Remote Sensing Scene Classification Using Multilayer Stacked Covariance Pooling

Nanjun He, *Student Member, IEEE,* Leyuan Fang, *Senior Member, IEEE,* Shutao Li, *Senior Member, IEEE,* Antonio Plaza, *Fellow, IEEE,* and Javier Plaza, *Senior Member, IEEE*

***Abstract*—This paper proposes a new method, called multilayer stacked covariance pooling (MSCP), for remote sensing scene classification. The innovative contribution of the proposed method is that it is able to naturally combine multi-layer feature maps, obtained by pre-trained convolutional neural network (CNN) models. Specifically, the proposed MSCP based classification framework consists of the following three steps. First, a pre-trained CNN model is used to extract multi-layer feature maps. Then, the feature maps are stacked together, and a covariance matrix is calculated for the stacked features. Each entry of the resulting covariance matrix stands for the covariance of two different feature maps, which provides a natural and innovative way to exploit the complementary information provided by feature maps coming from different layers. Finally, the extracted covariance matrices are used as features for classification by a support vector machine (SVM). The experimental results, conducted on three challenging data sets, demonstrate that the proposed MSCP method can not only consistently outperform the corresponding single-layer model, but also achieve better classification performance than other pre-trained CNN based scene classification methods.**

***Index Terms*—Remote sensing scene classification, pre-trained convolutional neural networks (CNN), multilayer feature maps, feature fusion.**

I. Introduction

REMOTE sensing scene classification has received considerable attention recently, as can beused in many practical applications, such as natural hazards detection, geographic image retrieval, urban planning, etc. [1]–[3]. Given a query remote sensing image, scene classification aims to assign a unique label (e.g., *industrial area* or *airport*) to the image, based on its contents. However, remote sensing scene classification is a challenging problem since the scene images often exhibit complex spatial structures with high intraclass and low interclass variabilities. To address this problem, many scene classification methods have been proposed over the past years [4]–[11]. An extensive review of remote sensing scene classification methods can be found in [1], [2].

Recently, inspired by the great success achieved by convolutional neural networks (CNN) in the computer vision community [12]–[14], a considerable number of CNN-based models have been proposed for remote sensing scene classification [15]–[19]. These models can achieve better classification performance than other traditional methods. The success of CNN-based scene classification methods is mainly due to the fact that pre-trained CNN models (e.g., AlexNet [20], VGG-VD16 [21], GoogleNet [22]) on the ImageNet [23] exhibit powerful generalization ability and can extract more representative features than traditional feature extraction methods (e.g., scale-invariant feature transform (SIFT) [24] or color histograms).

However, although these methods can obtain very good classification performance, the issue of how to utilize pre-trained CNN models effectively for remote sensing scene classification is still a open question. In this paper, we propose a new method, called multilayer stacked covariance pooling (MSCP), to combine the feature maps from different layers of a pre-trained CNN for remote sensing scene classification. The proposed MSCP scene classification framework includes three main steps. In the first step, a pre-trained CNN model (i.e., AlexNet or VGG-VD16) is used to extract multilayer feature maps. Then, the feature maps are stacked together and a covariance matrix is calculated. Each entry in the covariance matrix stands for the covariance between two different feature maps, which serves as a natural mechanism to fuse the feature maps from different layers. Finally, the obtained covariance matrices are used as features for classification using a support vector machine (SVM) classifier with linear kernel. We note that, in order to stack feature maps with different spatial dimensions together, downsampling is adopted. Moreover, channel-wise average fusion is proposed and applied on each convolutional layer to reduce computational complexity before stacking the maps together.

The main motivation of the proposed method is to be able to exploit the complementary information among different convolutional layers to further enhance classification performance. As pointed by Lecun *et al.* in [26], the basic idea behind the CNN model is to represent the image from raw to abstract using multi-level architectures. The sallower layers of the CNN model are more likely to reflect the low-level visual features (e.g., edges), while the deeper layers represent more

This paper was supported by the National Natural Science Fund of China for International Cooperation and Exchanges under Grant 61520106001, the National Natural Science Foundation for Young Scientist of China under Grant No. 61501180, and the Fund of Hunan Province for Science and Technology Plan Project under Grant 2017RS3024. This work was also supported by the China Scholarship Council.

N. He is with the College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China, and also with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, E-10003 Cáceres, Spain (henanjun@hnu.edu.cn).

L. Fang and S. Li are with the College of Electrical and Information Engineering, Hunan University, Changsha, 410082, China (e-mail: fangleyuan@gmail.com; shutao_li@hnu.edu.cn).

A. Plaza and J. Plaza are with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, E-10003 Cáceres, Spain (e-mail: aplaza@unex.es; jplaza@unex.es).

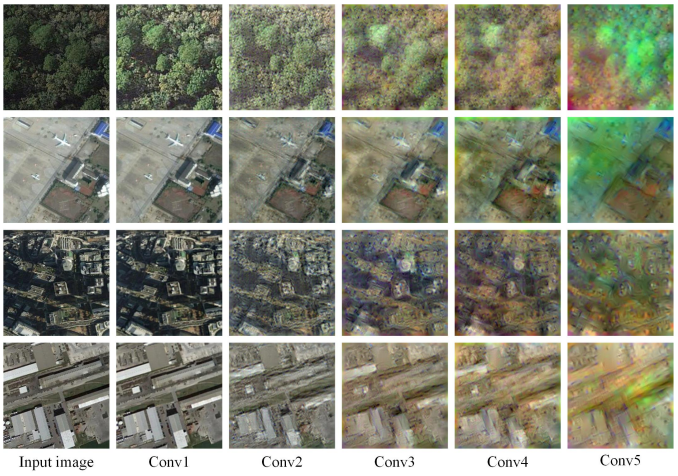

Fig. 1. An example illustrating the reconstruction [25] of feature maps from different convolutional layers (conv) of AlexNet [20]. From left to right we show the feature maps resulting from layers: conv1, conv2, conv3, conv4, and conv5, respectively. The maps at different levels are shown to convey complementary information that can be used to further improve the classification performance.

abstract information contained in the images. Fig. 1 provides a graphical illustration of the feature maps in different layers of a typical CNN model (i.e., AlexNet [20]). As can beobserved in Fig. 1, the AlexNet (with its hierarchical architecture) can extract various feature maps from the image, and these maps are expected to convey complementary information that can be used to further improve the classification performance. In addition, by using the so-called *shortcut* connections to combine different layers, from shallow to deep, the recently proposed ResNet [27] and DenseNet [28] can achieve state-of-the-art performance in different computer vision tasks, which also suggests that the combination of different layers from the CNN can be very useful. In this regard, our proposed method uses a similar approach to exploit the complementary information contained by multiple layers. Our experiments demonstrate that the proposed MSCP method can indeed exploit such complementary information and achieve better classification performance than several state-of-the-art approaches.

The reminder of this paper is organized as follows. Section II gives an overview of related works and presents the main innovative contributions of our proposed approach. Section III details the proposed MSCP framework. In Section IV, comprehensive experimental results are reported on three data sets, and an exhaustive comparison to other state-of-the-art methods is also given. Section V concludes the paper with some remarks and hints at plausible future research lines.

II. Related Works and Contributions

Generally, existing scene classification methods can be categorized into three classes: 1) Low level visual feature (LLF) oriented methods, 2) mid-level visual feature (MLF) oriented methods, and 3) high level visual feature (HLF) oriented methods. For the LLF oriented methods, a local or global feature descriptor is first extracted to represent the test images. Then, the obtained features are sent to a supervised classifier such as the SVM for label assignment. In [29], Yang and Newsam combine Gabor features with the maximum a posteriori (MAP) model for scene classification, where each test image is represented by a vector consisting of the mean and standard deviation of the corresponding Gabor feature. Moreover, global color histograms are used to characterize the image and the SVM is then utilized to classify the obtained feature vectors [30]. In [31], a sparse representation [32]–[36] is adopted to combine several LLFs (e.g., local binary patterns (LBP) [37] and histogram of oriented gradients (HOG) [38]) to enhance classification performance.

Bearing in mind that there may be semantic gaps between LLFs and the high level semantic meaning of images, MLF oriented methods are introduced to bridge these two levels [39], [40]. In [30], Yang *et al.* use a bag of visual words (BoVW) model to encode the SIFT descriptor for MLF extraction. The MLFs are fed to a SVM with intersection kernel for classification. To further take into account the spatial-contextual information, the spatial pyramid matching (SPM) is used to extend the BoVW model in [41], [42]. In [7], Zhu *et. al.* utilize the BovW model to combine both local and global features, extracted from the images, to enhance classification performance. With the introduction of *partlets*, which are a library of pre-trained part detectors used for mid-level visual element discovery, an effective and efficient MLF method was proposed in [4]. The probabilistic topic model is another popular technique to bridge the semantic gap between LLFs and the high level semantic meaning [43], by means of which the input scene is represented as a probability distribution of the visual words. In [5], a multiple topic model was proposed to combine several different complementary features in order to achieve a discriminative MLF feature extraction. In addition, a sparse topic model was recently proposed to integrate homogeneous and heterogeneous features for scene classification [44]. In [45], a multitask learning method is proposed to take both the multi-resolutions analysis and feature selection into account for scene classification. In [46], Du *et. al* propose a local structure learning framework to make use of the local topological construction of images for remote scene image retrieval. Instead of using handcrafted features such as SIFT, other unsupervised feature learning methods based on different concepts have also been recently proposed [6], [47], [48].

Inspired by the recent success achieved by CNNs in the computer vision community, CNN models have also been extended for remote sensing scene classification [16]. However, training a deep CNN model from scratch generally needs a huge amount of training data, while available off-the-shelf remote sensing scene image data sets are relatively small. For example, deep CNN models are usually trained on the ImageNet [23], which contains millions of images, while the NWPU-RESISC45 data set [1] (one of the biggest data sets for remote sensing scene classification) contains less than 35K images. Moreover, CNN models pre-trained on ImageNet show powerful generalization ability on different tasks (e.g., object detection and semantic segmentation [49], [50]). Under this context, the perspective of using off-the-shelf pre-trained CNN models such as AlexNet [20], VGG-VD16 [21] and GoogleNet [22] as a universal feature extractors

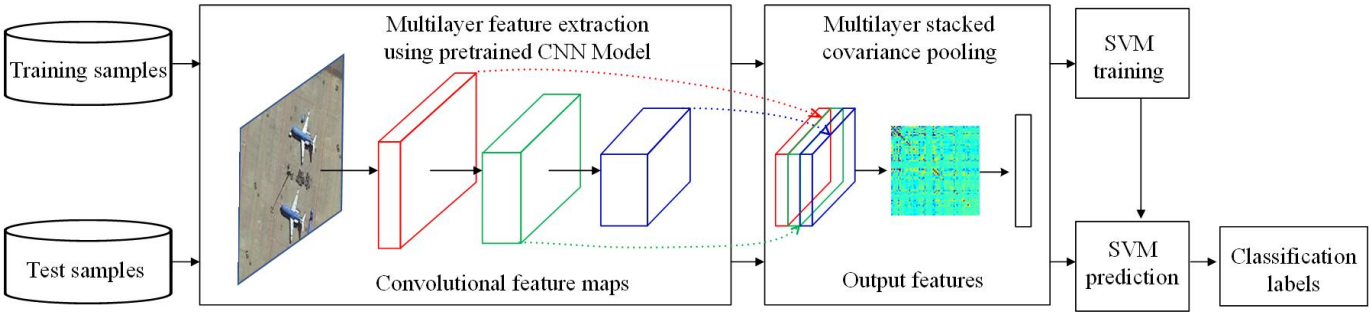

Fig. 2. Flowchart of the proposed MSCP classification framework. The proposed framework consists of three steps: 1) multi-layer feature extraction using a pre-trained CNN model, 2) multi-layer stacked covariance pooling, and 3) SVM classification. The dotted and colored lines denote downsampling and channel-wise average fusion, respectively.

has become an interesting approach for remote sensing scene classification. In [51], the GoogleNet is used for remote sensing scene classification, showing that pre-trained CNN models can outperform conventional handcrafted feature based methods by a large margin. In [15], Hu *et al*. considered two different scenarios to utilize a pre-trained CNN model (VGG-VD16). In the first scenario, the last few fully-connected layers are regarded as final image features for scene classification. In the second scenario, a traditional feature encoding method such as the improved Fisher kernel (IFK) [52] is used to encode the feature maps from the last convolutional layer for representing the input image. Both scenarios adopt the SVM as the final classifier. In [53], Gong *et al.* use the BoVW model to encode a single convolutional layer. In [17], the last two fully connection layers of the CNN model are fused together to represent the image. In [18], a multiscale IFK coding method is proposed to combine the feature maps from different layers.

More recently, as a second order pooling strategy, covariance pooling (CP) has been used in many computer vision tasks, such image segmentation [54], [55] and classification [56]. There are two main advantages of the CP approach. First, different from conventional pooling methods, the CP takes second-order statistics (i.e., covariance) into consideration and, therefore, obtains a more compact and discriminative representation. Second, each entry in the covariance matrix obtained by CP represents the covariance between two different feature maps. This offers a natural way to fuse complementary information coming from different feature maps.

The proposed method is different from existing pre-trained CNN based methods in the following two main innovative aspects. First, we utilize the different convolutional feature maps of the CNN (from shallow to deep layers) rather than the last one or two connection layers for representing the input image. As a result, the proposed approach can achieve better classification performance than the methods in [15], [17], [51], [53]. Second, we adopted a simple yet effective method (i.e., CP) to combine feature maps from different layers. Thus, the proposed method can run much faster than the method in [18], with very competitive classification performance, and is suitable to deal with relative large data sets such as the NWPU-RESISC45 data set in [1]).

III. Proposed Method

Fig. 2 illustrates the proposed MSCP based classification framework, which consists of the following three steps: 1) multi-layer feature extraction using a pre-trained CNN model, 2) multi-layer stacked covariance pooling, and 3) SVM based classification. In the following, we describe each one of the aforementioned steps in more details.

1) *Multi-layer Feature Extraction:* The CNN model can be thought as a composition of a number of functions as shown in Eq. (1), where each function f_l takes the data samples $\mathbf{X}_l$ and a filters bank $\mathbf{w}_l$ as inputs and produces $\mathbf{X}_{l+1}$, where $l = 1 \cdots L$ and L is the number of layers.

$$f(\mathbf{X}) = f_L(\cdots f_2(f_1(\mathbf{X}; \mathbf{w}_1); \mathbf{w}_2) \cdots, \mathbf{w}_L). \quad (1)$$

For a pre-trained CNN model, the filters bank $\mathbf{w}_l$ has been learned from some big data set (e.g., ImageNet [23]). Given an input image $\mathbf{X}$, the multi-layer features are extracted as shown follows: $\mathbf{M}_1 = f_1(\mathbf{X}; \mathbf{w}_1), \mathbf{M}_2 = f_2(\mathbf{M_1}; \mathbf{w}_2)$, etc. In this work, the AlexNet and VGG-VD16 are used as pre-trained CNN models. Specifically, three convolutional layers (i.e.,'conv3','conv4','conv5') of AlexNet are adopted, which are denoted by $\mathbf{M}_3, \mathbf{M}_4, \mathbf{M}_5$, respectively. Three convolutional layers (i.e., 'conv3-3','conv4-3','conv5-3') of VGG-VD16 are also used, denoted by $\mathbf{M}_{3,3}, \mathbf{M}_{4,3}, \mathbf{M}_{5,3}$, respectively. Note that the other layers (e.g., pooling layers) are omitted in Eq. (1) for simplicity.

2) *Multi-layer Stacked Covariance Pooling:* Usually, different convolutional layers have different spatial dimensions, and thus they cannot be stacked directly. If we take the VGG-VD16 as an example, we have $\mathbf{M}_{3,3} \in \mathbb{R}^{56\times56\times256}, \mathbf{M}_{4,3} \in \mathbb{R}^{28\times28\times512}, \mathbf{M}_{5,3} \in \mathbb{R}^{14\times14\times512}$. To address this problem, downsampling with bilinear interpolation is adopted in this work. Moreover, to reduce the computational complexity (the dimension of the covariance matrix is determined by the number of feature maps), a channel-wise average fusion method is further proposed and adopted on each convolutional layer before stacking them together. For a given convolutional layer $\mathbf{Y} \in \mathbb{R}^{H\times W\times L}$ and a predefined number of fused feature maps d, the channel-wise average fusion is conducted as follows. We firstly partition the L feature maps (each feature map is with size $H \times W$) into d subsets based on its original sequence. Then, the average feature maps of each subset are

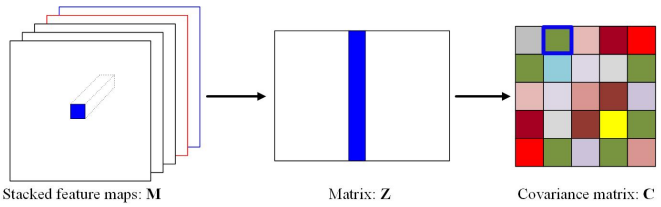

Fig. 3. Illustration of the concept of covariance pooling. The off-diagonal entries of $\mathbf{C}$ stands for the covariance between two different feature maps and the diagonal entries represent the variance of each feature map. For example, the blue square entry (i.e., the second entry in the first line) of covariance matrix $\mathbf{C}$ stands for the covariance of the last two feature maps (i.e., the red one and the blue one).

further calculated and are stacked together. Through the down-sampling and channel-wise average fusion operations, three preprocessed convolutional layers are obtained (i.e., $\hat{\mathbf{M}}_{3,3} \in \mathbb{R}^{s\times s\times d}, \hat{\mathbf{M}}_{4,3} \in \mathbb{R}^{s\times s\times d}, \hat{\mathbf{M}}_{5,3} \in \mathbb{R}^{s\times s\times d}$) and the stacked feature set is obtained as follows: $\mathbf{M} = [\hat{\mathbf{M}}_{3,3}; \hat{\mathbf{M}}_{4,3}; \hat{\mathbf{M}}_{5,3}] \in \mathbb{R}^{s\times s\times D}$, $D = 3d$ and s is the predefined downsampled spatial dimension. Finally, the CP of stacked feature $\mathbf{M}$ is expressed as follows.

$$\mathbf{C} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{z}_i - \boldsymbol{\mu})(\boldsymbol{z}_i - \boldsymbol{\mu})^T \in \mathbb{R}^{D\times D}. \tag{2}$$

where $[\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_N] \in \mathbb{R}^{D\times N}$ is the vectorization of $\mathbf{M}$ along the third dimension, $N = s^2$ and $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{z}_i \in \mathbb{R}^D$. A graphical illustration of the CP is showed in Fig. 3. The off-diagonal entries of covariance matrix $\mathbf{C}$ stands for the covariance between two different feature maps and the diagonal entries represent the variance of each feature map. For example, in Fig. 3, the blue square entry of covariance matrix $\mathbf{C}$ stands for the covariance of the last two feature maps (the red one and the blue one). From the definition of the CP (see Eq. (2)), following three advantages of CP can be concluded [57]–[59]. (Note that, in [57] and [58], the CP is firstly proposed as region feature descriptor, called region covariance descriptor, for texture classification and pedestrian detection.) Firstly, CP provides a natural way to fuse different feature maps. As we have mentioned above, each off-diagonal entry of the covariance matrix stands for the covariance of two different feature maps, which can fuse different feature maps effectively. Secondly, there is a average operation during covariance computation (see Eq. (2)), which can greatly filter the noise-corrupting individual samples. Last but not least, the computation of covariance matrix is independent from the ordering information of the samples (i.e., $\boldsymbol{z}_i$, $i = 1\cdots N$), which indicates the CP is robust to the rotation. In summary, the CP can not only make use of the second order information (i.e., covariance) to fuse different feature maps, but also be robust to the noise and rotation. Meanwhile, psychophysics research shows that the second order information plays an important role in the human visual recognition process [60] The above three distinctive advantages enable the CP becomes a very effective feature coding method and thus it could be expected that more discriminative representation can be achieved by CP, when compared to first-order pooling method (e.g., average pooling). More related works about CP and second-order pooling can be found in [55], [61], [62].

In addition, as pointed out by Arsigny *et al.*, the covariance matrices do not lie on the Euclidean space, but on the Riemannian manifold space. Thus, they cannot be processed by the SVM which is originally designed for data lying on the Euclidean space [63]. Fortunately, with the matrix logarithm operation, the covariance matrix can be mapped into Euclidean space while preserving the intrinsic geometric relationships as defined on the manifold as follows [63]:

$$\hat{\mathbf{C}} = \mathrm{logm}(\mathbf{C}) = \mathbf{U}\log(\mathbf{\Sigma})\mathbf{U}^T \in \mathbb{R}^{D\times D}, \tag{3}$$

where $\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ is the eigen-decomposition of the covariance matrix $\mathbf{C}$. More detail explanation about matrix logarithm operation can be found in [63], [64]. Note that the $\hat{\mathbf{C}}$ is a symmetric matrix and, therefore, only $\frac{D(D+1)}{2}$ entries need to be stored, which can further reduce the computational complexity. The $\frac{D(D+1)}{2}$ entries of $\hat{\mathbf{C}}$ comprise the final set of output features to represent the input image $\mathbf{X}$, denoted by $\boldsymbol{v}$. The channel-wise average fusion is an effective strategy to reduce the computational complexity. For example, the three original convolutional layers of AlexNet are with size as followings: conv3 ($13\times13\times384$), conv4 ($13\times13\times384$), conv5 ($13\times13\times256$). If we stack them together without channel-wise average fusion and then perform CP, the dimension of the obtained feature $\boldsymbol{v}$ for one single image is beyond 500K ($(384+384+256)^2/2$), which is hard to manage. By contrast, the dimension of the obtained feature $\boldsymbol{v}$ can be reduced significantly by channel-wise average fusion strategy with a small d, e.g., the dimension of $\boldsymbol{v}$ can be reduced to 29K ($(80+80+80)^2/2$) with $d = 80$, which is much less than 500K.

3) *SVM Classification:* The aforementioned operations are performed on both training samples and test samples. As such, the training set (i.e., $\{\boldsymbol{v}_i, y_i\}_{i=1\cdots n}$) and the testing set are now considered, where y_i are the corresponding labels and n is the number of training samples. Then, $\{\boldsymbol{v}_i, y_i\}_{i=1\cdots n}$ are used to train an SVM model as follows:

$$\begin{aligned}&\min_{\boldsymbol{\alpha},\xi,b}\left\{\frac{1}{2}\|\boldsymbol{\alpha}\|_2 + C\sum_i \xi_i\right\}, subject\ to\\ &y_i\left(\langle\phi(\boldsymbol{v}_i), \boldsymbol{\alpha}\rangle + b\right) \geqslant 1-\xi_i,\\ &\xi_i > 0, \forall i = 1,\cdots,n,\end{aligned} \tag{4}$$

where $\boldsymbol{\alpha}$ and b define a linear classifier, C is a regularization parameter that controls the generalization capacity of the classifier, ξ_i are positive slack variables to cope with outliers in the training set, and $\phi(\cdot)$ is the mapping function. A linear kernel is adopted in this work, with $K(\boldsymbol{v}_i, \boldsymbol{v}_j) = \boldsymbol{v}_i^T\boldsymbol{v}_j$. Finally, the prediction label of each test sample $\boldsymbol{v}$ is determined by means of a decision function, as shown below:

$$f(x) = sgn(\sum_{i=1}^{n} y_i\lambda_i K(\boldsymbol{v}_i, \boldsymbol{v}) + b), \tag{5}$$

where λ_i are the Lagrange multipliers. Note that the *one-against-all* strategy is adopted for solving the multi-class problem.

Fig. 4. Some examples of the UC data set.

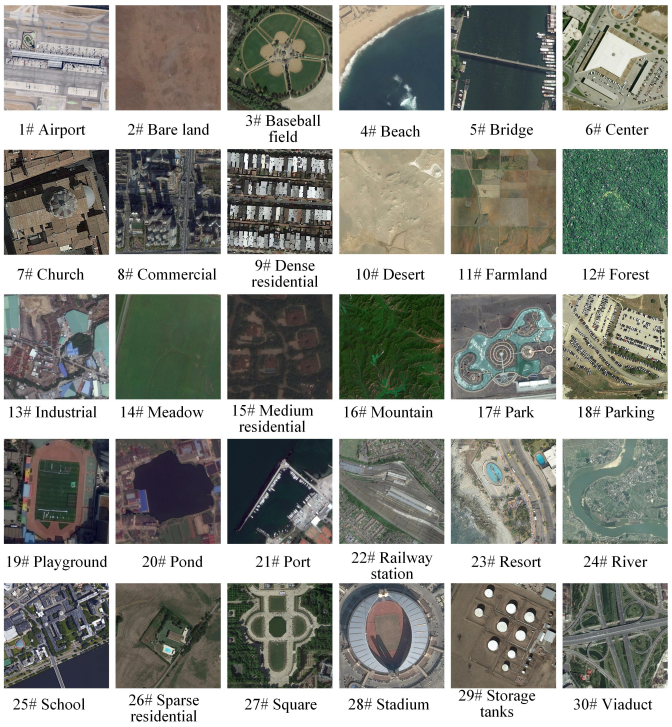

Fig. 5. Some examples of the AID data set.

IV. Experimental Results and Discussion

A. Experimental Data Sets

To evaluate the performance of proposed method, we conduct experiments on three challenging remote sensing scene image data sets.

1) *UC Merced Land Use Data Set:* The UC Merced Land Use (UC) [30] data set contains 2100 images divided into 21 scene classes. Each class consists of 100 images with size of 256×256 pixels in the RGB space. Each image has a pixel resolution of one foot. Fig. 4 shows some examples of the UC data set. As can be seen in Fig. 4, some categories have very high inter-class similarity (e.g., *forest* and *sparse residential*), which makes the UC data set a very challenging one.

2) *AID30:* The AID30 (AID) data set [2] contains 10000 images divided into 30 scene classes. Each class contain hundreds of images (ranging from 220 to 420) with size of 600×600 pixels in the RGB space. The spatial resolution changes from about 8m to 0.5m. Fig 5 shows some examples of the AID data set.

3) *NWPU-RESISC45 Data Set:* The NWPU-RESISC45 (NWPU) data set [1] contains 31500 images divided into 45 scene classes. Each class consists of 700 images, with size of 256×256 pixels in the RGB space. The spatial resolution changes from about 30m to 0.2m per pixel for most of the scene classes. This data set is of the largest available in terms of both the number of scene classes and the total number of images. Thus, it contains richer image variations, larger within-diversity, and higher inter-class similarity than the other considered data sets.

TABLE I
Convolution Layers Used by the Proposed Method and Corresponding Size of These Layers.

CNN Model	AlexNet	VGG-VD16
Size of input images	$227 \times 227 \times 3$	$224 \times 224 \times 3$
Convolutional layers	conv3 ($13 \times 13 \times 384$)	conv3-3 ($56 \times 56 \times 256$)
	conv4 ($13 \times 13 \times 384$)	conv4-3 ($28 \times 28 \times 512$)
	conv5 ($13 \times 13 \times 256$)	conv5-3 ($14 \times 14 \times 512$)
Size of each preprocessed convolutional layer	$13 \times 13 \times 80$	$14 \times 14 \times 90$
Size of stacked feature set	$13 \times 13 \times 240$	$14 \times 14 \times 270$

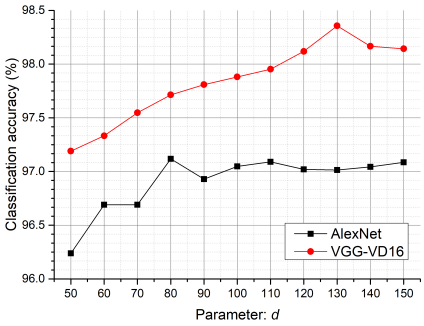

Fig. 6. Effect of parameter d on the proposed MSCP classification framework.

B. Experimental Setup

In our implementation, two popular CNN pre-trained models: AlexNet [20] and VGG-VD16 [21], are utilized to extract multi-layer features. Specifically, three convolutional layers (i.e.,'conv3','conv4','conv5') of AlexNet and three convolutional layers (e.g., 'conv3-3','conv4-3','conv5-3') of VGG-VD16 are used, respectively. Detailed information about the used convolutional layers of the CNN models is summarized in Table I. Before feeding the scene images into the pre-trained CNN model for feature extraction, the images are resized to the predefined size as shown in Table I (i.e., $227 \times 227 \times 3$ for AlexNet and $224 \times 224 \times 3$ for VGG-VG16), which are followed the experimental settings of [20] and [21], respectively. Both two models are pre-trained on ImageNet and downloaded from the homepage of MatConvNet [65] (a Matlab toolbox

TABLE II
COMPARISON BETWEEN THE CLASSIFICATION RESULTS (%) OBTAINED BY USING A SINGLE CONVOLUTIONAL LAYER PLUS COVARIANCE POOLING (CP) AND THE PROPOSED MULTI-LAYER STACKED COVARIANCE POOLING (MSCP) METHOD ON UC (TRAINING RATIO=80%) AND AID DATA SET (TRAINING RATIO=20%). '_PRE' MEANS THAT THE CONVOLUTIONAL LAYER HAS BEEN PREPROCESSED BY DOWNSAMPLING AND CHANNEL-WISE AVERAGE FUSION.

	AlexNet			VGG-VD16		
Data set	Method	Feature dimension	OA	Method	Feature dimension	OA
UC	Conv3_pre+CP	3K	95.71±0.83	Conv3-3_pre+CP	9K	96.93±0.46
	Conv4_pre+CP	3K	96.38±0.78	Conv4-3_pre+CP	9K	97.29±0.48
	Conv5_pre+CP	3K	95.24±0.76	Conv5-3_pre+CP	9K	96.76±1.00
	Conv3+CP	74K	96.11±0.55	Conv3-3+CP	33K	96.16±0.32
	Conv4+CP	74K	96.71±0.42	Conv4-3+CP	130K	97.59±0.43
	Conv5+CP	33K	96.69±0.55	Conv5-3+CP	130K	97.45±0.47
	MSCP	29K	**97.29±0.63**	MSCP	76K	**98.36±0.58**
AID	Conv3_pre+CP	3K	86.13±0.34	Conv3-3_pre+CP	9K	88.38±0.22
	Conv4_pre+CP	3K	87.40±0.28	Conv4-3_pre+CP	9K	89.85±0.37
	Conv5_pre+CP	3K	85.00±0.36	Conv5-3_pre+CP	9K	86.79±0.37
	Conv3+CP	74K	88.19±0.34	Conv3-3+CP	33K	87.53±0.32
	Conv4+CP	74K	88.48±0.42	Conv4-3+CP	130K	91.18±0.19
	Conv5+CP	33K	88.04±0.24	Conv5-3+CP	130K	89.32±0.23
	MSCP	29K	**88.99±0.38**	MSCP	76K	**91.52±0.21**

for CNN)[1]. Since random sampling is adopted to generate the training and test sets [1], all experiments are run ten times. The average and standard deviation of the obtained overall accuracies (OAs) are reported. Moreover, to avoid any possible experimental bias caused by random sampling, for each data set we first randomly obtain ten times splits and then apply the same ten splits on all experiments. All our experiments are conducted on a laptop with Matlab 2016, CPU (2.6GHz) and 16GB RAM, without any GPU acceleration. The LIBSVM library [66] (with default parameters setting) is used for the linear kernel based SVM. Our code will available online soon[2].

C. Parameters Setting

In the proposed method, there are two main parameters: the downsampled spatial dimension, s, and the number of feature maps after channel-wise average fusion, d. To avoid an exhausting search, s is simply set to the minimum spatial dimension among the used convolutional layers (i.e., s is set to 13 for the AlexNet and set to 14 for VGG-VD16). We have empirically tested that this configuration can always get satisfactory classification performance in our experiments. The parameter d is then analyzed and the UC data set with 80% samples randomly selected for training are used in this experiment. Fig. 6 shows the effect of using different values of d with the proposed MSCP method on the two considered CNN models. As can beobserved, on the AlexNet, when d grows from 50 to 80, there is an obvious improvement of OA, while further growing d will degrade the classification performance. Similar situation can be also observed on VGG-VD16 but with larger optimal d and the optimal d for VGG-VD16 is 130. The main reason maybe following two aspects. Firstly, a relative small d means more adjacent feature maps on the convolutional layer are fused together by the average operation, which could discard some useful information and thus decrease the classification performance. Secondly, a relative lager d will result in a covariance matrix with a larger dimension, which could weaken the discriminative ability and compactness of the covariance matrix, and therefore lead to relative worse classification results. The above parameters configuration are applied to all the test data sets and remained unchanged in the following experiments.

D. Effect of Combination of Different Layers

In this experiment, to demonstrate that the proposed MSCP method can fuse multi-layer feature maps effectively, a simplified version of MSCP are taken in consideration in this experiment, i.e., the single-layer feature maps of pre-trained CNN (plus CP), with and without preprocessing, and the resulting features are classified by a linear SVM (as the proposed MSCP). The UC data set and AID data set are used here for illustration purposes, where 80% training samples are randomly selected for UC data set and 20% training samples are randomly selected for AID data set. Table II shows the corresponding comparison results on the two data sets. As can be observed from Table II, there is a clear improvement of OA with the combination of different layers by the proposed MSCP. For example, on UC data set, the OAs of the preprocessed single-layer of VGG-VD16 are 96.93%, 97.29%, 96.76%, respectively, while the use of MSCP to combine these three layers increases the accuracy significantly (OA=98.36%). Similar results can also be observed on AID data sets. In addition, it can also be observed from Table II that the proposed MSCP method can outperform the original single-layer feature map (plus CP) with smaller feature dimension on both two test data sets. Moreover, Fig. 7 shows corresponding per-class accuracies obtained by MSCP and its simplified versions on the UC and AID data set. As can be observed in Fig. 7, by using the proposed MSCP strategy, the classification accuracies achieved for most classes exhibit obvious improvements on both two data sets. For example, on UC data set with AlexNet (see the first graph in the first column of Fig. 7), the accuracy obtained for the 14th class (*mobile homepark*) improves from 97% to 99.3%. The above conducted experiments suggest that there is indeed complementary information among different layers in the considered CNN architecture, and that the proposed

[1]http://www.vlfeat.org/matconvnet/
[2]https://sites.google.com/site/leyuanfang/home

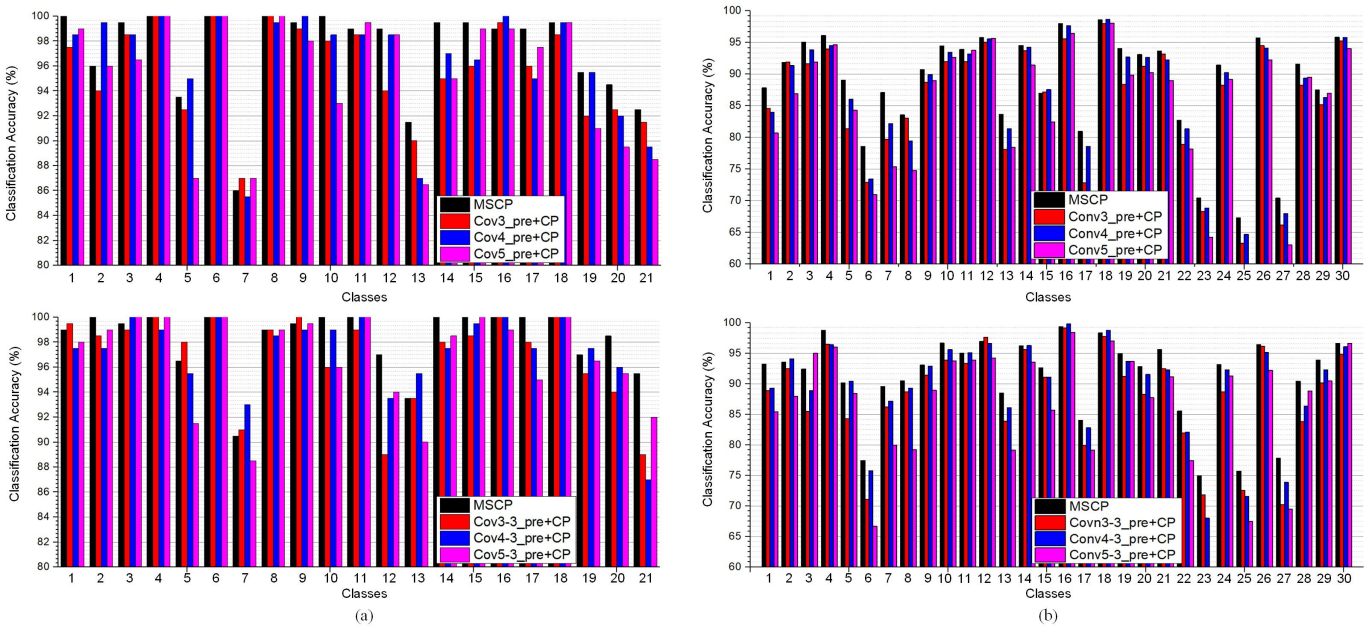

Fig. 7. Comparison of the per-class classification performance achieved by the proposed method and by a single-layer pre-trained CNN architecture (plus covariance pooling) with various pre-trained CNN models on different data sets: (a) UC data set (Training Ratio = 80%), (b) AID data set (Training Ratio = 20%). The first line is AlexNet and the second line is VGG-VD16. Here, '_pre' means that the convolutional layer has been pre-processed by downsampling and channel-wise average fusion.

MSCP can exploit such information to further improve the classification performance.

E. Compared with Other Pre-trained CNN based Methods

Firstly, the proposed MSCP method is compared to several pre-trained CNN model based classification methods on UC data set, including the two scenarios in [15], the method in [17], and the method in [18]. Specifically, in the first scenario of [15], only the last fully connected layer is used to represent the input image, and then classified by a linear SVM. In the second scenario of [15], only the last convolutional layer is encoded by some traditional coding methods (e.g., IFK) to represent the image, and then classified by a linear SVM. For the method in [17], the last two fully connected layers of the CNN are fused by means of discriminant correlation analysis (DCA) to represent the image, and then classified by a linear SVM. For the method in [18], a multiscale IFK strategy is adopted to fuse different layers of the CNN model for classification purposes. Table III shows the classification performance of the proposed method and the other four compared methods. As can beobserved in Table III, the MSCP (with pre-trained AlexNet and VGG-VD16) exhibits better classification performance than the methods in [15], [17]. Moreover, the MSCP with pre-trained VGG-VD16 is slower than the MSCP with pre-trained AlexNet. This is expected, since VGG-VG16 contains more layers than AlexNet and thus needs more time for forward propagation. Indeed, the method in [18] shows better classification performance than MSCP. However, the performances are similar and our method is almost 20 times faster. Fig 8 shows the confusion matrix for one of the experiments conducted by the proposed MSCP (with pre-trained VGG-VD16). As can beobserved in Fig 8,

TABLE III
COMPARISON OF THE CLASSIFICATION RESULTS (%) OBTAINED FOR THE UC DATA SET.

Method	OA	Times
Scenario (I) [15]	96.88±0.72	-
Scenario (II) [15]	96.90±0.77	-
DCA [17]	96.90±0.09	180
VGG_VD16+IFK [18]	98.57±0.34	14762
AleNet+MSCP (ours)	**97.29±0.63**	**134**
VGG_VD16+MSCP (ours)	**98.36±0.58**	**878**

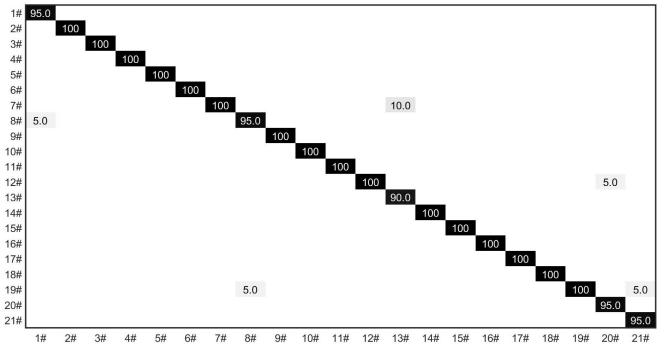

Fig. 8. Confusion matrix for the UC data set under a training rate of 80%, using the proposed MSCP method (with pre-trained VGG-VD16), in one single experiment (with OA = 98.57%).

the proposed method can get perfect classification performance on most classes except the following ones: 1#, 8#, 13#, 21# and 21#. It is worth noting that some of the misclassified images in this case may also be difficult to distinguish for a human interpreter (see Fig. 9).

We then compared the proposed method with several benchmark methods on AID data set: 1) a technique that uses the last

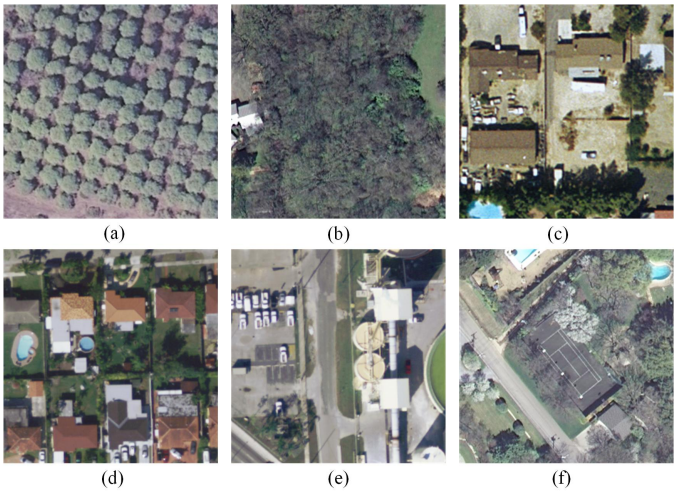

Fig. 9. Misclassified test images in one single experiment using the proposed MSCP method (with pre-trained VGG-VD16) on UC data set. (a) *agricultural* misclassified into (→) *forest*, (b) *forest* → *sparse residential*, (c) *medium residential* → *dense residential*, (d) *medium residential* → *dense residential*, (e) *storage tanks* → *intersection*, (f) *tennis court* → *sparse residential*.

TABLE IV
COMPARISON OF THE CLASSIFICATION RESULTS (%) OBTAINED FOR THE AID DATA SET

Method	Training ratio	
	20%	50%
VGG-VD16 [2]	86.59±0.29	89.64±0.36
GoogleNet [2]	83.44±0.40	86.39±0.55
Fusion by concatenation [17]	-	91.87±0.36
DCA [17]	-	89.71±0.33
AlexNet+MSCP (ours)	**88.99±0.38**	**92.36±0.21**
VGG-VD16+MSCP (ours)	**91.52±0.21**	**94.42±0.17**

fully connected layers of two pre-trained CNN models (i.e., VGG-VG16 and GoogleNet) as input to an SVM classifier [2], and 2) the two fusion methods proposed in [17] (i.e., fusion by concatenation and DCA) which only fuse the last two fully connected layers of the pre-trained CNN model. In addition, following the experimental setup in [2], two kinds of training rates are used for comparison. The first one considers 20% of the samples for training and the rest for testing. The other one considers 50% of the samples for training and the rest for testing. The corresponding results are given in Table IV. As can be seen in Table IV, the proposed method (both with pre-trained AlexNet and VGG-VD16) can obviously outperform the other tested methods, which demonstrates the effectiveness of the proposed MSCP approach. Moreover, Fig. 10 shows the confusion matrix obtained by the MSCP method (with pre-trained VGG-VD16) in one of the experiments conducted using a 50% training rate. As can be seen in Fig. 10, classes 6# (*center*), 23# (*resort*) and 25# (*school*) exhibit the lowest classification accuracies, which is mainly due to the fact that these classes usually share very similar objects (e.g., buildings) and thus making correct classification of these categories becomes very challenging. Some misclassified test images in this experiment are also given in Fig. 11 for illustrative purposes.

Last but not the least, we perform the proposed method on one of the largest and most challenging scene data set

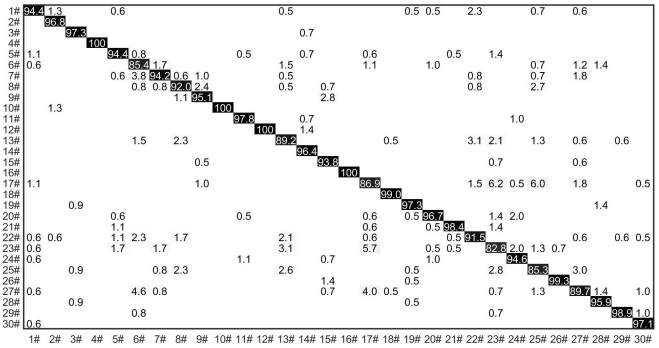

Fig. 10. Confusion matrix for the AID data set under a training rate of 50%, using the proposed MSCP method (with pre-trained VGG-VD16), in one of our experiments (with OA = 94.84%).

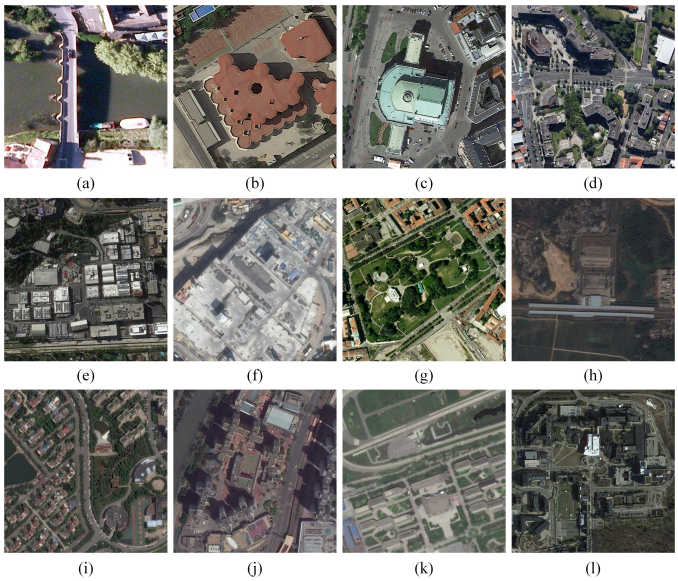

Fig. 11. Some examples of misclassified test images in one single experiment using the proposed MSCP method (with pre-trained VGG-VD16) on AID data set. (a) *bridge* misclassified into (→) *resort*, (b) *center* → *square*, (c) *church* → *square*, (d) *commercial* → *school*, (e) *industrial* → *school*, (f) *industrial* → *commercial*, (g) *park* → *square*, (h) *railway station* → *park*, (i) *resort* → *park*, (j) *school* → *commercial*, (k) *school* → *square*, (l) *school* → *commercial*.

(i.e.,NWPU data set) and make comparison with other two pre-trained CNN based methods [1] and [53]. The experimental setup is followed the description in [1], [53], two kinds of splits are used in this experiment. The first one considers 10% of samples for training and the rest for testing. The other one considers 20% of the samples for training and the rest for testing. Specifically, in [1], the last fully connected layer of three pre-trained CNN models (using AlexNet, VGG-VG16 and GoogleNet) is used to represent the image, and then the SVM is used for classification. In [53], the BoVW model is used to encode the last convolution layer, and the encoded features are also fed to an SVM for classification purposes. Table V reports the classification results obtained by all tested methods. It is clear that the proposed MSCP (with pre-trained VGG-VD16) can outperform the other methods with the two considered training rates.

TABLE V
COMPARISON OF THE CLASSIFICATION RESULTS (%) OBTAINED FOR THE NWPU DATA SET

Method	Training ratio	
	10%	20%
AlexNet [1]	76.69±0.21	79.85±0.13
VGG-VD16 [1]	76.47±0.18	79.79±0.15
GoogleNet [1]	76.19±0.38	78.48±0.26
AlexNet+BoVW [53]	55.22±0.39	59.22±0.18
VGG-VD16+BoVW [53]	82.65±0.31	84.32±0.17
AlexNet+MSCP (ours)	**81.70±0.23**	**85.58±0.16**
VGG-VD16+MSCP (ours)	**85.33±0.17**	**88.93±0.14**

F. MSCP Plus Multi-Resolutions Analysis to Further Improve Classification Accuracy

Here, we introduce a widely used method, i.e., multi-resolutions analysis (MSA) [21], [45], into our method to further improve the classification performance. The basic motivation using MSA in our method is based on the observation that most of the existing data sets contain a lot of image samples with large scales-variance. Fig. 12 shows some examples with large scales-variance from the NWPU data set. Specifically, in this experient, three different resolutions i.e., $\{227 \times 227 \times 3, 454 \times 454 \times 3, 908 \times 908 \times 3\}$ are considered for AlexNet and three different resolutions i.e., $\{224\times224\times3, 448\times448\times3, 896\times896\times3\}$ are considered for VGG-VD16. On the feature extraction stage, each image in the training set has been resize into the three different resolutions firstly. Then, the resized images are fed into the corresponding pre-trained CNN model and processed by the proposed MSCP method, successively. All the parameters (i.e.,the downsampled spatial dimension, s, and the number of feature map after channel-wise average fusion, d) are kept same to our previous experiment settings. As a result, three pooled features from different resolutions are obtained to represent the input image. The above operations are also conducted on each image in the test data set. It is worth noting that, the pooled features from different resolutions have the same dimension, which makes the images from different resolutions become comparable. On the train phase, all pooled features from different resolutions of the training set are used to train the SVM classifier. On the prediction phase, for each test image, the three pooled feature are classified by the SVM one by one. The final label of the test image is obtained based on the sum of the prediction probability from the three different resolutions.

The classification results achieved by our method with MRA and without MRA on the three test data sets (i.e., UC data set, AID data set, and NWPU data set) are reported in Table VI and are compared with two very recent related works [44] and [16]. Specifically, in [44], Zhu *et. al* propose a topical model based hand-crafted feature learning method, called sparse homogeneous-heterogeneous topic feature model (SHHTFM), for remote sensing scene image classification, which can simultaneously explore the homogeneous information (i.e., superpixels) and heterogeneous information (i.e., square window). In [16], by considering there are large variance within same class and high similarity among different classes, the authors introduce a discriminative loss term into pre-trained CNN models and then fine-tune whole

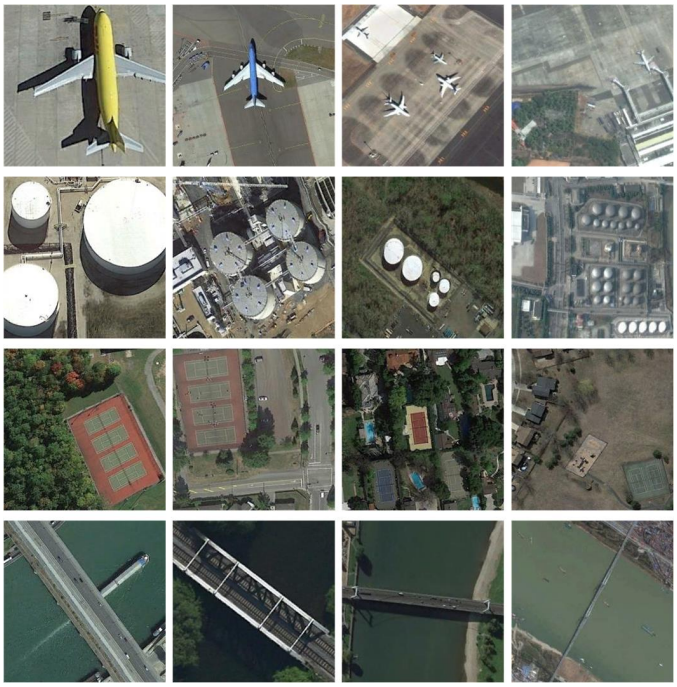

Fig. 12. Some examples with large scales-variance in NWPU data set. The first line is the category *airplane*, the second line the category *store tank*, the third line is the category *tennis court*, and the last line is the category *bridge*.

CNN model for end-to-end remote scene image classification, which called discriminative CNN model (DCNN). As can be observed from Table VI, our methods (VGG-VD16+MSCP and VGG-VD16+MSCP+MRA) are slightly better than the method in [44]. The reason is because that our methods use the deep CNN models for feature extraction, which can obtain more discriminative future than hand-crafted based feature learning method. In addition, our methods (with MRA) show competitive or better classification performance to the DCNN. For example, on the AID data set with training ratio is 20%, our methods (i.e., VGG-VD16+MSCP and VGG-VD16+MSCP+MRA) can outperform the DCNN. However, in most cases, DCNN shows slightly better classification results. This is mainly due to that DCNN can fine-tune the neural units (i.e., parameters) in the CNN models to match different images in different data sets and therefore achieve better classification results. In our future work, we will modify our method to a end-to-end classification framework and then adopt the fine-turning strategy to further improve the classification performance.

V. CONCLUSIONS AND FUTURE LINES

In this paper, we proposed a new method called multi-layer stacked covariance pooling (MSCP) to fuse the feature maps from different layers of a CNN architecture for remote sensing scene classification. The proposed MSCP based classification framework first performs feature extraction, using a pre-trained CNN model, and then performs feature fusion by covariance pooling. Since the proposed MSCP can take the second order information into consideration, more compact features are extracted for classification purposes. Moreover,

TABLE VI
COMPARISON OF THE CLASSIFICATION RESULTS (%) ACHIEVE BY OUR METHOD WITH MRA AND WITHOUT MRA AS WELL AS TWO VERY RECENT REMOTE SCENE CLASSIFICATION METHODS [44] [16]. 'TR' IS THE ABBREVIATION OF TRAINING RATIO

Method	UC	AID		NWPU	
	Tr=80%	Tr=20%	Tr = 50%	Tr=10%	Tr = 20%
SHHTFM [44]	98.33±0.98	-	-	-	-
DCNN [16]	98.93±0.10	90.82±0.16	96.89±0.10	89.22±0.50	91.89±0.22
AlexNet+MSCP (ours)	97.29±0.63	88.99±0.38	92.36±0.21	81.70±0.23	85.58±0.16
AlexNet+MSCP+MRA (ours)	97.32±0.52	90.65±0.19	94.11±0.15	83.31±0.23	87.05±0.23
VGG-VD16+MSCP (ours)	98.36±0.58	91.52±0.21	94.42±0.17	85.33±0.17	88.93±0.14
VGG-VD16+MSCP+MRA (ours)	98.40±0.34	92.21±0.17	96.56±0.18	88.07±0.18	90.81±0.13

each feature represents the covariance of two different feature maps, which captures the complementary information among different layers in a natural way. Our comprehensive experiments using three publicly available remote sensing image scene classification data sets, and the conducted comparisons with state-of-the-art approaches, verify the effectiveness of the proposed MSCP method. As a potential line of improvement, we realize that it can be useful to process the original layers with the MSCP approach. However, the dimension of features is difficult to manage taking into account the available off-the-shelf CNN models. In the future, we are planning to address this issue by designing a new end to end CNN model similar to the one presented in [28] with covariance pooling, which uses fewer feature maps in each of the layers.

ACKNOWLEDGMENT

The authors would like thank Prof. Cheng Gong and Prof. Guisong Xia for providing the NWPU and AID data sets on their homepages, respectively. In addition, we would like also to thank the editors and anonymous reviewers for their insightful comments and suggestions, which have significantly improved this paper.

REFERENCES

[1] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[2] G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[3] L. Fang, N. He, S. Li, P. Ghamisi, and J. A. Benediktsson, "Extinction profiles fusion for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1803–1815, Mar. 2018.

[4] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[5] B. Zhao, Y. Zhong, G. S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.

[6] 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.

[7] Q. Zhu, Y. Zhong, B. Zhao, G. S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[8] L. Huang, C. Chen, W. Li, and Q. Du, "Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors," *Remote Sens.*, vol. 8, no. 6, Jun. 2016.

[9] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens*, vol. 10, no. 6, pp. 2889–2901, Jun. 2017.

[10] G. Cheng, J. Han, L. Guo, and T. Liu, "Learning coarse-to-fine sparselets for efficient object detection and scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1173–1181.

[11] J. Zou, W. Li, C. Chen, and Q. Du, "Scene classification using local and global features with collaborative representation fusion," *Inf. Sci.*, vol. 348, pp. 209–226, Jun. 2016.

[12] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548 – 556, Jan. 2017.

[13] 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.

[14] W. Zhang, X. Lu, and X. Li, "A coarse-to-fine semi-supervised change detection for multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3587–3599, Jun. 2018.

[15] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14 680–14 707, 2015.

[16] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[17] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.

[18] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[19] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Conf. Adv. Neural Inform. Process. Syst.*, 2012, pp. 1097–1105.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[23] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[25] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5188–5196.

[26] L. Yann, B. Yoshua, and H. Geoffrey, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

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

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

[29] Y. Yang and S. Newsam, "Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery," in *Proc. IEEE Int. Conf. Image Proces.*, 2008, pp. 1852–1855.

[30] ——, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.
[31] M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, "Land-use classification with compressive sensing multifeature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2155–2159, Oct. 2015.
[32] Y. Feng, Y. Yuan, and X. Lu, "A non-negative low-rank representation for hyperspectral band selection," *Int. J. Remote Sens.*, vol. 37, no. 19, pp. 4590–4609, Aug. 2016.
[33] X. Lu, H. Wu, Y. Yuan, P. Yan, and X. Li, "Manifold regularized sparse nmf for hyperspectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2815–2826, May 2013.
[34] X. Lu, Y. Yuan, and X. Zheng, "Joint dictionary learning for multispectral change detection," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 884–897, Apr. 2017.
[35] L. Fang, H. Zhuo, and S. Li, "Super-resolution of hyperspectral image via superpixel-based sparse representation," *Neurocomputing*, vol. 273, pp. 171–177, Jan. 2018.
[36] L. Fang, S. Li, D. Cunefare, and S. Farsiu, "Segmentation based sparse reconstruction of optical coherence tomography images," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 407–421, Feb. 2017.
[37] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2005, pp. 886–893.
[39] K. Qi, H. Wu, C. Shen, and J. Gong, "Land-use scene classification in high-resolution remote sensing images using improved correlatons," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2403–2407, Dec. 2015.
[40] B. Zhao, Y. Zhong, and L. Zhang, "A spectral-structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 73–85, Jun. 2016.
[41] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1465–1472.
[42] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.
[43] B. Zhao, Y. Zhong, G. S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.
[44] Q. Zhu, Y. Zhong, S. Wu, L. Zhang, and D. Li, "Scene classification based on the sparse homogeneous-heterogeneous topic feature model," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2689–2703, May 2018.
[45] X. Lu, X. Li, and L. Mou, "Semi-supervised multitask learning for scene recognition," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1967–1976, Sep. 2015.
[46] Z. Du, X. Li, and X. Lu, "Local structure learning in high resolution remote sensing image retrieval," *Neurocomputing*, vol. 207, pp. 813 – 822, Sep. 2016.
[47] F. Hu, G. S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.
[48] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.
[49] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2014, pp. 1717–1724.
[50] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2014, pp. 580–587.
[51] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," *arXiv*, vol. 1508, pp. 1–11, Aug. 2015.
[52] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Proc. Europ. Conf. Comput. Vis.*, 2010, pp. 143–156.
[53] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.
[54] C. Ionescu, O. Vantzos, and C. Sminchisescu, "Matrix backpropagation for deep networks with structured layers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2965–2973.
[55] 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.
[56] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition?" in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2089–2097.
[57] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 589–600.
[58] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on riemannian manifolds," *IEEE Trans. on Pattern. Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.
[59] L. Fang, N. He, S. Li, A. J. Plaza, and J. Plaza, "A new spatial spectral feature extraction method for hyperspectral images using local covariance matrix representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3534–3546, Jun. 2018.
[60] B. Julesz, E. Gilbert, L. Shepp,and H. Frisch, "Inability of humans to discriminate between visual textures that agree in second-order statistics-revisited," *Perception*, vol. 2, no. 4, pp. 391–405, Dec. 1973.
[61] P. Li, J. Xie, Q. Wang, and Z. Gao, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, to be published, 2018.
[62] T. Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.
[63] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM J. Matrix Analysis and Appl.*, vol. 29, no. 1, pp. 328–347, Feb. 2007.
[64] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic Resonance in Medicine*, vol. 56, no. 2, pp. 411–421, 2006.
[65] A. Vedaldi and K. Lenc, "Matconvnet-convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[66] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

Nanjun He (S'17) received the B.S. degree from Central South University of Forestry and Technology, Changsha, China, in 2013. He is currently working toward the Ph.D. degree in the Laboratory of Vision and Image Processing, Hunan University, Changsha, China.

From October 2017 to October 2018, he is a visiting PH.D. student with Hyperspectral Computing Laboratory, University of Extremadura, Cáceres, Spain, supported by the China Scholarship Council. His research interests include remote sensing image classification, remote sensing object detection.

Leyuan Fang (S'10-M'14-SM'17) received the Ph.D. degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2015.

From September 2011 to September 2012, he was a visiting Ph.D. student with the Department of Ophthalmology, Duke University, Durham, NC, USA, supported by the China Scholarship Council. From August 2016 to September 2017, he was a Postdoctoral Research Fellow with the Department of Biomedical Engineering, Duke University, Durham, NC, USA. Since Jan. 2017, he has been an associate professor with the College of Electrical and Information Engineering, Hunan University. His research interests include sparse representation and multi-resolution analysis in remote sensing and medical image processing. He has won the Scholarship Award for Excellent Doctoral Student granted by Chinese Ministry of Education in 2011.

Shutao Li (M'07-SM'15) received the B.S., M.S., and Ph.D. degrees from Hunan University, Changsha, China, in 1995, 1997, and 2001, respectively. He was a Research Associate with the Department of Computer Science, the Hong Kong University of Science and Technology, Hong Kong, in 2011. From 2002 to 2003, he was a Post-Doctoral Fellow with the Royal Holloway College, University of London, London, U.K., with Prof. John Shawe-Taylor. In 2005, he visited the Department of Computer Science, Hong Kong University of Science and Technology as a Visiting Professor. He joined the College of Electrical and Information Engineering, Hunan University, in 2001. He is currently a Full Professor with the College of Electrical and Information Engineering, Hunan University. He has authored or co-authored over 160 refereed papers. His current research interests include compressive sensing, sparse representation, image processing, and pattern recognition.

He is now an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, and a member of the Editorial Board of the Information Fusion and the Sensing and Imaging. He was a recipient of two 2nd-Grade National Awards at the Science and Technology Progress of China in 2004 and 2006.

Antonio Plaza (M'05-SM'07-F'15) is the Head of the Hyperspectral Computing Laboratory at the Department of Technology of Computers and Communications, University of Extremadura, where he received the M.Sc. degree in 1999 and the PhD degree in 2002, both in Computer Engineering. His main research interests comprise hyperspectral data processing and parallel computing of remote sensing data. He has authored more than 600 publications, including more than 200 JCR journal papers (more than 150 in IEEE journals), 24 book chapters, and over 300 peer-reviewed conference proceeding papers. He has guest edited 10 special issues on hyperspectral remote sensing for different journals. Prof. Plaza is a Fellow of IEEE "for contributions to hyperspectral data processing and parallel computing of Earth observation data." He is a recipient of the recognition of Best Reviewers of the IEEE Geoscience and Remote Sensing Letters (in 2009) and a recipient of the recognition of Best Reviewers of the IEEE Transactions on Geoscience and Remote Sensing (in 2010), for which he served as Associate Editor in 2007-2012. He is also an Associate Editor for IEEE Access, and was a member of the Editorial Board of the IEEE Geoscience and Remote Sensing Newsletter (2011-2012) and the IEEE Geoscience and Remote Sensing Magazine (2013). He was also a member of the steering committee of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS). He is a recipient of the Best Column Award of the IEEE Signal Processing Magazine in 2015, the 2013 Best Paper Award of the JSTARS journal, and the most highly cited paper (2005-2010) in the Journal of Parallel and Distributed Computing. He received best paper awards at the IEEE International Conference on Space Technology and the IEEE Symposium on Signal Processing and Information Technology. He served as the Director of Education Activities for the IEEE Geoscience and Remote Sensing Society (GRSS) in 2011-2012, and as President of the Spanish Chapter of IEEE GRSS in 2012-2016. He has reviewed more than 500 manuscripts for over 50 different journals. He is served as the Editor-in-Chief of the IEEE Transactions on Geoscience and Remote Sensing journal for five years (2013-2017). Additional information: http://www.umbc.edu/rssipl/people/aplaza.

Javier Plaza (M'09-SM'15) is a member of the Hyperspectral Computing Laboratory at the Department of Technology of Computers and Communications, University of Extremadura, where he received the M.Sc. degree in 2004 and the PhD degree in 2008, both in Computer Engineering. He was the recipient of the Outstanding Ph.D. Dissertation Award at the University of Extremadura in 2008. His main research interests comprise hyperspectral data processing and parallel computing of remote sensing data. He has authored more than 150 publications, including over 50 JCR journal papers, 10 book chapters, and 90 peer-reviewed conference proceeding papers. He has guest edited 3 special issues on hyperspectral remote sensing for different journals. He is an Associate Editor for IEEE Geoscience and Remote Sensing Letters and an Associate Editor of the IEEE Remote Sensing Code Library. He is a recipient of the Best Column Award of the IEEE Signal Processing Magazine in 2015 and the most highly cited paper (2005-2010) in the Journal of Parallel and Distributed Computing. He received best paper awards at the IEEE International Conference on Space Technology and the IEEE Symposium on Signal Processing and Information Technology. Additional information: http://www.umbc.edu/rssipl/people/jplaza.