Learning to Pay Attention on Spectral Domain: A Spectral Attention Module-Based Convolutional Network for Hyperspectral Image Classification

Lichao Mou, *Student Member, IEEE*, and Xiao Xiang Zhu, *Senior Member, IEEE*

***Abstract*— Over the past few years, hyperspectral image classification using convolutional neural networks (CNNs) has progressed significantly. In spite of their effectiveness, given that hyperspectral images are of high dimensionality, CNNs can be hindered by their modeling of all spectral bands with the same weight, as probably not all bands are equally informative and predictive. Moreover, the usage of useless spectral bands in CNNs may even introduce noises and weaken the performance of networks. For the sake of boosting the representational capacity of CNNs for spectral-spatial hyperspectral data classification, in this work, we improve networks by discriminating the significance of different spectral bands. We design a network unit, which is termed as the spectral attention module, that makes use of a gating mechanism to adaptively recalibrate spectral bands by selectively emphasizing informative bands and suppressing less useful ones. We theoretically analyze and discuss why such a spectral attention module helps in a CNN for hyperspectral image classification. We demonstrate using extensive experiments that in comparison with state-of-the-art approaches, the spectral attention module-based convolutional networks are able to offer competitive results. Furthermore, this work sheds light on how a CNN interacts with spectral bands for the purpose of classification.**

***Index Terms*— Attention module, convolutional neural network (CNN), gating mechanism, hyperspectral image classification.**

I. Introduction

HYPERSPECTRAL images encompass hundreds of continuous observation spectral bands, which are capable of precisely differentiating various materials of interest. Hence, in the remote sensing community, hyperspectral images have already been considered a vital data source for object identification and classification tasks.

Consequently, numerous kinds of classification approaches, especially supervised models have been developed for hyperspectral data classification, as found in the literature. Among them, random forest [1]–[3] and support vector machine (SVM) [4]–[8] are two examples of supervised classification approaches, which have been exploited for solving varied and numerous classification problems. Random forests are basically a kind of ensemble bagging or averaging algorithm. It creates a set of decision trees using random subsamples of training data and then aggregates their predictions via a maximum a posterior (MAP) rule or voting to decide final classes of test samples. On the other hand, an SVM seeks for a hyperplane that is able to sort two-class data by the largest margin. However, the random forest and SVM are characterized as "shallow" models [9] as compared to deep networks which are able to extract hierarchical, deep feature representations.

Deep learning, which is mainly characterized by deep networks, has been quite successful in solving a wide range of problems (e.g., natural language processing [10]–[12], computer vision [13]–[25], and remote sensing [26]). In the hyperspectral community, some studies have been published recently on the use of convolutional neural networks (CNNs) [27]–[42] as well as recurrent neural networks (RNNs) [43]–[49] for pattern recognition tasks. For instance, Kussul *et al.* [27] addressed the classification problem of crop types by making use of 1-D and 2-D CNNs and found that the 2-D CNN is superior to the 1-D CNN, but several tiny objects in the classification map of the 2-D CNN are a little oversmoothed and misclassified. In [28], Song *et al.* studied feature fusion in a residual learning-based 2-D CNN, aiming to build a more discriminative network for hyperspectral data classification tasks. Following the recent developments in 3-D CNN for video analysis [50], where the third dimensionality is usually the time axis, 3-D CNNs have also been studied in hyperspectral data classification. Chen *et al.* [29] introduced a ℓ_2 regularized 3-D CNN for learning spectral-spatial features, while [30] followed a similar idea for the purpose of classification. Paoletti *et al.* [51] introduced an improved 3-D CNN consisting of 5 layers which make use of all the spatial-spectral information on the hyperspectral image.

To avoid overfitting, Zhao and Du [32] jointly used a dimension reduction method and a 2-D CNN for spectral-spatial

Manuscript received December 27, 2018; revised May 5, 2019; accepted June 23, 2019. Date of publication September 27, 2019; date of current version December 27, 2019. This work was supported in part by the German Research Foundation (DFG) under Grant ZH 498/7-2, in part by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (Acronym: *So2Sat*, www.so2sat.eu) under Grant [ERC-2016-StG-714087], and in part by the Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (www.sipeo.bgu.tum.de) under Grant VH-NG-1018. *(Corresponding author: Xiao Xiang Zhu.)*

The authors are with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany, and also with the Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: lichao.mou@dlr.de; xiaoxiang.zhu@dlr.de).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2019.2933609

feature extraction. Ghamisi *et al.* [33] first exploited a computational intelligence (particle swarm optimization) method to choose informative spectral bands and then train a 2-D CNN using the selected bands. In [34], to properly train a CNN with limited ground truth data, the authors devised a pixel-pair CNN that takes as input a pair of hyperspectral pixels. By doing so, the amount of training data is greatly augmented. Furthermore, in order to access a huge amount of unlabeled hyperspectral data, unsupervised feature learning via a CNN is of great interest. Romero *et al.* [35] presented a CNN to address the problem of unsupervised spectral-spatial feature extraction and estimated network weights via a sparse learning approach in a greedy layer-wise fashion. Mou *et al.* [37] proposed a residual learning-based fully conv-deconv network, aiming at unsupervised spectral-spatial feature learning in an end-to-end manner. Better classification network architecture from computer vision (e.g., ResNet [17], DenseNet [18], and CapsuleNet [52]) also provides new insights into hyperspectral image classification [37]–[39], [53]. Moreover, the integration of networks and other traditional machine learning models, e.g., conditional random field (CRF) and active learning, has also received attention recently [54], [55].

The unique asset of hyperspectral images is their rich spectral content in comparison with high-resolution aerial images and natural images in the computer vision field. Although there already exist a number of works that have focused on using CNNs for hyperspectral data classification, we notice that in the community, the following questions have not been well addressed until now.

1) Do all spectral bands contribute equally to *a CNN* for classification tasks?
2) If no, how to *task-drivenly* find informative bands that can help hyperspectral data classification *in an end-to-end network*?
3) Is it possible to improve classification results of a CNN by emphasizing informative bands and suppressing less useful ones in the network?

These questions give us an incentive to devise a novel network called spectral attention module-based convolutional network for hyperspectral image classification. Inspired by recent advances in the attention mechanism of networks [56]–[58], which enables feature interactions to contribute differently to predictions, we design a channel attention mechanism for analyzing the significance of different spectral bands and recalibrating them. More importantly, the significance analysis is automatically learned from tasks and hyperspectral data in an end-to-end network without any human domain knowledge. Experiments show that the use of the proposed spectral attention module in a CNN for hyperspectral data classification serves two benefits: it not only offers better performance but also provides an insight into which spectral bands contribute more to predictions. This work's contributions are threefold.

1) We propose a learnable spectral attention module that explicitly allows the spectral manipulation of hyperspectral data within a CNN. This attention module exploits the global spectral-spatial context for producing a series of spectral gates which reflects the significance of spectral bands. The recalibrated spectral information using these spectral gates can effectively improve the classification results.
2) We analyze and discuss why the proposed spectral attention module is able to offer better classification results from a theoretical perspective by diving into the backward propagation of the network. As far as we know, learning and analyzing such a spectral attention-based network for hyperspectral image classification have not been done yet.
3) We conduct experiments on four benchmark data sets. The empirical results demonstrate that our spectral attention module-based convolutional network is capable of offering competitive classification results, particularly in the situation of high dimensionality and inadequate training data.

The remainder of this article is organized as follows. After detailing hyperspectral image classification using CNNs in Section I, Section II introduces the proposed spectral attention module-based convolutional network. Section III verifies the proposed approach and presents the corresponding analysis and discussion. Finally, Section IV concludes the article.

II. Methodology

A. Problem Formulation

The spectral attention module in our model transforms a patch $\boldsymbol{x}$ of a hyperspectral image into a new representation $\boldsymbol{z}$ via the following mapping:

$$\boldsymbol{F} : \boldsymbol{x} \rightarrow \boldsymbol{z} \tag{1}$$

where $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^{H \times W \times C}$.

Our aim is to strengthen the representational capacity of a spectral-spatial classification network through explicitly modeling the significance of spectral bands. Therefore, we instantiate $\boldsymbol{F}$ as

$$\boldsymbol{z} = \boldsymbol{x} \odot \boldsymbol{g} \tag{2}$$

where $\odot$ is a channel-wise multiplication operation and $\boldsymbol{g} \in \mathbb{R}^{C}$ represents a set of *spectral gates* applied to individual spectral bands of the patch $\boldsymbol{x}$.

The motivation behind (2) is that we wish to make use of a gating mechanism to recalibrate the strength of different spectral bands of the input, i.e., selectively emphasize useful bands and suppress less informative ones, for image classification problems.

Fig. 1 illustrates the architecture of the spectral attention module-equipped convolutional network.

B. Modeling of Spectral Attention Module

The gating mechanism has been widely used in modeling and processing temporal sequences. For example, long short-term memory (LSTM)-based networks [59], [60] harness three gates to cope with vanishing gradients. Similarly, a gated recurrent unit (GRU) [61], [62] is designed to implement the modulation of information flow through the gating mechanism.

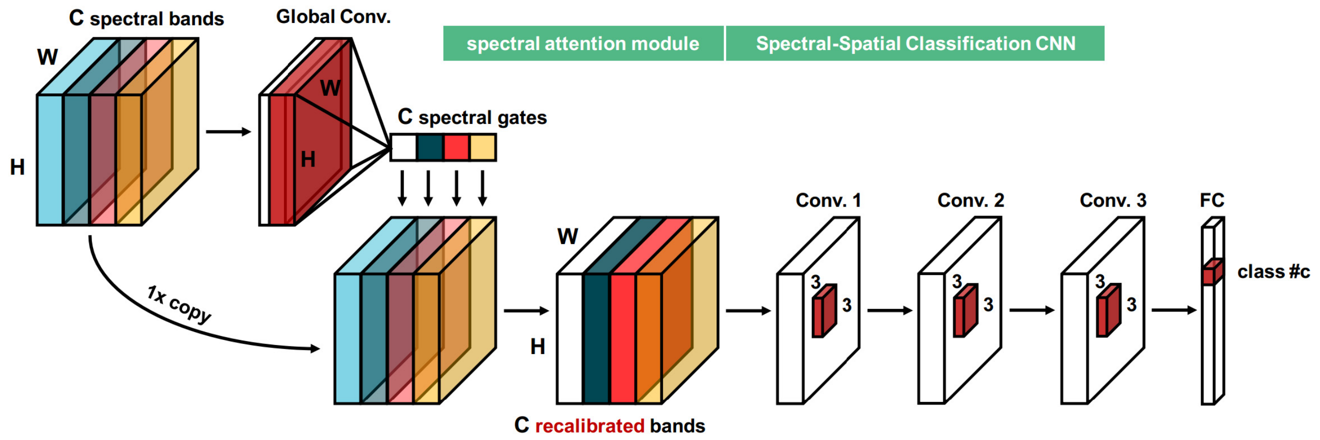


Fig. 1. The overall architecture of the proposed gating mechanism, spectral attention module, for hyperspectral classification problems. We would like to exploit this module to learn and recalibrate strengths of different spectral bands, i.e., selectively emphasize useful bands and suppress less informative ones, for image classification problems. To this end, we first learn a set of spectral gates by using global convolution and then apply them to individual spectral bands. Moreover, in Section II-C, we theoretically analyze and discuss why the proposed spectral attention module can help a spectral-spatial classification network (e.g., a 2-D CNN) for hyperspectral image classification tasks.

In addition, several recent works in computer vision have shown the benefit of introducing the gating mechanism to vision problems. To name a few, Wang *et al.* [56] proposed a gating mechanism that is capable of dynamically balancing contributions of the current event and its surrounding contexts in their model for dense video captioning tasks. Hu *et al.* [58] built a gated block for image classification tasks and demonstrated its good performance on large-scale image recognition. Liu *et al.* [57] addressed person re-identification tasks through utilizing a network module based on a soft gating mechanism, which enables the network to concentrate on significant local regions of an input image pair adaptively. In remote sensing, a very recently published, parallel work related to this article can be found in [63], where the authors introduced a visual attention technique that first calculates a mask and then applies it to features produced by a ResNet for hyperspectral data classification tasks.

Here, we would like to design our own gating mechanism, spectral attention module, for analyzing the significance of different spectral bands and recalibrating them. Besides, we hope this module is task-driven and can be adaptively learned in an end-to-end spectral-spatial classification network. To this end, we need a way to aggregate the spectral-spatial information of $\boldsymbol{x}$ across the spatial domain to produce a collection of spectral gates $\boldsymbol{g}$.

The convolution operation is an ideal candidate, as 1) it is able to spatially shrink the input patch and 2) its differential property allows end-to-end learning. In general, a convolutional filter operates with a local receptive field (e.g., 3 × 3 in VGG-16 network), which leads to the fact that the output is not capable of utilizing contextual information outside of this region. This is a severe issue for our case because the spectral gates $\boldsymbol{g}$ in our model are expected to be derived from the whole spectral-spatial information. To tackle this problem, we distill global spatial information into the spectral gates by using global convolution. Formally, let $\boldsymbol{f} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_C]$ denote a set of convolutional filters and their sizes are both $H \times W$, where $\boldsymbol{f}_c$ refers to the c-th filter. Thus, the c-th spectral gate g_c can be calculated as follows:

$$g_c = \boldsymbol{x} * \boldsymbol{f}_c = \sum_{i=1}^{C} \boldsymbol{x}_i * \boldsymbol{f}_c^i \tag{3}$$

where $*$ represents convolution and $\boldsymbol{f}_c^i$ and $\boldsymbol{x}_i$ are separately the i-th channels of the c-th filter and $\boldsymbol{x}$. Taking into account that the field of view of global convolution is equal to the spatial size of $\boldsymbol{x}$, g_c is actually calculated by the inner product of $\boldsymbol{x}_i$ and $\boldsymbol{f}_c^i$ (both $\boldsymbol{x}_i$ and $\boldsymbol{f}_c^i$ are vectorized into columns), i.e., (3) can be rewritten as follows:

$$g_c = \sum_{i=1}^{C} \langle \boldsymbol{x}_i, \boldsymbol{f}_c^i \rangle = \sum_{i=1}^{C} \boldsymbol{x}_i^T \boldsymbol{f}_c^i. \tag{4}$$

From (4), the spectral gates $\boldsymbol{g}$ can be considered as a series of global descriptors, which are capable of representing spectral-spatial features of $\boldsymbol{x}$.

Thus, according to (2), we can associate the c-th spectral gate g_c with the c-th spectral band of $\boldsymbol{x}$ to obtain the recalibrated z_c via

$$z_c = \boldsymbol{x}_c \sum_{i=1}^{C} \boldsymbol{x}_i^T \boldsymbol{f}_c^i. \tag{5}$$

So far, we can obtain an initial spectral attention module [as shown in (5)], but there still exist three issues which we should address:

1) Given the complex spectral-spatial properties of hyperspectral images, we wish that the spectral gates in this module are capable of learning a nonlinear mapping, instead of a linear one, from the input.
2) The attention module should model a nonmutually exclusive relationship between spectral bands, as we would like to ensure that multiple bands can be emphasized at the same time (unlike one-hot activation in softmax).

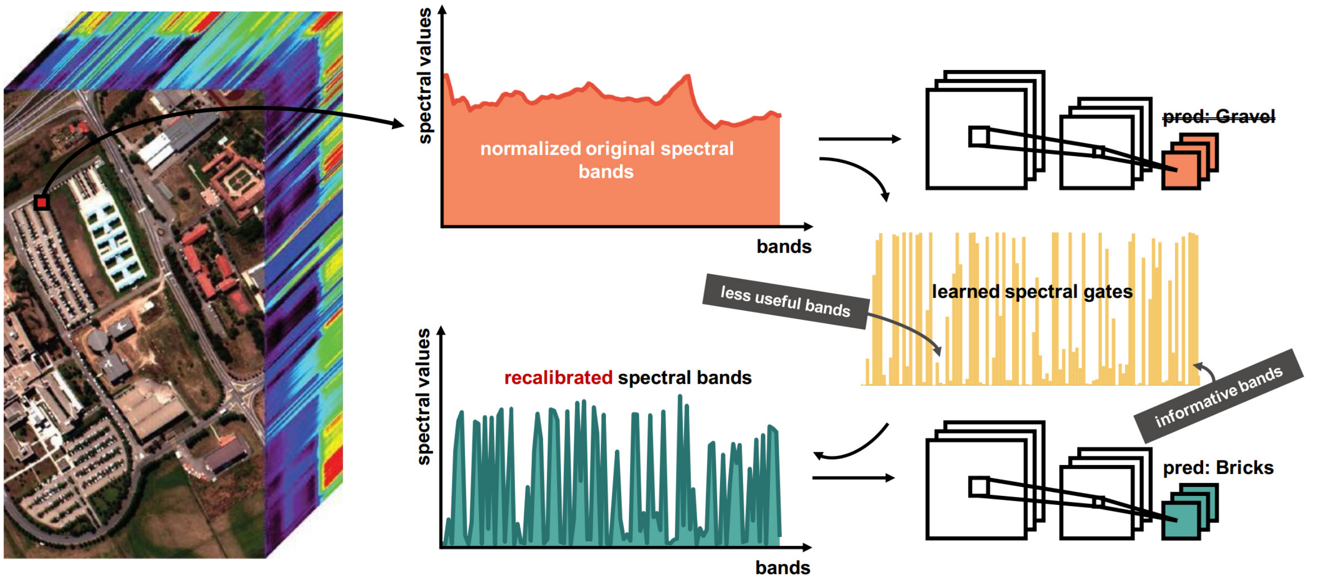


Fig. 2. Example showing how the proposed spectral attention module in a CNN correct a wrong prediction (gravel) to a right one (bricks) via learned spectral gates.

3) The gates should be bounded (e.g., between 0 and 1), easily differentiable, and monotonic (good for convex optimization).

To meet these three requirements, we modify spectral gates in the initial spectral attention module as follows:

$$g_c = \frac{1}{1+\exp(-\boldsymbol{x} * \boldsymbol{f}_c)} = \frac{1}{1+\exp\left(-\sum_{i=1}^{C} \boldsymbol{x}_i^T \boldsymbol{f}_c^i\right)}. \quad (6)$$

Hence, the final version of the spectral attention module can be written as

$$z_c = \boldsymbol{x}_c \frac{1}{1+\exp\left(-\sum_{i=1}^{C} \boldsymbol{x}_i^T \boldsymbol{f}_c^i\right)}. \quad (7)$$

Fig. 2 is an example showing how the proposed spectral attention module works in a CNN.

C. Why Does the Spectral Attention Module Work?

In our experiments, we observed that a 2-D CNN with our spectral attention module can offer better classification results. However, how exactly does this attention module help a spectral-spatial classification network for hyperspectral data classification? We dive into the backward propagation of the network to seek the answer to this question.

For notional simplicity, we subsequently drop the subscript c and rewrite the final expression of the spectral attention module as follows:

$$z = \boldsymbol{x} \odot \frac{1}{1+\exp(-\boldsymbol{x} * \boldsymbol{f})}. \quad (8)$$

Thus, the gradient of the spectral attention module can be written as

$$\nabla z = \nabla \boldsymbol{x} \odot \frac{1}{1+\exp(-\boldsymbol{x} * \boldsymbol{f})} + \boldsymbol{x} \odot \nabla \left(\frac{1}{1+\exp(-\boldsymbol{x} * \boldsymbol{f})} \right). \quad (9)$$

It can be seen that the term $\nabla \boldsymbol{x}$ is weighted by the spectral gates $(1/1+\exp(-\boldsymbol{x} * \boldsymbol{f}))$. This has the following interesting properties.

1) On the one hand, the existence of $\nabla \boldsymbol{x}$ ensures that the gradient information on spectral-spatial features can be backpropagated directly, which helps to prevent the vanishing gradient problem.
2) On the other hand, for spectral bands where the spectral gates are close to 0 (less useful bands), the gradient propagation vanished; on the contrary, for values that are close to 1, gradients (of informative bands) directly propagated from $\boldsymbol{z}$ to $\boldsymbol{x}$.

For the first point, a similar effect can be found in residual learning. He *et al.* [17] introduced the residual learning into CNNs for large-scale image classification tasks and exhibited significantly improved network training characteristics, e.g., allowing network depths that were previously unattainable. Formally, denote by $\boldsymbol{y}$ a random variable representing the output of a residual block. It can then be expressed as

$$\boldsymbol{y} = \boldsymbol{x} + \mathcal{F}(\boldsymbol{x}; \boldsymbol{w}) \quad (10)$$

where $\mathcal{F}$ is a residual function and usually implemented by a couple of stacked convolutional layers. Moreover, $\boldsymbol{w}$ represents learnable weights of this residual block. The gradient of a residual block can be calculated as

$$\nabla \boldsymbol{y} = \nabla \boldsymbol{x} + \nabla(\mathcal{F}(\boldsymbol{x}; \boldsymbol{w})). \quad (11)$$

TABLE I
CONFIGURATION OF A SPECTRAL ATTENTION MODULE-BASED CONVOLUTIONAL NETWORK FOR THE PAVIA UNIVERSITY DATA SET

Layer	Input Shape	Output Shape	#Filters	Connected to	Configuration
spec. attn. module	(16, 16, 103)	(16, 16, 103)	103	input	16×16 kernel
conv1-1	$(16, 16, 103)$	$(16, 16, 32)$	32	spec. attn. module	3×3 kernel, stride 1, padding 1, bn, relu
conv1-2	$(16, 16, 32)$	$(16, 16, 32)$	32	conv1-1	3×3 kernel, stride 1, padding 1, bn, relu
maxpool1	$(16, 16, 32)$	$(8, 8, 32)$	-	conv1-2	pool size 2×2, stride 2
conv2-1	$(8, 8, 32)$	$(8, 8, 64)$	64	maxpool1	3×3 kernel, stride 1, padding 1, bn, relu
conv2-2	$(8, 8, 64)$	$(8, 8, 64)$	64	conv2-1	3×3 kernel, stride 1, padding 1, bn, relu
maxpool2	$(8, 8, 64)$	$(4, 4, 64)$	-	conv2-2	pool size 2×2, stride 2
conv3-1	$(4, 4, 64)$	$(4, 4, 128)$	128	maxpool2	3×3 kernel, stride 1, padding 1, bn, relu
conv3-2	$(4, 4, 128)$	$(4, 4, 128)$	128	conv3-1	3×3 kernel, stride 1, padding 1, bn, relu
maxpool3	$(4, 4, 128)$	$(2, 2, 128)$	-	conv3-2	pool size 2×2, stride 2
gap	$(2, 2, 128)$	$(1, 1, 128)$	-	maxpool3	pool size 2×2
fc1	$(1, 1, 128)$	$(1024,)$	-	gap	1024 units, relu
fc2	$(1024,)$	$(9,)$	-	fc1	9 units, softmax

From (11), we can see that $\nabla \boldsymbol{y}$ is a sum of the gradient of the input $\nabla \boldsymbol{x}$ and the gradient $\nabla(\mathcal{F}(\boldsymbol{x}; \boldsymbol{w}))$, and as mentioned above, the term $\nabla \boldsymbol{x}$ is a key to avoiding the vanishing gradient problem. This is the same for the first property of our spectral attention module.

Instead of $\nabla \boldsymbol{x}$ in (9), $\nabla \boldsymbol{x}$ in (11) is not weighted – in other words, gradients of all spectral bands are indiscriminately backpropagated; in contrast, the spectral attention module has a selection mechanism regarding the significance of different spectral bands from the perspective of gradient.

D. Network Training

We insert the spectral attention module into a 2-D CNN (between the input and the first convolutional layer) and then train the whole network. Note that the spectral attention module and other layers are trained simultaneously. We use the TensorFlow framework to implement and train networks. All network weights are initialized by a Glorot uniform initializer [64]. The Nesterov Adam [65] algorithm is chosen to optimize networks, as for our experiments, compared to stochastic gradient descent (SGD) with momentum [66] or Adam [67], it is able to provide much faster convergence. Almost all parameters of this optimizer are set as recommended in [65]. We utilize a relatively small learning rate of $2e{-}04$. Finally, we train networks on an NVIDIA Tesla P100 16 GB GPU. Table I exhibits an example of a CNN with the proposed attention module.

III. EXPERIMENTS AND ANALYSIS

A. Data Description

1) Indian Pines Hyperspectral Data Set: The first data were collected by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor over Northwest Indiana, USA, 1992. It includes 145 × 145 pixels with a 20 m/pixel spatial resolution and 200 spectral bands covering from 400 to 2500 nm after removing 20 water absorption channels (220, 150-163, and 104-108). The ground truth includes 16 classes of interest, which are mostly various crops in different growth phases and are detailed in Table II. Since these 16 classes have similar spectral signatures, the precise classification of this scene is hard. The true-color composite image and the available ground truth data can be found in Fig. 3 (black color in the ground truth indicates unknown samples).

TABLE II
AMOUNTS OF TRAINING AND TEST DATA ON THE INDIAN PINES SCENE

Class No.	Class Name	Training	Test
1	Corn-notill	50	1384
2	Corn-min	50	784
3	Corn	50	184
4	Grass-pasture	50	447
5	Grass-trees	50	697
6	Hay-windrowed	50	439
7	Soybean-notill	50	918
8	Soybean-mintill	50	2418
9	Soybean-clean	50	564
10	Wheat	50	162
11	Woods	50	1244
12	Buildings-grass-trees	50	330
13	Stone-steel-towers	50	45
14	Alfalfa	15	39
15	Grass-pasture-mowed	15	11
16	Oats	15	5
	TOTAL	695	9671

2) Pavia University Hyperspectral Data Set: The second data set was acquired over the city of Pavia, Italy, 2002 by an airborne instrument – Reflective Optics Spectrographic Imaging System (ROSIS). The aircraft was operated by the German Aerospace Center (DLR) within the context of European Union funded HySens project. The data set is made up of 640 × 340 pixels with a 1.3 m/pixel spatial resolution and 103 bands covering from 430 to 860 nm after removing 12 noisy channels. Besides unknown pixels, 9 classes are manually annotated in the reference data. Fig. 4 displays a composite image of this data set and its reference map. Table III offers information on all 9 categories.

3) Salinas Hyperspectral Data Set: The third data set was also gathered by the AVIRIS sensor over the region of Salinas Valley, CA, USA and with a 3.7-m/pixel spatial resolution.

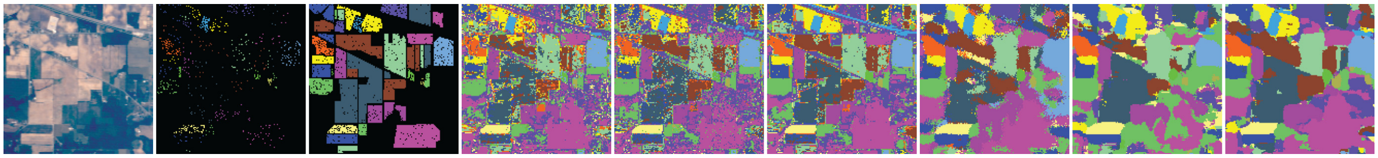


Fig. 3. Classification maps of different approaches for the Indian Pines data set. (Left to right) True-color composite image, training set, test set, RF-200, SVM-RBF, CCF-200, SICNN, 2-D CNN, and SpecAttenNet. Best zoomed-in view.

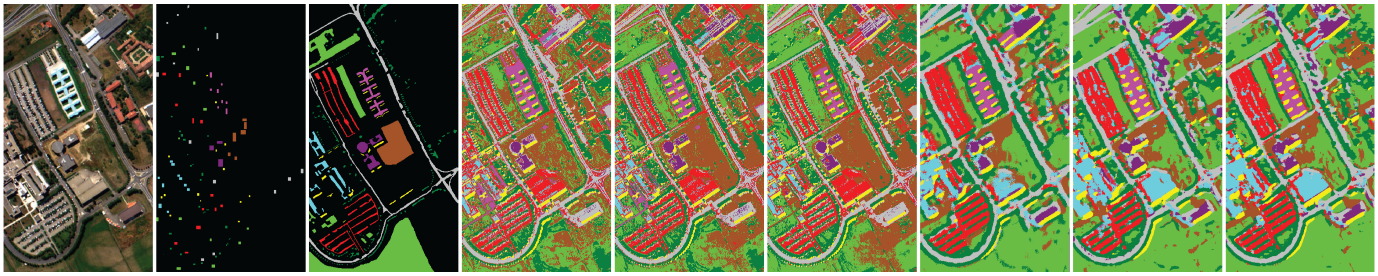


Fig. 4. Classification maps of different approaches for the Pavia University data set. (Left to right) Composite image, training samples, ground truth, RF-200, SVM-RBF, CCF-200, SICNN, 2-D CNN, and SpecAttenNet. Best zoomed-in view.

TABLE III

Amounts of Training and Test Data on the Pavia University Data Set

Class No.	Class Name	Training	Test
1	Asphalt	548	6631
2	Meadows	540	18649
3	Gravel	392	2099
4	Trees	524	3064
5	Metal sheets	265	1345
6	Bare Soil	532	5029
7	Bitumen	375	1330
8	Bricks	514	3682
9	Shadows	231	947
	TOTAL	3921	42776

TABLE IV

Amounts of Training and Test Data on the Salinas Data

Class No.	Class Name	Training	Test
1	Brocoli_green_weeds_1	50	1959
2	Brocoli_green_weeds_2	50	3676
3	Fallow	50	1926
4	Fallow_rough_plow	50	1344
5	Fallow_smooth	50	2628
6	Stubble	50	3909
7	Celery	50	3529
8	Grapes_untrained	50	11221
9	Soil_vinyard_develop	50	6153
10	Corn_senesced_green_weeds	50	3228
11	Lettuce_romaine_4wk	50	1018
12	Lettuce_romaine_5wk	50	1877
13	Lettuce_romaine_6wk	50	866
14	Lettuce_romaine_7wk	50	1020
15	Vinyard_untrained	50	7218
16	Vinyard_vertical_trellis	50	1757
	TOTAL	800	53329

The Salinas scene is composed of 224 spectral bands and 512 × 217 pixels. Like the Indian Pines data set, 20 water absorption bands (224, 154-167, and 108-112) of the Salinas scene have been discarded. The data set presents 16 classes related to vegetables, vineyard fields, and bare soils. Table IV shows the amounts of training and test data on this data set.

4) Houston Hyperspectral Data Set: The fourth data set was acquired over the University of Houston campus and its neighboring urban area. It was collected with an ITRES-CASI 1500 sensor on June 23, 2012 between 17:37:10 and 17:39:50 UTC. The average altitude of the sensor was about 1676 m, which results in 2.5-m spatial resolution data consisting of 349 by 1905 pixels. The hyperspectral imagery consists of 144 spectral bands ranging from 380 to 1050 nm and was processed (radiometric correction, attitude processing, GPS processing, geo-correction, and so on) to yield the final geo-corrected image cube representing the sensor spectral radiance. Table V provides information about all 15 classes of this data set with their corresponding training and test samples. This data set was kindly made available by the Image Analysis and Data Fusion Technical Committee of IEEE GRSS in 2012.

B. Experiment Setup

To quantitatively compare different models for hyperspectral data classification tasks from various aspects, the following measurements are considered.

1) *Overall Accuracy (OA):* This criterion is calculated as the fraction of test samples that are differentiated correctly.
2) *Per-Class Accuracy:* To assess the performance with respect to each category in a data set, we also compute per-class accuracy. This measurement is particularly useful when class labels are not uniformly distributed.
3) *Average Accuracy (AA):* This criterion is computed as the average of all per-class accuracies.
4) *Kappa Coefficient:* This statistic criterion is a robustness measurement with the degree of agreement.

Furthermore, we make use of a statistical test to validate the significance of classification accuracies produced by

TABLE V
AMOUNTS OF TRAINING AND TEST DATA ON THE HOUSTON DATA SET

Class No.	Class Name	Training	Test
1	Grass Healthy	198	1053
2	Grass Stressed	190	1064
3	Grass Synthetic	192	505
4	Tree	188	1056
5	Soil	186	1056
6	Water	182	143
7	Residential	196	1072
8	Commercial	191	1053
9	Road	193	1059
10	Highway	191	1036
11	Railway	181	1054
12	Parking Lot 1	192	1041
13	Parking Lot 2	184	285
14	Tennis Court	181	247
15	Running Track	187	473
	TOTAL	2832	12197

various methods. Given that samples used for two classification models are not independent, McNemar's test can be harnessed to estimate the significance of the difference in two classification maps, and the McNemar's test can be performed by

$$z_{12} = \frac{f_{12} - f_{21}}{\sqrt{f_{12} + f_{21}}} \tag{12}$$

where f_{ij} is the amount of data correctly recognized by method i and incorrectly recognized by j. McNemar's test is a statistical test for paired nominal data, and we can use McNemar's test to compare the predicted accuracies of two models. In McNemar's test, the null hypothesis, which means none of the two models performs better than the other, is rejected at $p = 0.05$ ($|z| > 1.96$), which indicates the significance level.

Below are methods included in our comparison.

1) *RF-200:* A random forest composed of 200 decision trees.
2) *SVM-RBF:* An SVM[1] having the widely used radial basis function (RBF) kernel. We make use of five-fold cross validation to search optimal hyper-parameters γ (spread of the RBF kernel) and C (controlling the magnitude of penalization during the model optimization) in the range of $\gamma = 2^{-3}, 2^{-2}, \cdots, 2^4$ and $C = 10^{-2}, 10^{-1}, \cdots, 10^4$.
3) *CCF-200:* A canonical correlation forest (CCF)[2] [68], [69] with 200 trees.
4) *SICNN:* A CNN model, which makes an attempt at solving the curse of dimensionality by first utilizing a computational intelligence (particle swarm optimization) algorithm to choose informative spectral bands and then training a 2-D CNN using the selected bands. The used network is made up of three convolutional layers. The first two convolutional layers are followed by max-pooling layers and their fields of view are 4×4 and 5×5, respectively. The last convolutional layer is equipped with 4×4 filters. Moreover, 32, 64, and 128 convolutional filters are used separately for those three convolutional layers. For more details, refer to [33].
5) *2-D CNN:* To demonstrate the superiority of the proposed method, we perform an ablation study, i.e., designing a 2-D CNN which has no spectral attention module, but other parts are the same as the proposed network (cf. Table I). The exact architecture of the 2-D CNN is a VGG-like network, in which we utilize three convolutional blocks and 3×3 filters for all the blocks. Spatial shrinkage is operated by three max-pooling layers following the convolutional blocks. Each convolutional block in this 2-D CNN has two convolutional layers, and 32, 64, and 128 filters are used for convolutional layers of those three blocks, respectively. Overall, we keep the architecture of 2-D CNN and that of the following network consistent.
6) *SpecAttenNet:* The proposed spectral attention module-based convolutional network (cf. Table I).

Note that, in order to make our model completely comparable with other investigated approaches, we use standard training and test sets for the Indian Pines, Pavia University, and Houston data sets. For the Salinas scene, training samples are generated by a simple random sampling. In both hyperspectral data sets, 10% samples of the training set are randomly selected as validation samples. In other words, in the network training phase, we use 90% samples of the training set to iteratively update and optimize network weights and the remaining ones as validation to tune hyperparameters of networks. Prior to training, we normalize each channel of the hyperspectral data to the range between 0 and 1. In addition, network architecture for these data sets is the same.

C. Ablation Study

To validate the effectiveness of the proposed module, we perform ablation experiments. As we have mentioned above, the competitor 2-D CNN is a network that has no spectral attention module, but other parts are the same as the proposed SpecAttenNet. From Tables VI–IX, we can see that SpecAttenNet outperforms 2-D CNN on all indexes on all four data sets. Specifically, SpecAttenNet increases accuracies significantly by 7.46% of OA, 4.75% of AA, and 0.0849 of Kappa coefficient on the Indian Pines data set; by 2.21% of OA, 1.28% of AA, and 0.0293 of Kappa coefficient on the Pavia University data set; by 2.76% of OA, 2.87% of AA, and 0.0303 of Kappa coefficient on the Salinas scene; and by 3.1% of OA, 4.93% of AA, and 0.0333 of Kappa coefficient on the Houston scene. This shows that recalibrated spectral bands obtained by our gating mechanism become more separable for a spectral-spatial classification network, as informative bands have been emphasized, and less useful ones have been suppressed.

D. Results and Discussion

Tables VI–IX give information about per-class accuracies, OAs, AAs, and kappa coefficients obtained by various spectral

[1] https://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[2] https://github.com/twgr/ccfs

TABLE VI

ACCURACY COMPARISONS FOR THE INDIAN PINES SCENE. BOLD NUMBERS INDICATE THE BEST PERFORMANCE

Class No.	Class Name	RF-200	SVM-RBF	CCF-200	SICNN	2D CNN	SpecAttenNet
1	Alfalfa	56.65	71.39	76.37	79.84	54.77	**90.46**
2	Corn-notill	55.48	71.05	77.93	92.47	**96.94**	92.22
3	Corn-min	82.07	86.96	94.57	99.46	99.46	**100**
4	Corn	85.23	91.72	94.41	93.29	**96.87**	93.96
5	Grass-pasture	80.20	85.80	91.39	92.68	94.12	**95.55**
6	Grass-trees	94.99	93.85	97.04	96.58	96.81	**99.77**
7	Grass-pasture-mowed	77.02	75.38	90.96	86.82	**91.29**	89.65
8	Hay-windrowed	57.94	59.88	69.48	69.52	**93.05**	88.79
9	Oats	62.94	76.24	**89.01**	83.69	87.59	85.64
10	Soybean-notill	95.06	96.91	98.77	**100**	**100**	**100**
11	Soybean-mintill	88.67	79.58	93.73	**96.70**	68.57	96.22
12	Soybean-clean	53.33	74.84	74.55	96.97	88.48	**98.79**
13	Wheat	97.78	97.78	**100**	**100**	**100**	**100**
14	Woods	56.41	79.49	**97.44**	94.87	82.05	94.87
15	Buildings-grass-trees	81.82	**100**	90.91	**100**	**100**	**100**
16	Stone-steel-towers	**100**	**100**	**100**	**100**	**100**	**100**
OA	-	69.31	74.24	82.87	85.13	84.76	**92.22**
AA	-	76.60	83.80	89.78	92.68	90.62	**95.37**
Kappa	-	0.6538	0.7093	0.8059	0.8313	0.8261	**0.9110**

TABLE VII

ACCURACY COMPARISONS FOR THE PAVIA UNIVERSITY SCENE. BOLD NUMBERS INDICATE THE BEST PERFORMANCE

Class No.	Class Name	RF-200	SVM-RBF	CCF-200	SICNN	2D CNN	SpecAttenNet
1	Asphalt	81.54	82.37	86.59	84.21	83.85	**86.71**
2	Meadows	55.39	67.87	72.33	91.10	96.09	**98.47**
3	Gravel	53.07	69.18	71.75	64.36	**81.47**	77.47
4	Trees	98.76	98.37	**99.09**	95.53	96.12	96.83
5	Metal Sheets	99.11	99.41	**99.78**	97.70	98.74	98.81
6	Bare Soil	79.10	93.64	**97.26**	56.53	49.79	53.11
7	Bitumen	84.36	91.20	**91.88**	77.29	79.32	77.82
8	Bricks	91.39	92.59	94.92	**95.57**	88.89	94.43
9	Shadows	97.47	96.94	**98.73**	96.20	94.19	96.30
OA	-	71.53	79.89	83.36	85.25	86.93	**89.14**
AA	-	82.24	87.95	**90.26**	84.28	85.38	86.66
Kappa	-	0.6504	0.7491	0.7905	0.8041	0.8242	**0.8535**

TABLE VIII

ACCURACY COMPARISONS FOR THE SALINAS DATA. BOLD NUMBERS INDICATE THE BEST PERFORMANCE

Class No.	Class Name	RF-200	SVM-RBF	CCF-200	2D CNN	SpecAttenNet
1	Brocoli_green_weeds_1	99.29	98.98	**99.49**	71.57	94.84
2	Brocoli_green_weeds_2	99.21	99.67	99.95	99.86	**99.97**
3	Fallow	97.72	98.70	99.43	88.89	**99.64**
4	Fallow_rough_plow	97.62	97.77	**99.33**	98.14	98.88
5	Fallow_smooth	97.95	98.33	98.82	98.17	**99.81**
6	Stubble	99.41	99.72	99.80	**100**	99.69
7	Celery	99.23	99.46	99.66	97.00	**99.69**
8	Grapes_untrained	61.92	70.37	67.56	70.79	**84.34**
9	Soil_vinyard_develop	98.70	98.59	99.19	**99.45**	98.39
10	Corn_senesced_green_weeds	85.56	93.74	93.80	**96.19**	95.14
11	Lettuce_romaine_4wk	91.75	94.70	95.87	96.37	**98.82**
12	Lettuce_romaine_5wk	98.24	99.89	99.95	**100**	99.63
13	Lettuce_romaine_6wk	97.69	97.81	98.15	**100**	**100**
14	Lettuce_romaine_7wk	92.25	97.35	96.86	98.33	**99.90**
15	Vinyard_untrained	70.32	71.53	80.77	**91.22**	79.36
16	Vinyard_vertical_trellis	96.98	**98.18**	**98.18**	93.00	96.93
OA	-	86.02	88.82	89.72	90.25	**93.01**
AA	-	92.74	94.67	95.43	93.69	**96.56**
Kappa	-	0.8450	0.8757	0.8858	0.8918	**0.9221**

and spectral-spatial classification methods on the four data sets. For spectral classification approaches, CCF-200 outperforms RF-200 and SVM-RBF. With respect to the obtained classification results, deep networks, including SICNN, 2-D CNN, and the proposed SpecAttenNet show better performance than "shallow" models (i.e., random forest, SVM,

TABLE IX
ACCURACY COMPARISONS FOR THE HOUSTON SCENE. BOLD NUMBERS INDICATE THE BEST PERFORMANCE

Class No.	Class Name	RF-200	SVM-RBF	CCF-200	2D CNN	SpecAttenNet
1	Healthy grass	82.53	82.24	**83.10**	82.91	78.63
2	Stressed grass	83.46	83.18	83.46	**100**	85.81
3	Synthetic grass	97.82	99.80	**100**	75.05	94.65
4	Trees	91.38	92.05	91.48	89.49	**97.82**
5	Soil	96.59	98.58	98.67	99.53	**100**
6	Water	98.60	**99.30**	**99.30**	93.71	89.51
7	Residential	74.81	78.92	**87.59**	76.12	79.10
8	Commercial	32.48	48.81	46.34	71.32	**78.92**
9	Road	69.41	77.90	73.84	80.64	**87.72**
10	Highway	43.73	62.07	66.41	53.96	**70.37**
11	Railway	69.83	81.31	**84.63**	76.57	74.67
12	Parking Lot 1	53.70	81.75	85.98	**88.86**	76.08
13	Parking Lot 2	61.40	71.23	73.68	85.96	**90.53**
14	Tennis Court	99.19	**100**	98.79	81.38	98.38
15	Running Track	97.89	97.04	**98.10**	68.50	95.77
OA	-	72.86	80.80	82.15	81.40	**84.50**
AA	-	76.85	83.61	84.76	81.60	**86.53**
Kappa	-	0.7085	0.7933	0.8069	0.7985	**0.8318**

TABLE X
ASSESSMENTS OF THE SIGNIFICANCE OF CLASSIFICATION ACCURACIES OF THE PROPOSED METHOD COMPARED TO OTHER INVESTIGATED APPROACHES FOR THE FOUR DATA SETS.

Data Set	RF-200	SVM-RBF	CCF-200	SICNN	2D CNN
Indian Pines	40.953	35.169	21.278	19.280	19.255
Pavia University	64.010	36.743	24.161	22.904	16.895
Salinas	2.021	25.101	22.943	-	31.336
Houston	27.389	9.720	6.742	-	8.377

and CCF) in regard to OA and kappa coefficient, mainly because: 1) they are capable of extracting hierarchical, deep feature representations; 2) spatial information can be fully exploited in them. These two properties make the deep networks more robust in finding appropriate decision boundaries and enable the models to handle nonlinearly separable data more efficiently.

On the other hand, in comparison with SICNN that selects the most informative spectral bands as inputs of a CNN using a band selection approach, SpecAttenNet is capable of achieving accuracy increments of 7.09%, 2.69%, and 0.0797 for OA, AA, and Kappa coefficient, respectively, on the Indian Pines scene. Regarding the Pavia University scene, the accuracy increments on OA, AA, and Kappa coefficient are, respectively, 3.89%, 2.38%, and 0.0494. This observation reveals that compared to conventional band selection methods, our data- and task-driven spectral attention mechanism can offer better results.

Table X demonstrates the results of McNemar's test, in which we compute our method and other competitors in terms of the significance of the difference between their classification results. We can see that on both data sets, the improvement of accuracies yielded by our approach is statistically significant as compared with other methods. Figs. 3–5 show classification maps produced by RF-200, SVM-RBF, CCF-200, SICNN, 2-D CNN, and SpecAttenNet on three scenes. As displayed in these figures, spectral classifiers (i.e., random forest, SVM, and CCF) lead to salt and pepper noised classification maps, while this issue is addressed in spectral-spatial classification networks (SICNN, 2-D CNN, and SpecAttenNet) by removing noisy scattered points of misclassification.

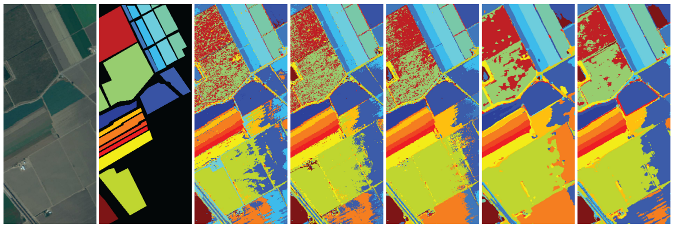


Fig. 5. Classification maps of different approaches for the Salinas data set. From left to right: true-color composite of the hyperspectral image, reference data, RF-200, SVM-RBF, CCF-200, 2-D CNN, and SpecAttenNet. Best zoomed-in view.

Moreover, we observe that the use of the spectral attention module alleviates the problem of misclassification. For instance, misclassification in the Indian Pines data set lies in similar objects (with extremely similar spectral characteristics), such as Alfalfa and Hay-windrowed. SpecAttenNet achieves the best average accuracy of 89.625% on these two classes, while the second best average accuracy is only 74.68%, as obtained by SICNN.

E. Analysis of the Spectral Attention Module

One challenge in hyperspectral data classification is that due to complex light scattering mechanism, some pixels of a hyperspectral image, which belong to the same land cover class, have different spectral signatures. Therefore, an approach that is capable of making spectral signals of those pixels that are more similar should be able to offer a more accurate classification result. Here, to quantitatively verify the effectiveness of the spectral attention module, an index called within-class similarity measures is used. The within-class similarity measure is defined as the trace of the

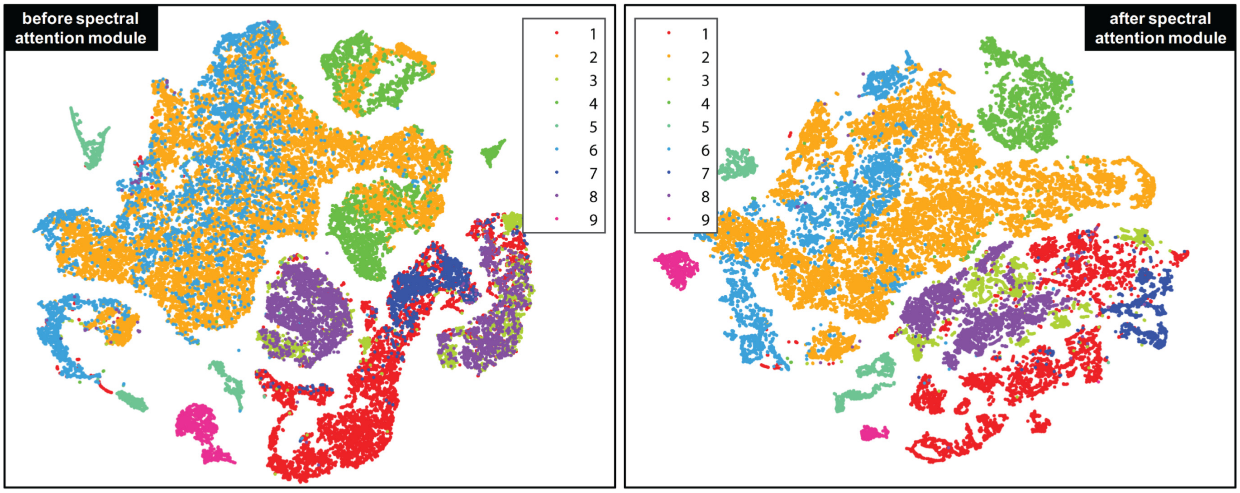


Fig. 6. Visualization of original samples and recalibrated ones by the spectral attention module of the Pavia University data set by t-SNE [70]. Different colors represent different categories. As shown in this figure, after the attention module, samples of some classes (e.g., class 2 and class 6) gather together and come into several groups, which means outputs of the module are more useful for tasks like classification. This is mainly because by making use of the proposed gating mechanism, bands that provide discriminative information are emphasized, while the others are suppressed.

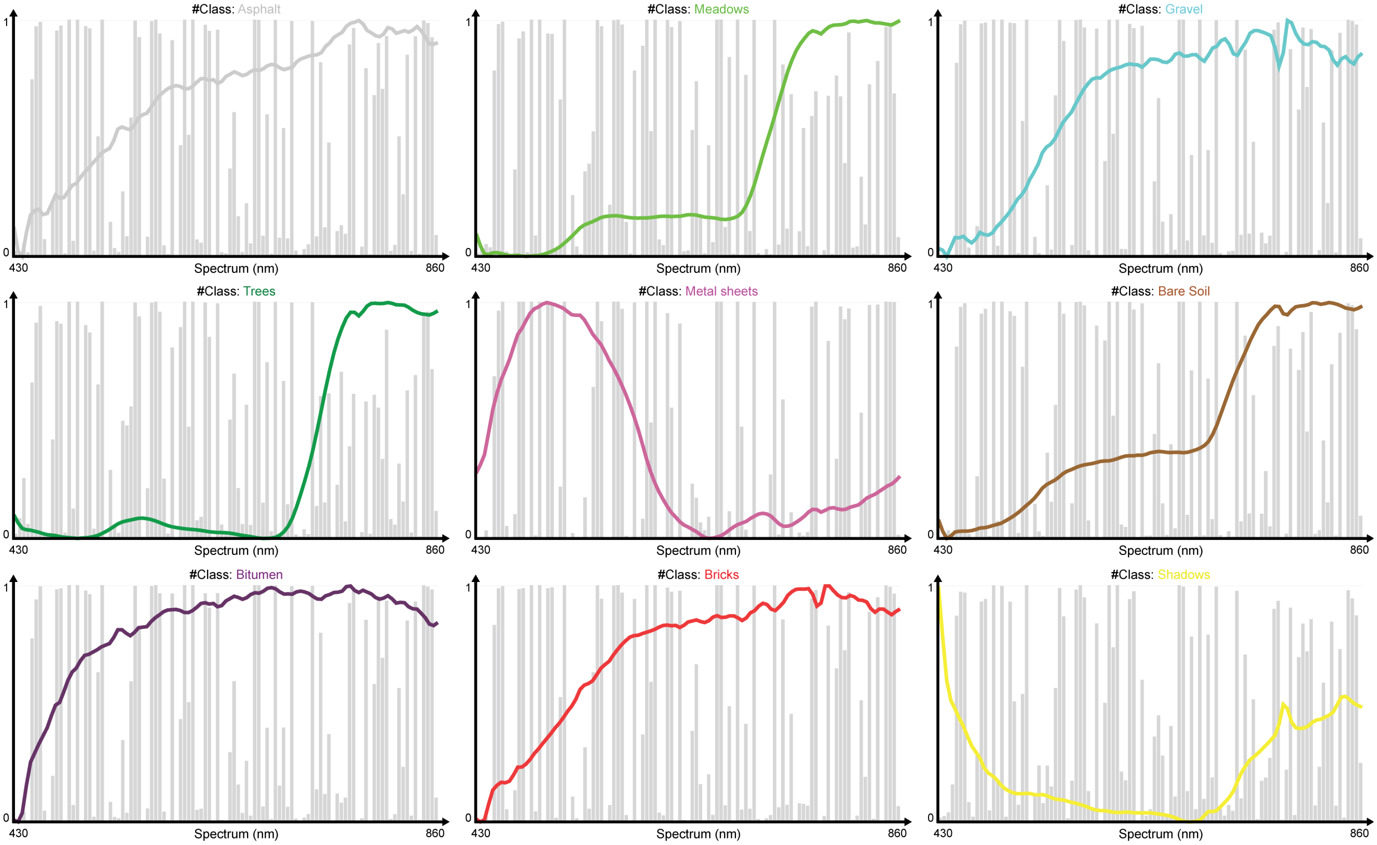


Fig. 7. Average reflectance spectrum and average spectral gates of each class on the Pavia University data set.

within-class scatter matrix, which can be calculated as follows:

$$\boldsymbol{S}_w = \sum_c \sum_{i \in c} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^T \tag{13}$$

where

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{i \in c} \boldsymbol{x}_i \tag{14}$$

and N_c denotes the amount of test data belonging to the c-th category.

Table XI reports calculated within-class similarity measures of features before and after the spectral attention module in our network on both data sets. We can observe that recalibrated spectra (i.e., outputs of the spectral attention module) in the same category have higher similarity.

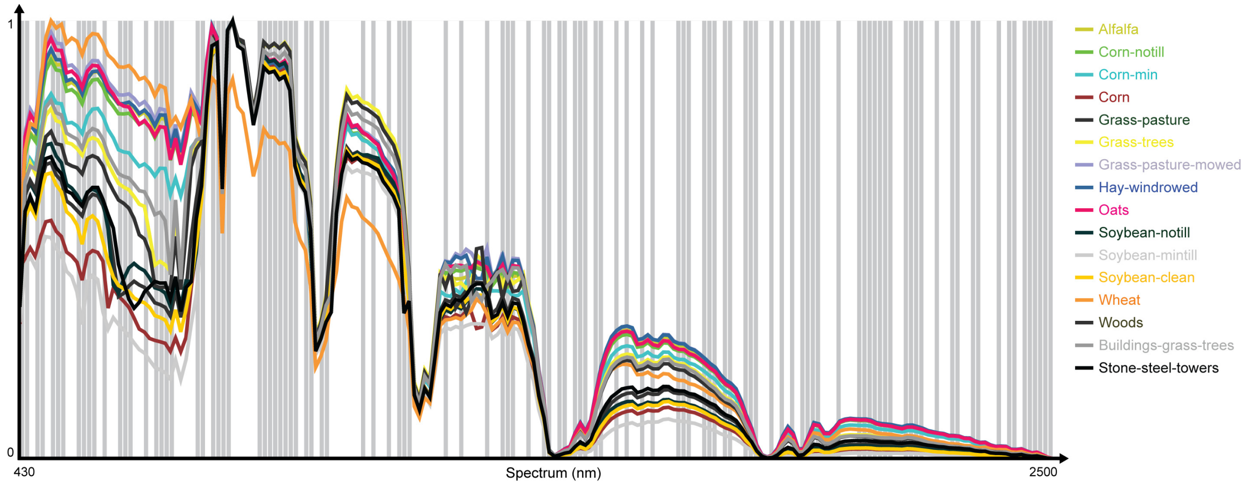


Fig. 8. Average reflectance spectrum of each class and learned spectral gates on the Indian Pines data set.

TABLE XI

WITHIN-CLASS SIMILARITY MEASURES OF FEATURES BEFORE AND AFTER THE SPECTRAL ATTENTION MODULE ON THE INDIAN PINES, PAVIA UNIVERSITY, AND SALINAS DATA SETS. SMALLER IS BETTER

Data Set	Before	After
Indian Pines	17.089	**9.403**
Pavia University	2.289	**1.058**
Salinas	2.240	**0.198**

Hence, the results demonstrate that the recalibrated spectra are more discriminative.

Furthermore, we use t-SNE [70] technique to visualize spectra before and after this module on the Pavia University scene in Fig. 6. As shown in this figure, after the attention module, samples of some classes (e.g., class 2 and class 6) gather together and come into several groups, which means outputs of the module are more useful for tasks like classification. This is mainly because by making use of the proposed gating mechanism, bands that provide discriminative information are emphasized, while others are suppressed.

Since the designed spectral attention mechanism is data- and task-driven, according to (3), different inputs have different spectral gates. For each class, we calculate the average of spectral gates of test samples belonging to this class and name it average spectral gate. Fig. 7 exhibits the average reflectance spectrum and the average spectral gate learned by our attention module of each class on the Pavia University scene. As shown in this figure, classes with similar spectral signatures (e.g., Gravel and Bricks) have extremely similar spectral gates, while these similar classes can be differentiated in detail; for example, we can see that activations of some gates on the Gravel class and the Bricks class are different. In Fig. 8, we also display the average reflectance spectrum of each class and learned spectral gates on the Indian Pines data set. Note that since spectral gates of all test samples learned on this scene are almost the same, we visualize the average spectral gate of all samples instead of each class. Interestingly, the learned spectral gate on this data set is nearly completely binary and quite different from the gates on the Pavia University scene. From Fig. 8, we can observe that the spectral attention module mainly pays attention on spectral bands that provide visual cues to distinguish different categories.

IV. CONCLUSION

This work proposed a simple, yet effective end-to-end trainable spectral attention module to make a spectral-spatial classification CNN learn a channel attention mechanism, i.e., how to pay attention on the spectral domain, for hyperspectral image classification. Our spectral attention module enhances the network by learning the importance of spectral bands with a gating mechanism and performing a dynamic band-wise recalibration, which improves not only the representational capacity but also the interpretability of the network. Extensive experiments validate the effectiveness of our network.

In the future, we will carry out further research and try to figure out the band importance induced by the spectral attention module, which may be helpful to related fields, e.g., band selection and hyperspectral data classification network pruning for model compression.

REFERENCES

[1] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[2] S. R. Joelsson, J. A. Benediktsson, and J. R. Sveinsson, “Random forest classifiers for hyperspectral data,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2005, p. 4.

[3] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, “Random Forests for land cover classification,” *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, Mar. 2006.

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

[5] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

[6] 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.

[7] B. Waske and J. A. Benediktsson, "Fusion of support vector machines for classification of multisensor data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 12, pp. 3858–3866, Dec. 2007.

[8] B. Waske, S. van der Linden, J. Benediktsson, A. Rabe, and P. Hostert, "Sensitivity of support vector machines to random feature selection in classification of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2880–2889, Jul. 2010.

[9] G. F. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio, "Large margin deep networks for classification," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2018, pp. 850–860.

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 3104–3112.

[11] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1746–1751.

[12] K. Cho, B. van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–15.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1–9.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. IEEE Int. Conf. Learn. Represent. (ICLR)*, Apr. 2015, pp. 1–14.

[15] Y. Yuan, L. Mou, and X. Lu, "Scene recognition by manifold regularized deep learning architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2222–2233, Oct. 2015.

[16] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

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

[18] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[20] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," Jun. 2016, *arXiv:1606.00915*. [Online]. Available: https://arxiv.org/abs/1606.00915

[21] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.

[23] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.

[24] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," Jun. 2017, *arXiv:1706.06905*. [Online]. Available: https://arxiv.org/abs/1706.06905

[25] L. Mou and X. X. Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," Feb. 2018, *arXiv:1802.1024*. [Online]. Available: https://arxiv.org/abs/1802.10249

[26] 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.

[27] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

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

[29] 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.

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

[31] X. Lu, W. Zhang, and X. Li, "A hybrid sparsity and distance-based discrimination detector for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1704–1717, Mar. 2018.

[32] 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.

[33] 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.

[34] 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.

[35] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

[36] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[37] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 391–406, Jan. 2018.

[38] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep&dense convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 10, no. 9, p. 1454, 2018.

[39] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019. doi: 10.1109/TGRS.2018.2860125.

[40] L. Mou, P. Ghamisi, and X. X. Zhu, "Fully conv-deconv network for unsupervised spectral-spatial feature extraction of hyperspectral imagery via residual learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5181–5184.

[41] X. Ma, A. Fu, J. Wang, H. Wang, and B. Yin, "Hyperspectral image classification based on deep deconvolution network with skip architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4781–4791, Aug. 2018.

[42] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[43] 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.

[44] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," Mar. 2018, *arXiv:1803.02642*. [Online]. Available: https://arxiv.org/abs/1803.02642

[45] M. Rußwurm and M. Körner, "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshop*, Jul. 2017, pp. 1496–1504.

[46] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 6, p. 506, 2016.

[47] H. Lyu *et al.*, "Long-term annual mapping of four cities on different continents by applying a deep information learning method to Landsat data," *Remote Sens.*, vol. 10, no. 3, p. 471, 2018.

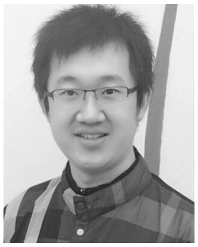
[48] H. Wu and S. Prasad, "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sens.*, vol. 9, no. 3, p. 298, 2017.

[49] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.

[50] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[51] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "A new deep convolutional neural network for fast hyperspectral image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 120–147, Nov. 2018.

[52] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic routing between capsules,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1–11.
[53] M. E. Paoletti *et al.*, “Capsule networks for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2145–2160, Apr. 2019.
[54] F. I. Alam, J. Zhou, A. W.-C. Liew, X. Jia, J. Chanussot, and Y. Gao, “Conditional random field and deep feature learning for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1612–1628, Mar. 2019.
[55] 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.
[56] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, “Bidirectional attentive fusion with context gating for dense video captioning,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7190–7198.
[57] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, “End-to-end comparative attention networks for person re-identification,” *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.
[58] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.
[59] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[60] A. Graves, “Generating sequences with recurrent neural networks,” Aug. 2013, *arXiv:1308.0850*. [Online]. Available: https://arxiv.org/abs/1308.0850
[61] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” in *Proc. 8th Workshop Syntax, Semantics Struct. Stat. Transl. (SSST)*, Oct. 2014, pp. 103–167.
[62] Y. Gal and Z. Ghahramani, “A theoretically grounded application of dropout in recurrent neural networks,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1019–1027.
[63] J. M. Haut, M. E. Paoletti, J. Plaza, A. Plaza, and J. Li, “Visual attention-driven hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, to be published. doi: 10.1109/TGRS.2019.2918080.
[64] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2010, pp. 249–256.
[65] T. Dozat. *Incorporating Nesterov Momentum Into Adam*. Accessed: Sep. 22, 2019. [Online]. Available: http://cs229.stanford.edu/proj2015/054_report.pdf
[66] Y. LeCun *et al.*, “Backpropagation applied to handwritten zip code recognition,” *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.
[67] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. IEEE Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.
[68] T. Rainforth and F. Wood, “Canonical correlation forests,” Jul. 2015, *arXiv:1507.05444*. [Online]. Available: https://arxiv.org/abs/1507.05444
[69] J. Xia, N. Yokoya, and A. Iwasaki, “Hyperspectral image classification with canonical correlation forests,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 421–431, Jan. 2017.
[70] L. van der Maaten, “Accelerating t-SNE using tree-based algorithms,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014.



Lichao Mou (S’16) received the bachelor’s degree in automation from the Xi’an University of Posts and Telecommunications, Xi’an, China, in 2012, and the master’s degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Wessling, Germany, and also with the Technical University of Munich (TUM), Munich, Germany.

In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with the University of Cambridge, Cambridge, U.K. His research interests include remote sensing, computer vision, and machine learning, especially deep networks and their applications in remote sensing.

Mr. Mou was a recipient of the first place in the 2016 IEEE GRSS Data Fusion Contest and finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.

Xiao Xiang Zhu (S’10–M’12–SM’14) received the master’s (M.Sc.), D.E. (Dr.-Ing.), and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, in 2009; Fudan University, Shanghai, China, in 2014; The University of Tokyo, Tokyo, Japan, in 2015; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2016. Since 2019, she has been co-coordinating the Munich Data Science Research School. She is also leading the Helmholtz Artificial Intelligence Cooperation Unit (HAICU)–Research Field “Aeronautics, Space and Transport.” She is currently a Professor of signal processing in earth observation with the Technical University of Munich (TUM) and also with the German Aerospace Center (DLR); also the Head of the Department “EO Data Science,” DLR’s Earth Observation Center; and also the Head of the Helmholtz Young Investigator Group “SiPEO,” DLR and TUM. Her research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of young academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.