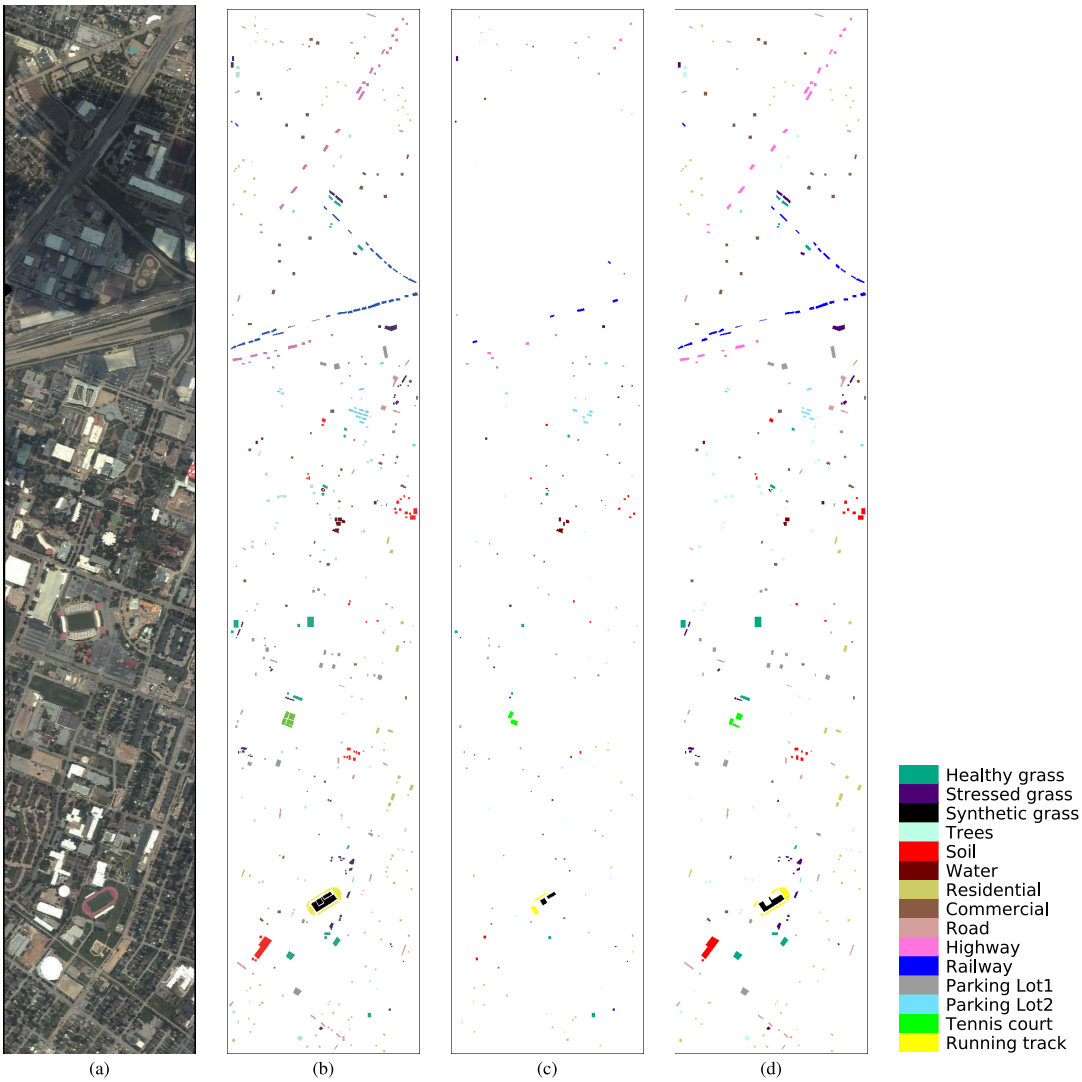

Fig. 5. University of Pavia data sets. (a) False-color composite image. (b) Ground-truth map. (c) Disjoint training samples. (d) Disjoint test samples.

TABLE III
NUMBER OF TRAINING AND TEST SAMPLES USED FOR THE INDIAN PINE DATA SET

Class name	Random		Disjoint		Total
	Training	Test	Training	Test	
Alfalfa	5	41	29	17	46
Corn-notill	143	1285	762	666	1428
Corn-mintill	83	747	435	395	830
Corn	24	213	146	91	237
Grass-pasture	48	435	232	251	483
Grass-trees	73	657	394	336	730
Grass-pasture-moved	3	25	16	12	28
Hay-windrowed	48	430	235	243	478
Oats	2	18	10	10	20
Soybeans-notill	97	875	470	502	972
Soybeans-mintill	246	2209	1424	1031	2455
Soybeans-clean	59	534	328	265	593
Wheat	21	184	132	73	205
Woods	127	1138	728	537	1265
Bldg-grass-tree-drivers	39	347	291	95	386
Stone-steel-towers	9	84	57	36	93
Total (16 classes)	1027	9222	5689	4560	10249

GTX 1070 (8 GB), Intel Xeon E3 CPU, and 32-GB RAM. The source codes will be available online.[2] The experimental results are reported by averaging the outputs of twenty independent runs. Some quantitative metrics, including overall accuracy (OA), average accuracy (AA), Kappa coefficient (κ), training time (s), test time (s), prediction time (s), and memory usage (MB), are adopted to evaluate the classification performance.

[2]https://github.com/snowzm/A-SPN

C. Analysis of Initial Learning Rate

In the first experiment, we adopt a swift strategy [58] to obtain an appropriate value for the initial learning rate. Specifically, we determine a sequence of learning rates to train A-SPN, i.e., [1e-5, 3e-5, 5e-5, 1e-4, 3e-4, 5e-4, 1e-3, 3e-3, 5e-3, 1e-2, 3e-2, 5e-2, 1e-1, 3e-1, 5e-1].

As depicted in Fig. 6, the training loss decreases as the increase in the learning rate before learning rate less than 1e-1. The minimum value of training loss occurs when the learning rate is equal to 1e-1 for the three data sets. The training loss suddenly increases when the learning rate is larger than 1e-1. We choose 1e-1 as the initial learning rate since the turning point represents an ideal choice.

TABLE IV
NUMBER OF TRAINING AND TEST SAMPLES USED FOR THE UNIVERSITY OF PAVIA DATA SET

Class name	Random		Disjoint		Total
	Training	Test	Training	Test	
Asphalt	332	6299	548	6083	6631
Meadows	932	17717	540	18109	18649
Gravel	105	1994	392	1707	2099
Trees	153	2911	524	2540	3064
Painted metal sheets	67	1278	265	1080	1345
Bare soil	251	4778	532	4497	5029
Bitumen	67	1263	375	955	1330
Self-Blocking Bricks	184	3498	514	3168	3682
Shadows	47	900	231	716	947
Total (9 classes)	2138	40638	3921	38855	42776

TABLE V
NUMBER OF TRAINING AND TEST SAMPLES USED FOR THE UNIVERSITY OF HOUSTON DATA SET

Class name	Random		Disjoint		Total
	Training	Test	Training	Test	
Healthy grass	50	1201	198	1053	1251
Stressed grass	50	1204	190	1064	1254
Synthetic grass	50	647	192	505	697
Trees	50	1194	188	1056	1244
Soil	50	1192	186	1056	1242
Water	50	275	182	143	325
Residential	50	1218	196	1072	1268
Commercial	50	1194	191	1053	1244
Road	50	1202	193	1059	1252
Highway	50	1177	191	1036	1227
Railway	50	1185	181	1054	1235
Parking Lot1	50	1183	192	1041	1233
Parking Lot2	50	419	184	285	469
Tennis court	50	378	181	247	428
Running track	50	610	187	473	660
Total (15 classes)	750	14279	2832	12197	15029

D. Analysis of Window Size

In the second experiment, we analyze the impact of patch size on classification accuracy. First, we investigate the impact of using randomly selected samples. As shown in Fig. 7(a), the OAs increase as the patch size also increases for different cases. When the patch size is smaller than 9×9, the increase in OAs is significant. Then, the OAs remain stable when the patch size becomes larger than 9×9. For example, in the case of the Indian pines data set, the OA reaches 99.24% when the patch size is 9×9. Second, we investigate the impact of using spatially disjoint samples. As shown in Fig. 7(b), the OAs increase as the patch size also increases. However, the overall OAs are lower than that of using randomly selected samples in different cases. For example, in the case of the University of Houston data set, the OA only reaches 88.40% when the patch size increases to 9×9. Although a relatively larger patch size may lead to higher accuracy for some data sets, the between-class boundary will be unavoidably mixed in the classification map. In this context, we choose 9×9 as the suitable patch size in the following experiments.

E. Generalization Performance

In the third experiment, we evaluate the generalization performance of different methods based on limited training samples. As shown in Fig. 8, A-SPN outperforms others in most of the cases for different data sets. For the Indian Pines data set, A-SPN obtains an OA of 90.57±1.74% when the number of labeled samples per class is 15, which is

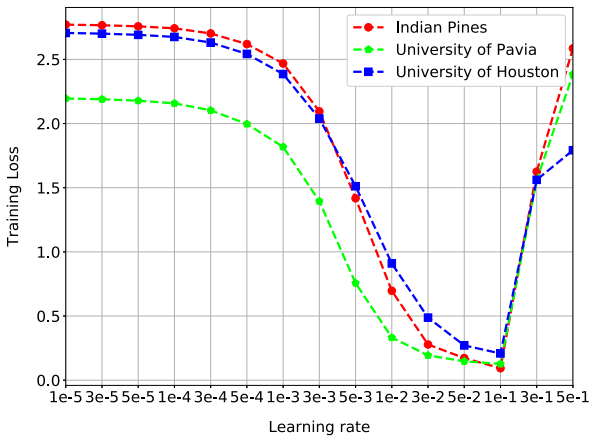

Fig. 6. Evolution of training loss as a function of initial learning rate.

4.91%–33.50% higher than other methods. For the other two data sets, SSRN provides competitive performance. However, clear improvements of OA obtained by A-SPN can still be found when the number of labeled samples per class is 10 and 15 for the University of Pavia data set. Similar results can be found when the number of labeled samples per class is 15, 20, and 25 for the University of Houston data set.

F. Classification Results With Randomly Selected Samples

In the fourth experiment, we compare the classification performance of different methods based on enough training samples, i.e., 10% per class for the Indian Pines data set, 5% per class for the University of Pavia data set, and 50 samples per class for the University of Houston data set.

For the Indian Pines data set, the classification accuracies are reported in Table VI. A-SPN obtains the highest OA of 99.24±0.19%, which is 0.80%–21.24% higher than the other methods. In addition, A-SPN also produces significant improvements in terms of AA and κ compared with others. As for the class-specific accuracies, A-SPN obtains higher accuracies for a total of 15 classes. The corresponding classification maps are visually depicted in Fig. 9, where A-SPN yields more accurate and smooth results, especially for the Grass-pasture-moved and Hay-windrowed classes.

For the University of Pavia data set, the classification accuracies are reported in Table VII. A-SPN obtains the highest OA of 99.65±0.08%, which is 0.13%–9.09% higher than the other methods. For AA and κ, A-SPN also produces higher accuracies than others. As for class-specific accuracies, the proposed method obtains better classification results for a total of five classes. Especially, for the Meadows class that is the toughest class for accurate classification, A-SPN obtains 100% accuracy. The classification maps are shown in Fig. 10, where A-SPN yields more accurate and smooth results. The corresponding regions for the Meadows and Painted metal sheets classes are very accurate compared with the ground-truth map, which is in accordance with the class-specific accuracies listed in Table VII.

As for the University of Houston data set, A-SPN obtains an OA of 97.27±0.60%, which is 1.38%–18.23% higher than

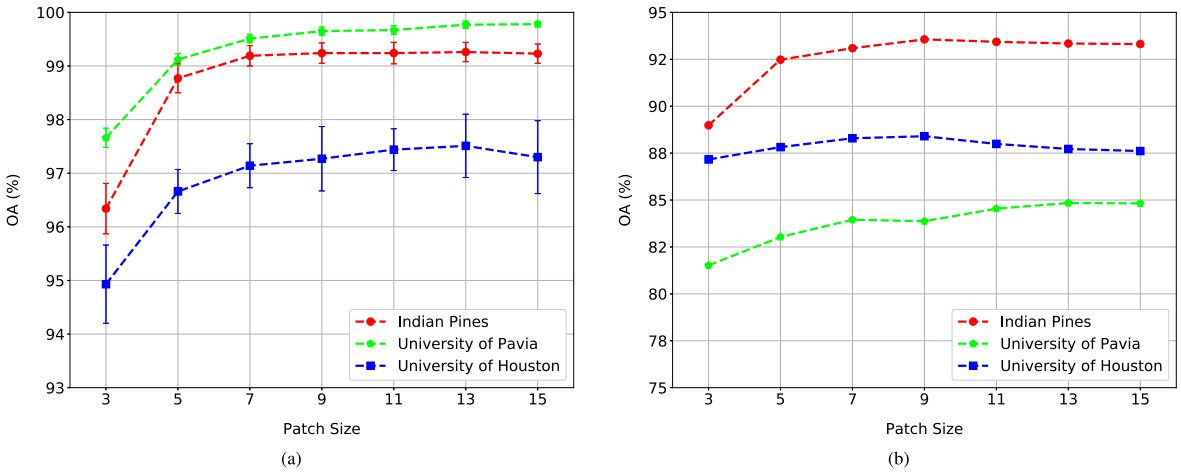

Fig. 7. Evolution of A-SPN of OA as a function of patch size using (a) randomly selected samples and (b) spatially disjoint samples. Error bars represent standard deviations.

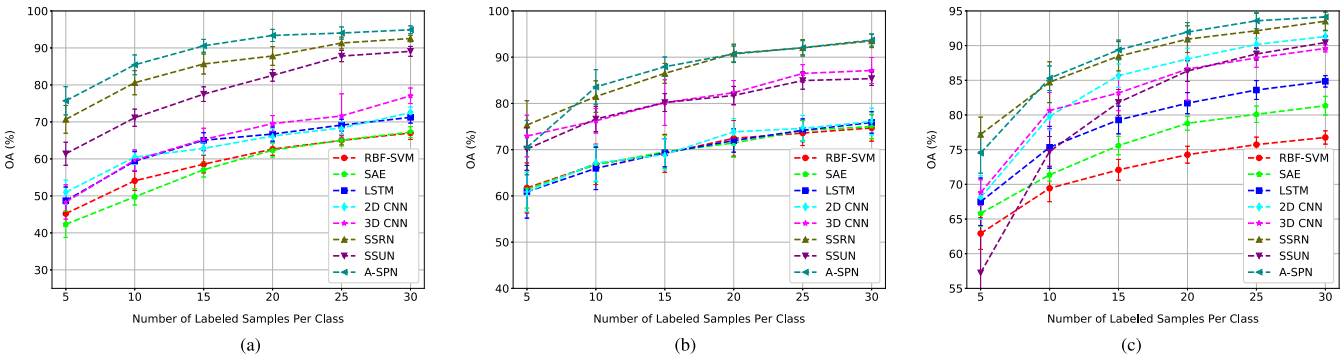

Fig. 8. Evolution of OA as a function of number of randomly selected labeled samples per class. Error bars represent standard deviations. (a) Indian Pines. (b) University of Pavia. (c) University of Houston.

TABLE VI
CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT METHODS FOR THE INDIAN PINES DATA SET
(10% LABELED SAMPLE PER CLASS USED FOR TRAINING)

Class	RBF-SVM	SAE	LSTM	2D CNN	3D CNN	SSRN	SSUN	A-SPN
Alfalfa	22.07±9.08	54.32±14.6	61.71±13.1	79.27±14.1	87.07±8.73	94.88±4.93	97.32±5.71	**98.78±1.22**
Corn-notill	71.03±2.15	70.15±3.06	79.79±2.03	82.51±2.70	90.83±3.12	97.41±1.90	97.93±0.50	**98.75±0.81**
Corn-mintill	52.74±4.81	68.27±5.01	73.04±2.88	88.50±2.55	91.41±1.73	98.35±0.82	97.66±0.98	**99.26±0.62**
Corn	42.61±6.82	59.25±10.6	64.25±5.76	85.31±7.23	90.85±4.09	98.73±1.98	96.20±3.31	**99.18±1.21**
Grass-pasture	84.94±2.94	85.83±1.74	87.93±2.75	93.43±3.09	95.24±2.25	97.75±1.15	96.71±2.11	**97.92±1.77**
Grass-trees	95.94±1.76	94.43±2.52	95.93±1.35	95.94±1.93	97.32±1.06	99.30±0.82	99.38±0.44	**99.95±0.07**
Grass-pasture-moved	65.40±13.87	72.80±8.54	78.80±9.22	64.80±15.6	89.20±13.63	86.80±16.8	91.20±9.93	**100.00±0.0**
Hay-windrowed	98.67±0.81	97.63±0.58	97.48±1.41	98.86±0.83	99.84±0.28	97.72±5.92	99.38±0.44	**100.00±0.0**
Oats	7.78±11.16	31.11±17.4	62.50±13.5	**98.33±3.56**	62.22±18.05	82.78±14.8	95.56±10.2	93.33±8.71
Soybeans-notill	70.46±11.17	73.11±5.03	80.27±2.75	90.48±1.94	90.03±3.33	97.78±1.58	97.35±1.24	**98.80±0.98**
Sobeans-mintill	87.24±1.36	79.68±1.18	83.68±1.00	93.11±0.93	95.64±2.15	97.94±1.03	99.12±0.48	**99.37±0.62**
Sobeans-clean	55.88±3.03	67.91±1.36	75.56±2.58	84.99±5.37	84.42±4.93	96.74±2.09	97.55±2.26	**98.32±0.73**
Wheat	95.11±2.49	90.76±1.88	98.23±1.11	98.21±1.52	98.86±1.36	98.97±1.18	99.24±0.50	**99.54±0.49**
Woods	96.67±0.86	91.65±2.25	93.65±2.12	98.17±0.84	98.65±0.60	99.86±0.18	99.67±0.27	**99.95±0.12**
Bldg-grass-tree-driver	46.24±3.36	53.89±3.68	63.29±5.86	90.11±3.40	92.33±7.52	98.47±2.81	99.11±0.90	**99.12±1.43**
Stone-steel-towers	80.12±6.11	88.57±4.16	89.88±5.10	95.60±4.02	88.21±9.37	95.60±4.55	96.79±3.77	**99.94±0.26**
OA (%)	78.00±0.72	78.58±0.69	83.41±0.72	91.29±0.42	93.75±0.99	98.09±0.63	98.44±0.24	**99.24±0.19**
AA (%)	67.06±1.85	73.71±2.56	80.37±1.38	89.79±1.08	90.76±1.78	96.19±1.94	97.52±0.83	**98.89±0.60**
$\kappa\times100$	74.64±0.85	75.57±0.83	81.06±0.83	90.07±0.49	92.87±1.12	97.83±0.71	98.23±0.27	**99.11±0.22**

the other methods according to Table VIII. In addition, A-SPN produces an AA of 97.74% and a $\kappa \times 100$ of 97.04, which are, respectively, 1.12%–17.70% and 1.49–19.71 higher than others. As for the class-specific accuracies, A-SPN obtains the highest accuracies for a total of 11 classes, especially for the class of Synthetic Grass, Soil, Water, Tennis court, and Running track, and the accuracies are 100%. Significant improvements in terms of classification accuracy can be visually inspected from Fig. 11, where A-SPN yields more clear classification results for the north part of this scene. Those performance improvements obtained by our method are precious considering the shadowing region (located in the

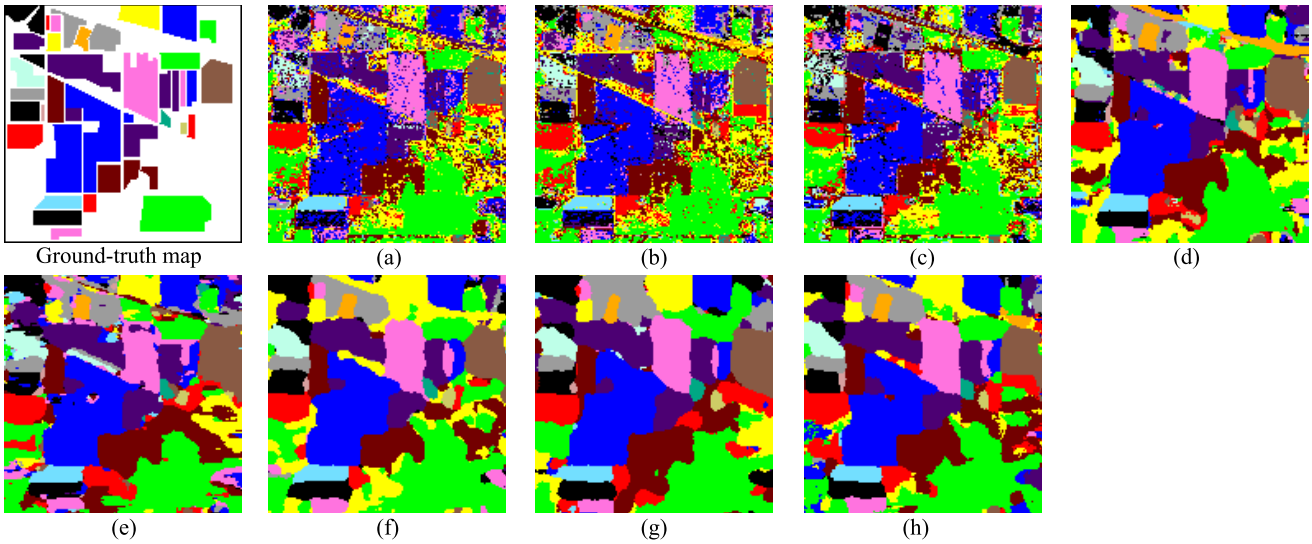

Fig. 9. Classification maps obtained by different methods for the Indian Pines data set. (a) RBF-SVM (78.00%). (b) SAE (78.58%). (c) LSTM (83.41%). (d) 2-D CNN (91.29%). (e) 3-D CNN (93.75%). (f) SSRN (98.09%). (g) SSUN (98.44%). (h) A-SPN (99.24%).

TABLE VII
CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT METHODS FOR THE UNIVERSITY OF PAVIA DATA SET (5% LABELED SAMPLE PER CLASS USED FOR TRAINING)

Class	RBF-SVM	SAE	LSTM	2D CNN	3D CNN	SSRN	SSUN	A-SPN
Asphalt	91.99±1.16	93.74±0.62	93.33±0.69	96.02±0.68	97.13±1.53	**99.87±0.07**	99.11±0.27	99.79±0.15
Meadows	98.90±0.22	97.03±0.38	97.58±0.25	98.08±0.24	99.19±0.44	99.61±0.49	99.84±0.06	**100.00±0.0**
Gravel	61.28±5.15	78.49±1.38	78.82±2.46	74.18±1.80	92.59±2.50	**98.45±1.38**	96.41±1.36	97.96±1.24
Trees	91.17±1.10	91.46±1.06	92.90±1.11	99.06±0.37	98.49±0.45	98.67±1.11	**99.22±0.34**	98.59±0.43
Painted metal sheets	99.24±0.31	99.39±0.21	99.45±0.26	99.88±0.17	**100.00±0.0**	**100.00±0.0**	99.80±0.12	**100.00±0.0**
Bare soil	71.08±1.57	88.45±1.42	89.63±0.87	88.77±1.72	92.94±3.02	99.86±0.53	99.59±0.24	**99.99±0.03**
Bitumen	69.91±8.41	85.10±0.79	85.19±3.34	82.84±2.01	93.44±1.79	99.65±0.53	96.59±2.22	**99.91±0.17**
Self-Blocking Bricks	90.35±1.82	85.30±1.52	87.10±1.25	93.62±0.94	94.09±2.52	98.55±1.14	98.15±0.52	**98.68±0.59**
Shadows	99.86±0.12	99.18±0.43	99.58±0.28	99.92±0.11	99.52±0.64	**99.95±0.05**	99.54±0.32	99.90±0.04
OA (%)	90.56±0.37	92.86±0.37	93.55±0.24	94.80±0.33	97.18±0.53	99.52±0.22	99.23±0.13	**99.65±0.08**
AA (%)	85.98±1.04	90.91±0.41	91.51±0.47	92.49±0.47	96.38±0.60	99.40±0.20	98.70±0.34	**99.42±0.14**
$\kappa\times100$	87.26±0.52	90.62±0.49	91.42±0.33	93.21±0.44	96.25±0.72	99.37±0.29	98.98±0.16	**99.53±0.11**

north part of this scene), which is very tough for accurate classification.

Although A-SPN has achieved the highest OA, AA, and κ on three data sets, there are exceptions for certain classes. For example, CNN performs better on the class Oats than A-SPN on the Indian Pines data set, and SSRN performs better on several classes than A-SPN on the University of Pavia data set and the University of Houston data set. There are some probable explanations for this phenomenon. First, A-SPN does not manipulate local features, thus ignoring local spatial information that is intensively utilized by convolutional layers in 2-D CNN or SSRN. Second, 2-D CNN has fewer parameters than A-SPN, and it may generalize better on Oats with very limited training samples. Finally, SSRN is a deeper network than A-SPN, and it may yield more abstract and higher level features that are beneficial to recognize some specific classes. Accordingly, it may be interesting to explore the cooperation of A-SPN with traditional CNN-based methods.

G. Classification Results With Spatially Disjoint Samples

In the final experiment, we compare the classification performance of different methods with spatially disjoint training and test samples of the three data sets. The experimental results (in terms of OA, AA, and κ) are reported on Table IX. As we can see, the degradation of performance for almost all methods is significant compared with previous classification results even though the disjointed training set actually contains more training samples.

As for the Indian Pines data set, it can be noticed that SAE, 2-D CNN, 3-D CNN, SSUN, SSRN, and A-SPN suffer an accuracy deterioration. In addition, the performance of 2-D CNN and 3-D CNN endures a drastic decline. There are two possible reasons for this phenomenon. For one thing, compared with disjointed samples, randomly selected samples may lead to a serious spatial overlap on the training and test samples, which leads to an overestimation of classification performance. For another, 2-D or 3-D convolutions overly focus on the spatial features, while the spatial resolution of this data set is relatively low. Although A-SPN also experiences performance degradation, it still achieves the highest OA of 93.56%, AA of 84.44, and κ of 0.93, which are, respectively, 6.17%–20.82%, 4.37%–18.73%, and 0.06–0.22 higher than others.

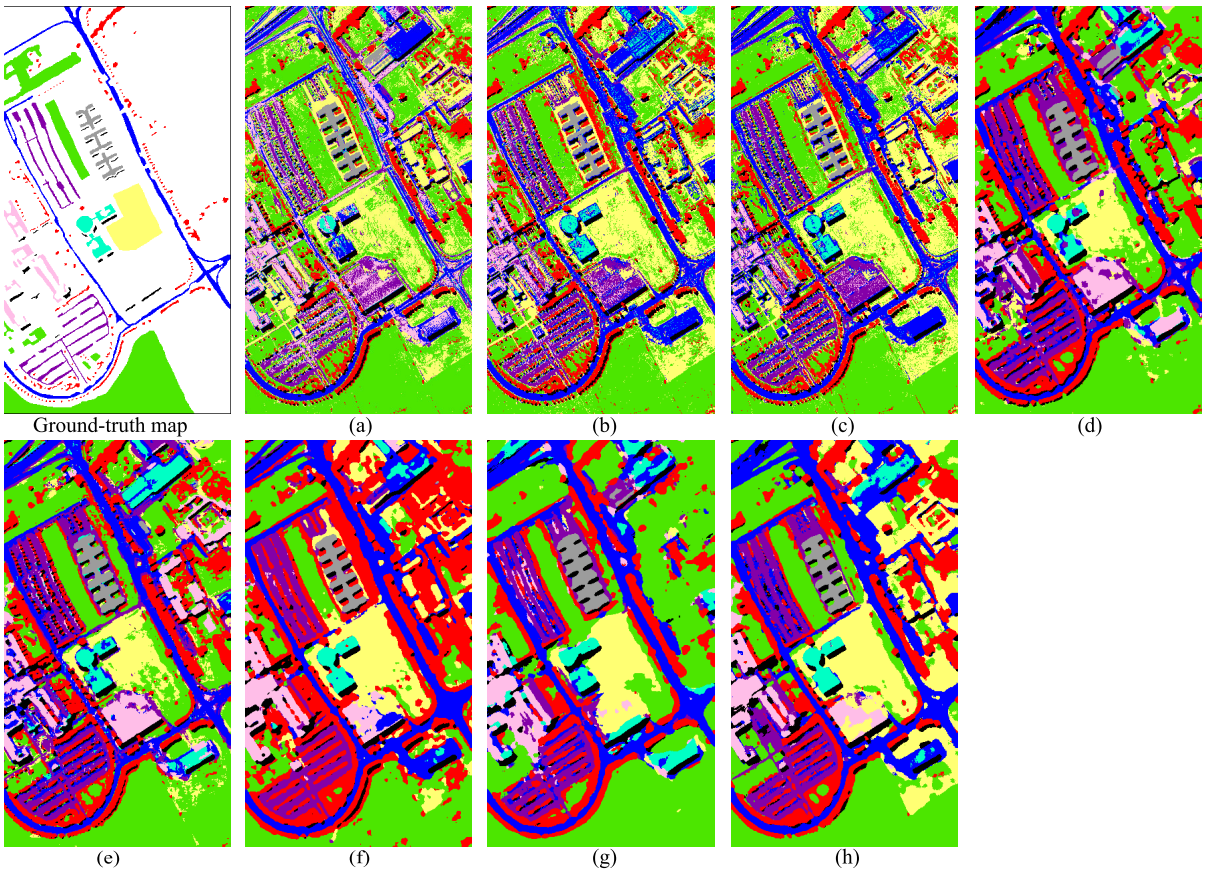

Fig. 10. Classification maps obtained by different methods for the University of Pavia data set. (a) RBF-SVM (90.56%). (b) SAE (92.86%). (c) LSTM (93.55%). (d) 2-D CNN (94.80%). (e) 3-D CNN (97.18%). (f) SSRN (99.52%). (g) SSUN (99.23%). (h) A-SPN (99.65%).

TABLE VIII

CLASSIFICATION ACCURACIES OBTAINED BY DIFFERENT METHODS FOR THE UNIVERSITY OF HOUSTON DATA SET (50 LABELED SAMPLE PER CLASS USED FOR TRAINING)

Class	RBF-SVM	SAE	LSTM	2D CNN	3D CNN	SSRN	SSUN	A-SPN
Healthy grass	94.36±5.00	**97.97±1.14**	95.76±2.60	97.01±1.49	95.08±6.02	97.69±1.48	94.08±3.46	96.68±2.97
Stressed grass	91.60±5.60	97.28±1.29	97.03±1.13	97.63±1.28	98.87±0.72	**99.14±0.63**	92.62±4.18	98.35±1.48
Synthetic grass	98.99±0.38	98.02±1.50	98.67±0.80	99.20±1.04	98.87±0.98	99.91±0.10	99.88±0.18	**100.00±0.0**
Trees	92.93±0.40	94.46±1.28	97.14±1.44	98.58±1.65	98.61±1.50	**99.19±1.65**	95.26±2.24	99.04±1.68
Soil	98.05±0.98	95.50±3.06	97.23±0.79	97.79±1.32	99.43±1.12	99.87±0.28	97.97±0.80	**100.00±0.0**
Water	86.67±4.47	94.98±2.42	98.29±0.98	98.11±0.90	98.33±1.40	99.42±0.91	98.47±2.42	**100.00±0.0**
Residential	80.42±2.71	73.75±2.42	82.16±4.43	90.36±1.62	88.78±2.64	95.76±3.54	95.09±2.04	**95.82±2.57**
Commercial	51.33±6.07	73.82±6.43	79.42±3.45	88.52±2.69	88.40±2.25	83.27±7.09	82.91±2.81	**91.72±3.66**
Road	76.46±1.87	71.51±1.67	77.95±3.09	89.18±2.37	87.56±5.42	87.03±12.2	92.64±3.33	**92.81±2.48**
Highway	70.20±8.36	86.64±3.89	89.22±2.53	93.08±2.61	89.02±6.23	96.67±4.64	99.13±0.58	**99.20±1.14**
Railway	62.58±3.87	76.07±3.79	81.81±2.39	92.37±3.31	88.07±3.15	97.50±1.83	94.24±3.81	**98.77±2.06**
Parking Lot1	55.08±8.73	70.74±5.41	79.70±2.29	95.25±1.39	82.87±17.15	95.50±5.52	91.64±2.66	**95.74±2.74**
Parking Lot2	44.65±7.83	49.79±3.26	59.09±4.08	98.35±1.72	98.14±1.10	**98.40±1.54**	97.64±2.13	98.03±2.31
Tennis court	98.31±1.14	96.72±2.42	99.26±1.33	99.05±1.85	99.84±0.24	**100.00±0.0**	99.92±0.17	**100.00±0.0**
Running track	98.94±0.52	99.08±0.57	99.10±0.42	98.98±2.24	99.30±0.80	**100.00±0.0**	96.46±3.45	**100.00±0.1**
OA (%)	79.04±0.84	84.66±1.15	88.39±0.71	94.77±0.26	92.87±2.70	95.89±1.89	94.35±0.76	**97.27±0.60**
AA (%)	80.04±0.65	85.09±0.97	88.78±0.53	95.56±0.27	94.08±2.14	96.62±1.55	95.20±0.65	**97.74±0.50**
$\kappa\times100$	77.33±0.90	83.42±1.24	87.51±0.83	94.34±0.28	92.28±2.92	95.55±2.05	93.89±0.82	**97.04±0.65**

As for the University of Pavia data set, all methods suffer certain performance deterioration. In particular, SSRN and SSUN are significantly affected by this sampling strategy. By comparison, A-SPN, 2-D CNN, and 3-D CNN are less affected. In this case, A-SPN achieves the highest OA of 83.87% and κ of 0.80, which are, respectively, 1.71%–11.63% and 0.02–0.13 higher than others, whereas the 3-D CNN achieves the highest AA, which is 0.94% higher than that of A-SPN.

As for the University of Houston data set, the performance degradation is exhibited in all methods. Among which, 2-D CNN and SSUN suffer serious influence. This deterioration can also be seen from A-SPN. Nevertheless, A-SPN obtains the highest OA of 88.40%, AA of 89.72%, and κ of 0.87,

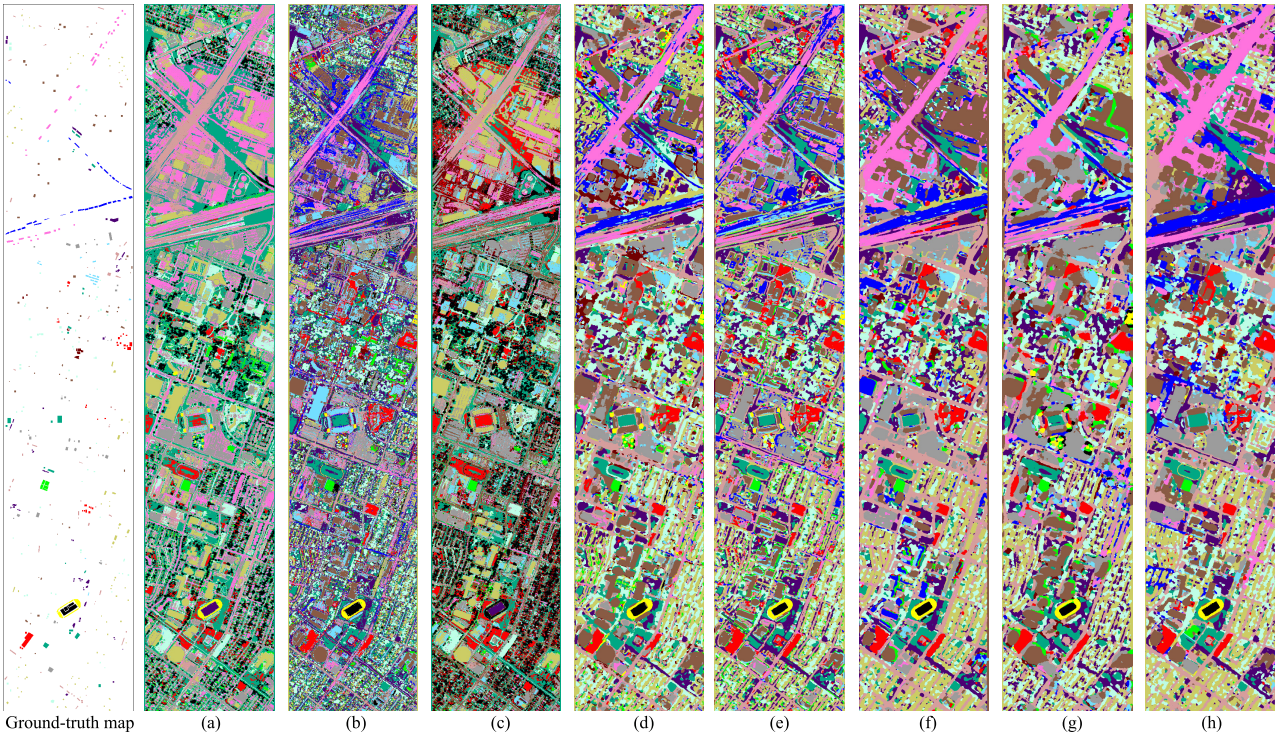

Fig. 11. Classification maps obtained by different methods for the University of Houston data set. (a) RBF-SVM (79.04%). (b) SAE (84.66%). (c) LSTM (88.39%). (d) 2-D CNN (94.77%). (e) 3-D CNN (92.87%). (f) SSRN (95.89%). (g) SSUN (94.35%). (h) A-SPN (97.27%).

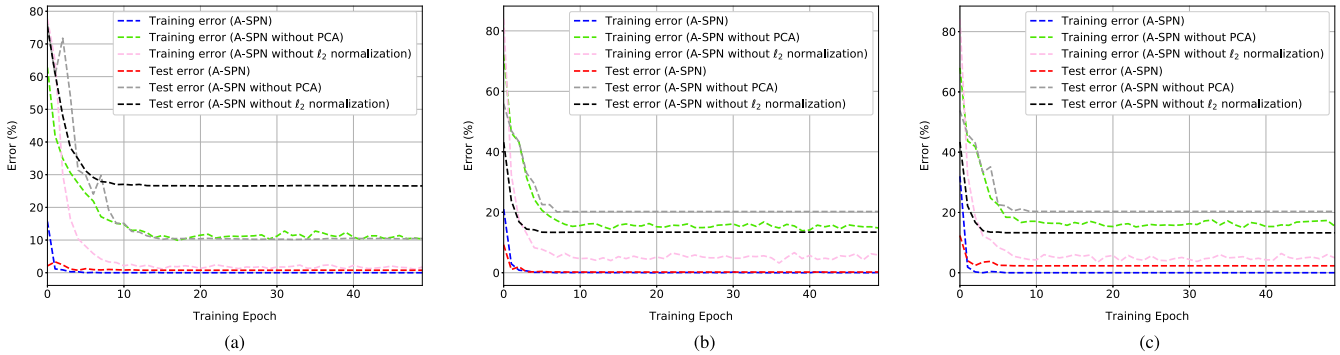

Fig. 12. Effects of PCA and ℓ_2 normalization for the convergence of A-SPN. (a) Indian Pines. (b) University of Pavia. (c) University of Houston.

which are, respectively, 3.54%–12.98%, 2.41%–12.32%, and 0.03–0.13 higher than others. As a result, when using spatially disjoint samples, which is a challenging scenario to evaluate classification performance, A-SPN is better than other counterparts.

V. Discussion

A. Effects of Different Components of A-SPN

In order to exploit the factors affecting convergence of A-SPN, we design two variations: A-SPN without PCA and A-SPN without ℓ_2 normalization. Note that we only analyze the effects of ℓ_2 normalization because the usage of two other normalizations, including batch normalization and Frobenius normalization, is referred to previous work [41]. Fig. 12 illustrates the convergence of A-SPN and its two variations based on training and test errors. In general, both PCA and ℓ_2 normalization have significant effects on the convergence of A-SPN. Precisely, without PCA and ℓ_2 normalization, A-SPN converges to a very unstable level with significant fluctuations, and the converged training and test errors are relatively higher than the normal level, i.e., the converged training errors are larger than 10% and the test errors are around 5% for the three data sets. In addition, without PCA and ℓ_2 normalization, A-SPN converges very slowly, i.e., the two variations meet the convergence level until the training epoch reaches 10, whereas the training epoch is less than 5 for A-SPN. This is due to the fact that PCA transformation can accelerate convergence since it enables a larger initial learning rate, and the ℓ_2 normalization can remove the scale-dependence of the input features [54]. They jointly affect the generalization capacity during the test phase. Therefore, PCA and ℓ_2 normalization are beneficial for A-SPN to reach a better convergence level.

TABLE IX
CLASSIFICATION RESULTS USING SPATIALLY DISJOINT SAMPLES DURING THE TRAINING AND TEST STAGES

Method	Indian Pines			University of Pavia			University of Houston		
	OA (%)	AA (%)	κ	OA (%)	AA (%)	κ	OA (%)	AA (%)	κ
RBF-SVM	82.25	76.93	0.81	72.24	74.45	0.67	80.75	83.62	0.79
SAE	78.43	73.59	0.77	76.46	70.44	0.71	76.26	79.67	0.75
LSTM	85.04	80.07	0.84	74.05	72.09	0.69	79.63	82.45	0.78
2D CNN	72.74	68.48	0.71	79.78	74.94	0.76	75.42	77.40	0.74
3D CNN	74.40	65.71	0.73	82.16	**80.23**	0.78	82.75	85.85	0.82
SSRN	87.39	75.39	0.87	77.53	75.59	0.73	84.86	87.31	0.84
SSUN	83.76	80.01	0.83	75.01	72.31	0.70	82.09	84.85	0.81
A-SPN	**93.56**	**84.44**	**0.93**	**83.87**	79.29	**0.80**	**88.40**	**89.72**	**0.87**

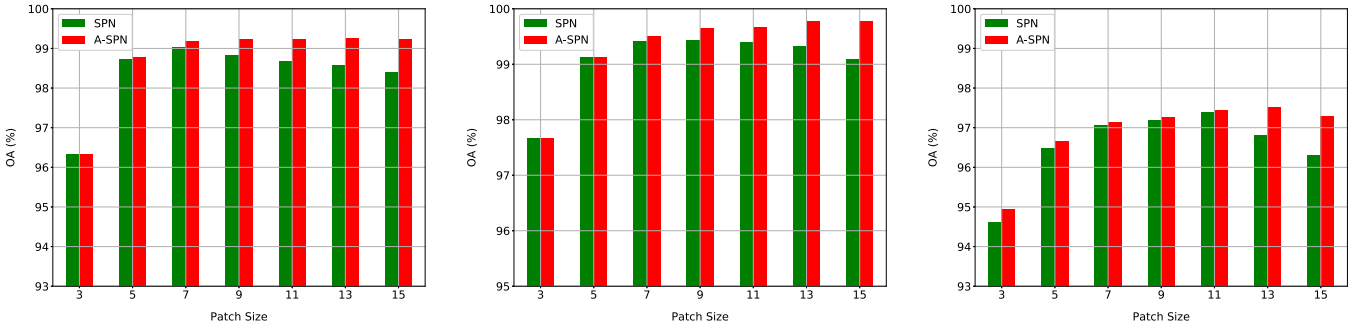

Fig. 13. Comparison of A-SPN with SPN in terms of OA using randomly selected samples. (a) Indian Pines. (b) University of Pavia. (c) University of Houston.

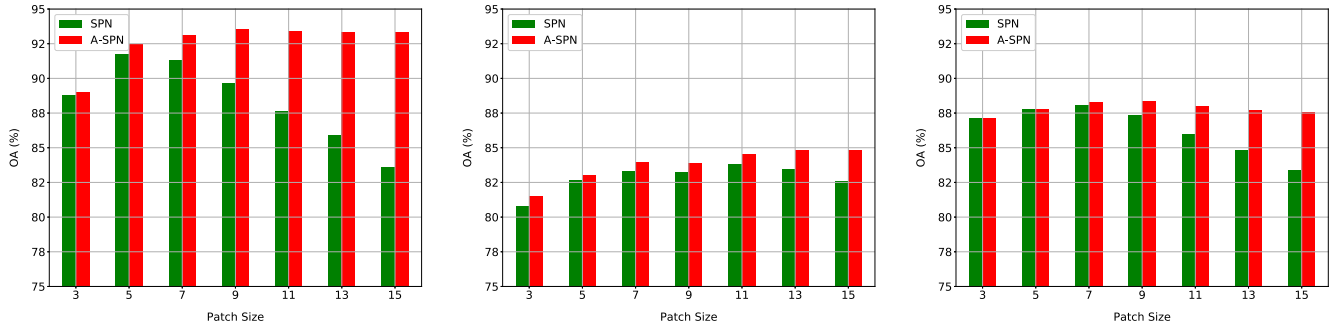

Fig. 14. Comparison of A-SPN with SPN in terms of OA using spatially disjoint samples. (a) Indian Pines. (b) University of Pavia. (c) University of Houston.

Moreover, we design A-SPN without attention (SPN) to analyze the effect of attention mechanism on classification accuracy. As shown in Fig. 13, A-SPN obviously performs better than SPN when using randomly selected samples. For instance, A-SPN obtains an OA of 99.27% for the Indian Pines data set with a patch size of 9×9, whereas SPN obtains an OA of 98.82%. Moreover, SPN is more sensitive to patch size, whereas A-SPN is more stable. For example, the OAs obtained by A-SPN stably increase as the patch size becomes larger, whereas the OAs of SPN significantly decline when the patch size becomes larger. Furthermore, we compare A-SPN with SPN when using spatially disjoint samples. The results are shown in Fig. 14. As we can see, A-SPN is still better than SPN, and their difference is also more salient when the patch size becomes larger. This phenomenon is more conspicuous regarding the Indian Pines data set. For example, A-SPN obtains an OA of 93.31% for the Indian Pines data set with a patch size of 15×15, whereas SPN obtains an OA of 83.62%. Similar results can be observed for the other two data sets. Accordingly, with the attention mechanism, A-SPN is superior over SPN, and it is more stable to patch size.

B. Representative Analysis of A-SOP Operator

To uncover the representative power of the presented A-SPN method, we qualitatively analyze the feature maps obtained by the proposed A-SOP operator. As shown in Fig. 15, we present the origin image patches (two for each classes) with size of 27×27, the corresponding attention weight $\mathbf{W}$, and the obtained second-order features $\mathbf{F}_{\mathrm{A-SOP}}$ (rescaled to 27×27 for readability).

As we can see, the A-SOP operator focuses on the local salient and most representative features. By inspecting the images with a column hand of $\mathbf{W}$, A-SOP is capable of capturing the class-oriented information for different image patches,

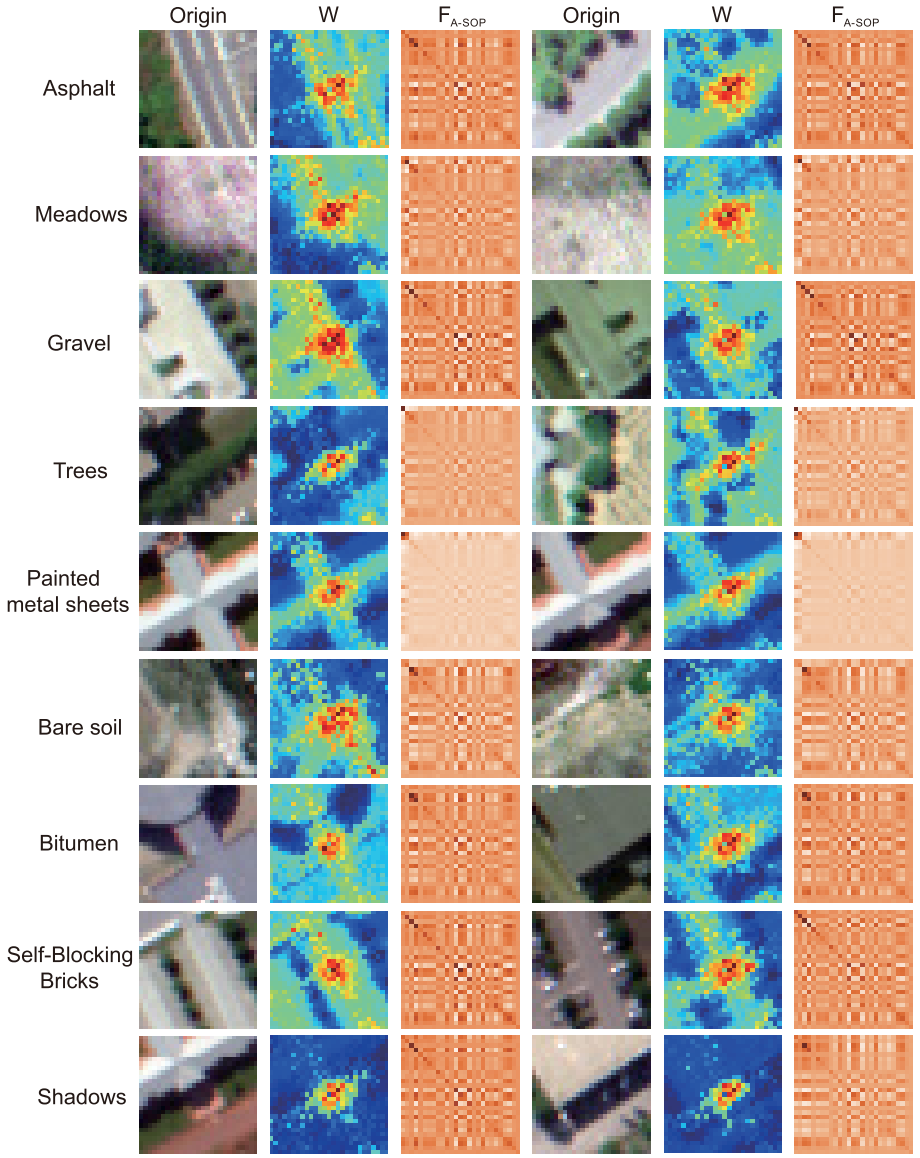

Fig. 15. Visualization of spatial attention weights and feature maps learned by A-SOP for the University of Pavia data set.

resulting in higher attention weights for the neighboring pixels belonging to the central pixel whereas lower attention weights for those pixels whose class labels are different from the central pixel. In this context, A-SOP models discriminative information between-class. By inspecting the images with a column hand of $\mathbf{F}_{\mathrm{A-SOP}}$, the obtained second-order features for each two image patches from the same class are quite similar, demonstrating good generalization capacity of A-SOP. In this context, A-SOP also considers the variations within-class.

C. Complexity Analysis

We analyze the complexity of the proposed method in terms of running time (s) and parameter size (MB). The computational complexity of A-SPN in the training phase can be formulated as

$$O\left(\mathrm{QCB}^2\mathrm{iteration} + \mathrm{BN}^2\mathrm{iteration}\right) \tag{6}$$

where Q is the batch size. Accordingly, the computational complexity of A-SPN in the test or prediction phase can be formulated as

$$O\left(\mathrm{SCB}^2 + \mathrm{SN}^2\right) \tag{7}$$

where S is the number of unseen sample.

As for memory burden, the parameter size of A-SPN can be formulated as

$$O\left(\mathrm{CB}^2 + N\right). \tag{8}$$

We only report the running time and parameter size obtained by different methods using randomly selected samples. As listed in Table X, compared with other DL methods, including SAE, LSTM, CNN, SSRN, and SSUN, A-SPN has a remarkable improvement on the training time. For instance, A-SPN only costs 4.57 s for the Indian Pines data set, which is more than seven times faster than 2-D CNN, 24 times faster than SSUN, nearly 100 times faster than SSRN, and

TABLE X
RUNNING TIME (S) AND PARAMETER SIZE (MB) OF DIFFERENT METHODS

Data set	Time	RBF-SVM	SAE	LSTM	2D CNN	3D CNN	SSRN	SSUN	A-SPN
Indian Pines	Training	0.27	466.09	52.71	29.42	4044.14	380.14	96.65	4.57
	Test	2.05	0.21	0.56	0.38	44.77	6.33	1.10	1.13
	Prediction	4.48	0.40	1.06	0.70	101.70	12.64	1.59	1.90
	Parameter size	-	0.09	0.12	0.11	4.09	0.35	1.22	0.64
University of Pavia	Training	0.15	1197.40	107.27	56.06	2337.34	669.80	195.63	7.91
	Test	4.66	1.43	2.29	1.49	60.21	12.80	3.52	4.33
	Prediction	24.06	6.01	10.39	7.21	306.77	57.09	15.80	18.47
	Parameter size	-	0.02	0.10	0.10	2.78	0.20	1.20	0.10
University of Houston	Training	0.05	337.95	38.72	20.87	687.32	333.19	72.92	5.22
	Test	1.10	0.22	0.95	0.56	17.03	7.26	1.50	1.66
	Prediction	52.02	7.79	33.63	23.02	789.89	287.71	49.71	36.84
	Parameter size	-	0.09	0.11	0.11	3.49	0.26	1.2	0.31

nearly 1000 times faster than 3-D CNN. As for parameter size, all methods have a tolerable parameter size.

Note that there are two reasons to explain the superior efficiency of A-SPN. First, A-SPN is a straightforward architecture rather than a traditional hierarchical structure composed of several interleaved convolution and pooling layers. Second, A-SPN has some normalization layers, which can effectively improve the convergency level.

VI. CONCLUSION

In this article, we propose a novel A-SPN for HSI classification. The innovative contribution of A-SPN lies in the design of the A-SOP operator, which is very powerful by using second-order statistics and data-adaptive attention weights to model discriminative and representative features. Experimental results conducted on three HSIs validated the superiority of A-SPN compared with traditional and other state-of-the-art DL-based HSI classification methods. Generally, A-SPN has fewer hyperparameters, generalizes better with limited training samples, provides higher OAs, converges better, and runs faster than the other counterparts.

Although our experimental results are encouraging, we will plan to expand the proposed method from three perspectives: 1) design spectral attention mechanism; 2) exploit advanced higher order statistical operators; and 3) investigate the class-imbalance problem.

ACKNOWLEDGMENT

The authors would like to thank Prof. D. Landgrebe for making the Airborne Visible/Infrared Imaging Spectrometer Indian Pines hyperspectral data set available to the community, Prof. P. Gamba for providing the Reflective Optics Spectrographic Imaging System data over Pavia, Italy, and the IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Technical Committee for providing the University of Houston data sets.

REFERENCES

[1] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Process. Mag.*, vol. 19, no. 1, pp. 17–28, Jan. 2002.

[2] Y. Gu, J. Chanussot, X. Jia, and J. A. Benediktsson, "Multiple kernel learning for hyperspectral image classification: A review," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 11, pp. 6547–6565, Nov. 2017.

[3] P. Ghamisi *et al.*, "Advances in hyperspectral image and signal processing a comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

[4] 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.

[5] P. Ghamisi *et al.*, "New frontiers in spectral–spatial hyperspectral image classification: The latest advances based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 3, pp. 10–43, Sep. 2018.

[6] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data a technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[7] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[8] N. Audebert, B. Le Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.

[9] 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.

[10] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 279–317, Dec. 2019.

[11] 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.

[12] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, Jul. 2015.

[13] Y. S. Chen, X. Zhao, and X. P. 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] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[17] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, Jun. 2015.

[18] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016.

[19] 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.

[20] P. Ghamisi, Y. Chen, and X. X. Zhu, "A self-improving convolution neural network for the classification of hyperspectral data," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 10, pp. 1537–1541, Oct. 2016.

[21] E. Aptoula, M. C. Ozdemir, and B. Yanikoglu, "Deep learning with attribute profiles for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1970–1974, Dec. 2016.

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

[23] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised spatial–spectral feature learning by 3D convolutional autoencoder for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, Sep. 2019.

[24] 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.

[25] W. Li, C. Chen, M. Zhang, H. Li, and Q. Du, "Data augmentation for hyperspectral image classification with deep CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 593–597, Apr. 2019.

[26] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2018.

[27] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, "Active learning with convolutional neural networks for hyperspectral image classification using a new Bayesian approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6440–6461, Nov. 2018.

[28] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4604–4616, Jul. 2020.

[29] S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, "Learning sensor-specific spatial-spectral features of hyperspectral images via convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4520–4533, Aug. 2017.

[30] J. X. Yang, Y. Q. Zhao, and J. C. W. Chan, "Learning and transferring deep joint spectral–spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017.

[31] X. Zhou and S. Prasad, "Deep feature alignment neural networks for domain adaptation of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5863–5872, Oct. 2018.

[32] L. Windrim, A. Melkumyan, R. J. Murphy, A. Chlingaryan, and R. Ramakrishnan, "Pretraining for hyperspectral convolutional neural network classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2798–2810, May 2018.

[33] X. He, Y. Chen, and P. Ghamisi, "Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3246–3263, May 2020.

[34] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, "Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5585–5599, Oct. 2017.

[35] 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.

[36] 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.

[37] 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.

[38] Z. Gong, P. Zhong, Y. Yu, W. Hu, and S. Li, "A CNN with multiscale convolution and diversified metric for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3599–3618, Jun. 2019.

[39] S. Wan, C. Gong, P. Zhong, B. Du, L. Zhang, and J. Yang, "Multiscale dynamic graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3162–3177, May 2020.

[40] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Free-form region description with second-order pooling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1177–1189, Jun. 2015.

[41] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1449–1457.

[42] N. He *et al.*, "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 755–769, Feb. 2019.

[43] Y. Sun, Z. Fu, and L. Fan, "A novel hyperspectral image classification pattern using random patches convolution and local covariance," *Remote Sens.*, vol. 11, no. 16, p. 1954, Aug. 2019.

[44] J. Zheng, Y. Feng, C. Bai, and J. Zhang, "Hyperspectral image classification using mixed convolutions and covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 522–534, Jan. 2020.

[45] 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.

[46] 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.

[47] Z. Zheng, Y. F. Zhong, A. L. Ma, and L. P. Zhang, "FPGA: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5612–5626, Aug. 2020.

[48] J. Wang, J. Zhou, and W. Huang, "Attend in bands: Hyperspectral band weighting and selection for image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 4712–4727, Dec. 2019.

[49] 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.

[50] B. Fang, Y. Li, H. K. Zhang, and J. C. W. Chan, "Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism," *Remote Sens.*, vol. 11, no. 2, p. 18, Jan. 2019.

[51] P. D. Wu, Z. G. Cui, Z. L. Gan, and F. Liu, "Residual group channel and space attention network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 12, p. 27, Jun. 2020.

[52] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3024–3033.

[53] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11065–11074.

[54] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 143–156.

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

[56] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Jan. 1995.

[57] Z. L. Zhong, J. Li, Z. M. 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.

[58] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 464–472.

Zhaohui Xue (Member, IEEE) received the B.S. degree in geomatics engineering from Shandong Agricultural University, Tai'an, China, in 2009, the M.E. degree in remote sensing from the China University of Mining and Technology, Beijing, China, in 2012, and the Ph.D. degree in cartography and geographic information system from Nanjing University, Nanjing, China, in 2015.

He is a Youth Professor with the School of Earth Sciences and Engineering, Hohai University, Nanjing. His research interests include hyperspectral image classification, time-series image analysis, pattern recognition, and machine learning.

Dr. Xue has been honored as an Outstanding Graduate for the B.S., M.E., and Ph.D. degrees in 2009, 2012, and 2015, respectively. He was a recipient of the National Scholarship for Doctoral Graduate Students granted by the Ministry of Education of the People's Republic of China in 2014. He received the Best Reviewer for the IEEE GEOSCIENCE AND REMOTE SENSING SOCIETY. He is an Editorial Board Member in *Journal of Remote Sensing* from 2020 to 2024. He has been a Reviewer for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, *ISPRS Journal of Photogrammetry and Remote Sensing*, *International Journal of Remote Sensing*, etc.

Mengxue Zhang received the B.S. degree in geomatics engineering from Hohai University, Nanjing, China, in 2018, where he is pursuing the M.E. degree in geographical information engineering.

His research interests include hyperspectral image classification, radar automatic target recognition, and machine learning.

Yifeng Liu received the Ph.D. degree in electronic engineering from Wuhan University, Wuhan, China, in 2016.

He is a Senior Engineer with the China Academy of Electronics and Information Technology, Beijing, China, and the Deputy Director of the National Engineering Laboratory for Risk Perception and Prevention (RPP), Beijing. He is a selected candidate of the 5th Youth Talent Promotion Project of China Association for Science and Technology. The artificial intelligence video analysis system that he led the team to develop, has attracted more than 100 million RMB funds, which has been reported by the *People's Daily*. He has over 20 publications primarily in cyberspace and data science. His research interests include around computer vision, machine learning, and knowledge engineering.

Dr. Liu has won the first prize of Shijingshan District Science and Technology Award in 2016.

Peijun Du (Senior Member, IEEE) received the Ph.D. degree in geodesy and survey engineering from the China University of Mining and Technology, Xuzhou, China, in 2001.

He is a Professor of remote sensing and geographical information science with Nanjing University, Nanjing, China. He has authored over 70 articles in international peer-reviewed journals and over 100 articles in international conferences and Chinese journals. His research interests include remote sensing image processing and pattern recognition, hyperspectral remote sensing, and applications of geospatial information technologies.

Dr. Du is an Associate Editor of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He also served as the Co-Chair of the Technical Committee of URBAN in 2009, the International Association of Pattern Recognition Workshop on Pattern Recognition in Remote Sensing (IAPR-PRRS) in 2012, the International Workshop on Earth Observation and Remote Sensing Applications (EORSA) in 2014, the Co-Chair of the Local Organizing Committee of Joint Urban Remote Sensing Event (JURSE) in 2009, the Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS) in 2012, and EORSA in 2012, and a member of the Scientific Committee or the Technical Committee of other international conferences, including WHISPERS from 2010 to 2016, URBAN in 2011, 2013, and 2015, MultiTemp in 2011, 2013, and 2015, the International Symposium on Image and Data Fusion (ISIDF) in 2011, and the International Society for Optics and Photonics (SPIE) European Conference on Image and Signal Processing for Remote Sensing from 2012 to 2016.