# Deep Feature Fusion for VHR Remote Sensing Scene Classification

Souleyman Chaib, *Student Member, IEEE*, Huan Liu, *Student Member, IEEE*, Yanfeng Gu, *Senior Member, IEEE*, and Hongxun Yao, *Member, IEEE*

***Abstract*— The rapid development of remote sensing technology allows us to get images with high and very high resolution (VHR). VHR imagery scene classification has become an important and challenging problem. In this paper, we introduce a framework for VHR scene understanding. First, the pretrained visual geometry group network (VGG-Net) model is proposed as deep feature extractors to extract informative features from the original VHR images. Second, we select the fully connected layers constructed by VGG-Net in which each layer is regarded as separated feature descriptors. And then we combine between them to construct final representation of the VHR image scenes. Third, discriminant correlation analysis (DCA) is adopted as feature fusion strategy to further refine the original features extracting from VGG-Net, which allows a more efficient fusion approach with small cost than the traditional feature fusion strategies. We apply our approach to three challenging data sets: 1) UC MERCED data set that contains 21 different areal scene categories with submeter resolution; 2) WHU-RS data set that contains 19 challenging scene categories with various resolutions; and 3) the Aerial Image data set that has a number of 10 000 images within 30 challenging scene categories with various resolutions. The experimental results demonstrate that our proposed method outperforms the state-of-the-art approaches. Using feature fusion technique achieves a higher accuracy than solely using the raw deep features. Moreover, the proposed method based on DCA fusion produces good informative features to describe the images scene with much lower dimension.**

***Index Terms*— Discriminant correlation analysis (DCA), features fusion, scene classification, unsupervised features learning.**

## I. Introduction

RECENTLY, the fast development of remote sensing instruments has provided us the opportunities to capture high and very high resolution (VHR) images. The research for appropriate and efficient methods for scene presentation and classification becomes a critical task in the remote sensing image community. The high dimensionality and the statistical characteristics of these acquired images are the most challengeable acts for VHR scene understanding [1]. In the last decade, several methods were proposed in order to analyze and classify the scene of VHR images. One of the most popular methods used for scene classification is bag of visual words (BoVW) [2], which is developed for feature encoding. The weakness of the BoVW technique in remote sensing imagery is that it ignores the spatial and spectral information. Many methods are developed to overcome the BoVW limitation. The spatial pyramid matching kernel approach based on the approximate global geometric correspondence was proposed in [3]. Principal component analysis (PCA) [4] is widely used for feature dimension reduction. Xia *et al.* [5] proposed a new method based on two models for hyperspectral image classification: 1) supervised probabilistic PCA and 2) semisupervised probabilistic PCA. Recently, unsupervised feature learning is a hot topic for VHR image scene classification. Cheriyadat [6] proposed a recent approach for unsupervised feature coding, which extracted a dense scale invariant feature transform (SIFT) as a feature descriptor, and then generated a new sparse feature descriptor based on sparse coding (SC). This method outperforms the previous results with high accuracy. Lu *et al.* [7] introduced a new formwork by analyzing the images scene in different scales to construct multiresolution features, and then constructed a model of sparse features selection-based manifold regularization. However, these methods need to extract low-level feature descriptors such as SIFT, which ignore the information from the original data. Due to the limit of the aforementioned methods, Zhang *et al.* [8] developed an efficient framework based on two fundamental steps: 1) patches sampling and 2) a sparse autoencoder for unsupervised feature learning. In the first step, they proposed a saliency detection method named context-aware saliency as a sampling strategy to extract representative patches from the VHR images. In the second step, each patch was represented by the pixel intensity values; the major limit of this method is that it ignores the semantic meaning of the VHR images.

To overcome this limit, recently, Cheng *et al.* [9] have proposed two novel methods for scene understanding and scene representation, which are named multiclass geospatial object detection and geographic image classification based on the collection of part detectors. And then they introduced an efficient midlevel visual elements oriented land-use classification method based on "partlets" for VHR remote sensing images in [10], where the two proposed methods are based on part detection and they achieve a high accuracy on the public UC Merced data set.

In the last decade, deep learning-based approaches have been developed in VHR mage scene classification. Vaduva *et al.* [11] presented a deep learning algorithm for

Manuscript received November 1, 2016; revised January 11, 2017 and March 27, 2017; accepted April 28, 2017. Date of publication May 25, 2017; date of current version July 20, 2017. This work was supported by the Natural Science Foundation of China under Grant 61371180 and Grant 61522107. *(Corresponding author: Yanfeng Gu.)*

S. Chaib and H. Yao are with the School of Computer Science, Harbin Institute of Technology, Harbin 150001, China (e-mail: chaib@hit.edu.cn).

H. Liu and Y. Gu are with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: guyf@hit.edu.cn).

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

Digital Object Identifier 10.1109/TGRS.2017.2700322

VHR image scene classification. The research in [12] proposed a hybrid deep convolutional neural network (CNN) for vehicle detection from satellite images. And the research in [13] introduced the stacked autoencoder to extract spatial informative features for hyperspectral images classification. Zou *et al.* [14] proposed deep learning to select informative features for remote sensing scene classification. The research in [1] introduced a recent method based on CNN, which combined the greedy layer-wise unsupervised pretraining [15] and the efficient enforcing population and lifetime sparsity algorithm [16] to learn sparse feature representation from the satellite images. Despite the robustness and efficiency of deep learning network, the aforementioned methods did not overcome the highest accuracy.

CNN is one of the most successful deep learning methods due to its remarkable performance on ImageNet large-scale visual recognition competition. The success of CNN is due to its capacity to learn hierarchical representation to describe the image scene. Most recently, CNN has been developed in the context of remote sensing image analysis, which has become more and more popular for image scene representation. In [17], a deep ensemble framework based on CNN was proposed for scene classification. Unfortunately, it is difficult to train a new CNN because it needs very large labeled data set and consumes a long cost time [18]. Some literatures have demonstrated that CNN can facilitate transfer learning between different domains and work remarkably well. Moreover, due to the small size of training samples in remote sensing scene classification task, it is hard to fully train CNN model. Therefore, pretrained CNNs have been transferred for scene classification. There are two transferring ways for scene classification. One is to directly transfer pretrained CNNs as feature extractors with fixed parameters of CNNs [19]–[21]. The other is to use pretrained CNNs and then fine-tune them on VHR data set [22], [23]. Penatti *et al.* [20] show that pretrained CNNs can be used to recognize everyday objects and generalize well to classify remote sensing scenes. Moreover, more strategies based on pretrained CNNs have been developed to form better representation for scene classification. Othman *et al.* [19] explored a new method for VHR scene classification by coding the convolutional features by sparse autoencoder to represent the image scene. The research in [22] proposed an effective method for optical object detection by learning rotation invariant CNN. Marmanis *et al.* [24] used the pretrained OverFeat model as a feature extractor, and then transfer features into a supervised CNN for classification. Hu *et al.* [21] introduced two scenarios based on the pretrained CNN for VHR images scene classification: 1) the second fully connected layer is regarded as a final feature descriptor of the scene images and 2) extract dense features from the last convolutional layer at multiscales and then encode the dense features into global representation by feature coding approaches.

The efficient strategy to overcome the aforementioned limits is feature fusion. Sheng *et al.* [25] introduced the fusion of the color histogram and the SIFT features. The research in [26] and [27] combined more than three feature descriptors to represent the images. Recently, the research in [28] introduced the combination between the spectral and structural information of VHR images. The spectral information was represented by the first-order and second-order statistics extracted from the set of sampled patches. And the dense SIFT feature descriptor was used as the structural information. More recently, the research in [29] proposed the combination of the local and global features to calculate the BoVW. How to get good feature descriptors to represent the VHR images for scene classification is still a critical task for VHR image scene understanding. According to the success of pretrained CNNs in computer vision fields, and the great success of visual geometry group network (VGG-net) in the feature extraction task, we introduce a new framework based on pretrained VGG-Net model to automatically learn the feature descriptor for the VHR images.

As mentioned before, the feature fusion is an efficient step in scene understanding. We proposed to combine between the outputs of CNN algorithm, where the final features can effectively represent the scene images. In order to reduce dimension of features and use an adequate method for feature fusion, we propose to use the discriminant correlation analysis (DCA) in this paper. The major contributions of this paper are threefolds.

目的

本文提出的方法是DCA

1) We employ the pretrained deep CNN models for VHR images scene classification, where we used VGG-Net as the feature extractor by selecting useful layers in order to get a good representation of the images scene. It allowed us to achieve superior results compared with the previous works.
2) We are the first to combine between different full connected layers of VGG-Net model, where the output of each layer is supposed as a feature descriptor and combined to construct final feature representation of the input image. The fused deep feature learning performs better than other feature representation methods such as SIFT, speeded up robust features (SURF), and histogram of oriented gradients (HOG), and current methods based on pretrained CNNs.
3) We introduced DCA to represent the fused features in a very low dimension, which allowed achieving a good classification performance and speeding up the classification task.

The remainder of this paper is organized as follows. Section II describes the proposed framework and the processing way to extract and fuse the feature descriptor for VHR image scene classification. First, this section briefly introduces the pretrained deep feature learning model to automatically extract features from the VHR images. Second, it presents the fusion method used to combine the features extracted by CNN. We use the features concatenation as a fused method. Besides, DCA is introduced as the future fusion and dimension reduction method. The experimental results on several databases are presented in Section III. Finally, Section IV lists the conclusions along with the discussion of the results.

## II. Method Description

In this section, we describe the proposed framework for VHR scene classification using deep VGG-Net as a feature extractor to describe VHR image scene with representative

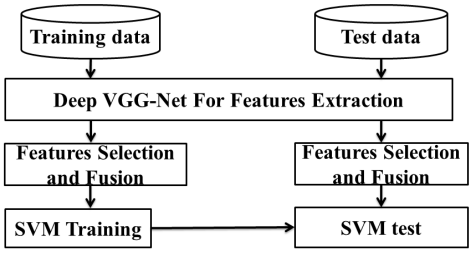


Fig. 1. Overall architecture of the proposed method.

features and DCA as a feature fusion method. Our proposed approach consists of the following three main steps as shown in Fig. 1: 1) feature extraction with deep VGG-Net; 2) using the DCA approach to fuse the extracted features; 3) using a support vector machine (SVM) as a classifier for VHR scene classification.

### *A. Deep VGG-Net for Feature Extraction*

In the last decade, several CNN models have been developed for large scale image classification and object detection such as AlexNet [30], R-CNN [31], CaffeNet [32], and VGG-Net [31]. In contrast to the most CNNs that usually have five or seven layers, the proposed framework is based on the VGG-Net for feature extraction. VGG-Net has a much deeper architecture (up to 19 weight layers) and hence can provide much informative features.

The VGG-Net outperforms the previous generation of CNN models, which is trained with the public ImageNet data set and achieves the state-of-the-art results on classification challenges [33]. Different from the most scene classification methods, which are based only on low-level features such as SIFT, SURF, HOG, or deep learned features, our framework is based on the fusion of features learned by the VGG-Net model. Then the outputs of some selected layers are supposed as a feature descriptor of the input, in the aim to describe the images scene by informative and significant features. The deep VGG-Net facilitates the feature extraction from different layers to describe the VHR image scene with informative features. Fig. 2 describes the architecture of VGG-net used in our framework. It comprises five convolutional layers, where each one followed by a pooling layer, and three fully connected layers, where the input images are resized to $224 \times 224 \times 3$ (more details are in [34]). In this paper, we used the first and second output fully connected layers of the pretrained net as the feature descriptor of the image scene for VHR image classification.

### *B. Features Fusion Based on Discriminant Correlation Analysis*

Features fusion for VHR images scene classification is a robust and efficient strategy, where the fused features contain rich information to describe the image scene well. The aim of feature fusion for VHR image scene classification is to combine two or more relevant features extracted from the images scene into a single feature vector with more discriminant information than the input feature vectors.

How to properly combine two or more features is becoming a critical task. Several methods have been developed in the literature for feature fusion, in the aim to get more informative descriptor to represent the images scene. We found two famous strategies for feature fusion developed.

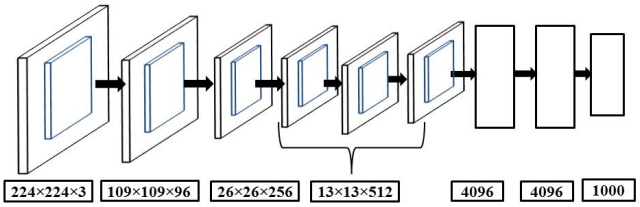


Fig. 2. VGG-net architecture used in this paper. The boxes show the size of each feature layer, fully connected layers, and the size of the output.

1) Serial feature fusion [35] that simply concatenates two set of features into one single feature. It supposes that $x$ and $y$ are two features extracted form an input image with $p$, $q$ vector dimension, respectively, and then the fused feature is $z$ with size equal to $(p+q)$.

2) Parallel strategy [36], [37] that combines the two features vector into a complex vector $z = x + iy$ where $i$ is the imaginary unit.

Sun *et al.* [38] have introduced a recent method for feature fusion based on canonical correlation analysis (CCA) in the area of image recognition. Feature fusion based on CCA uses the correlation between two sets of features to calculate two sets of transformation where the transformed features have higher correlation than the two feature sets. Several methods are improved for feature fusion based on CCA in [39] and [40].

Suppose that $X \in R^{p\times n}$ and $Y \in R^{q\times n}$ are two features matrices, where $n$ denotes the number of trained features and $p$ and $q$ represent the feature dimension of $X$ and $Y$, respectively.

Let $S_{xx} \in R^{p\times p}$ and $S_{yy} \in R^{q\times q}$ represent the covariance matrices of $X$ and $Y$, respectively, and $S_{xy} \in R^{p\times q}$ is the between-sets covariance matrix, where $S_{yx} = S_{xy}^T$.

The overall covariance matrix $S \in R^{(p+q)\times(p+q)}$ is then computed

$$S = \begin{pmatrix} \text{cov}(X) & \text{cov}(x,y) \\ \text{cov}(y,X) & \text{cov}(y) \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}. \tag{1}$$

As mentioned in [41], it is hard to understand the relationship between these two features from matrix $S$. The goal of CCA is to define a linear combination $X^* = W_x^T X$ and $Y^* = W_y^T Y$, which maximizes the pair-wise correlation across the two feature sets

$$\text{corr}(X^*, Y^*) = \frac{\text{cov}(X^*, Y^*)}{\text{var}(X^*) \cdot \text{var}(Y^*)} \tag{2}$$

where $\text{cov}(X^*, Y^*) = W_x^T S_{xy} W_y$, $\text{var}(X^*) = W_x^T S_{xx} W_x$, and $\text{var}(Y^*) = W_y^T S_{yy} W_{xy}$. The maximization problem of the covariance between $X^*$ and $Y^*$ is solved in [41] using Lagrange multipliers to satisfy the flowing constraint $\text{var}(X^*) = \text{var}(Y^*) = 1$.

As developed in [38], the combination between the transformed features is performed by concatenation or summation as follows:

$$Z = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_X & 0 \\ 0 & W_Y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{3}$$

or

$$Z = X^* + Y^* = \begin{pmatrix} W_X \\ W_Y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{4}$$

where $Z$ is the canonical correlation discriminant features.

CCA的主要限制是它忽略了图像数据集之间的类结构之间的关系这也为下文提出DCA做了铺垫

However, the major limitation of CCA is that it ignores the relation between class structures among the images data sets, where we are interested to maximize the correlation between the feature sets and also need to separate the classes within each set of features.

More recently, a great solution has been presented to overcome the CCA weakness [40], which introduced the DCA. DCA maximizes the correlation between features across the two feature sets and in the same time maximizes the difference between classes.

Let $D$ denote the images data set, and $c$ is the number of categories belong to $D$, where $X$ represents the data matrix of the set of features collected from the image data set $D$. $x_{ij} \in X$ denotes the features vector extracted from the $i$th image of the $j$th category. Suppose that $\bar{x_i}$ and $\bar{x}$ denote the mean of the $x_{ij}$ vector under the $i$th class in the whole data matrix $X$

$$\bar{x_i} = \frac{1}{n_i} \sum_{j=1}^{n} x_{ij} \tag{5}$$

where $n$ represents the number of images that belong to the $i$th category

$$\bar{x} = \frac{1}{n_i} \sum_{i=1}^{c} n_i \bar{x_i}. \tag{6}$$

The between-class scatter matrix is defined in [42] as follows:

$$S_{bx_{(p\times q)}} = \Phi_{bx} \Phi_{bx}^T \tag{7}$$

where

$$\Phi_{bx_{(p\times c)}} = \left[\sqrt{n}_1 (\bar{x}_1 - \bar{x}), \sqrt{n}_2 (\bar{x}_2 - \bar{x}), \ldots, \sqrt{n}_c (\bar{x}_c - \bar{x})\right]. \tag{8}$$

As defined in [41], the significant eigenvectors of $(\Phi_{bx}\Phi_{bx}^T)_{p\times p}$ can be computed by mapping the eigenvector of $(\Phi_{bx}^T \Phi_{bx})_{c\times c}$. To well separate the classes, matrix $\Phi_{bx}^T \Phi_{bx}$ should be a diagonal matrix [42] as follows:

$$P^T (\Phi_{bx}^T \Phi_{bx}) P = \vec{\bigwedge} \tag{9}$$

where $P$ is the orthogonal eigenvectors matrix, $\vec{\bigwedge}$ is the diagonal matrix of real nonnegative eigenvalues, and $r$ is the largest nonzero eigenvectors matrix that is represented by $\varphi_{(c\times r)}$. For matrix $P$, we have

$$\varphi^T \left(\Phi_{bx}^T \Phi_{bx}\right) \varphi = \bigwedge_{(r\times r).} \tag{10}$$

As proposed in [41], the significant eigenvectors of $S_{bx}$ can be calculated by mapping $\Phi \to \Phi_{bx}\varphi$ as follows:

$$(\Phi_{bx}\varphi)^T S_{bx} (\Phi_{bx}\varphi) = \bigwedge_{(r\times r).} \tag{11}$$

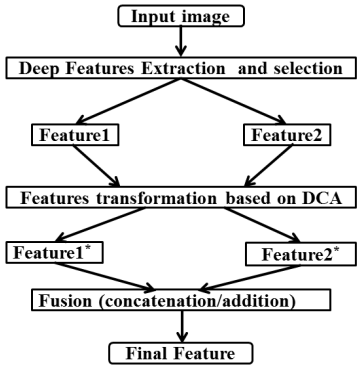


Fig. 3. Schematic of deep features fusion based on DCA.

$w_{bx} = \Phi_{bx}\varphi \bigwedge^{-1/2}$ is the transformation of the data matrix $X$ that unitizes $S_{bx}$ and reduces the dimensionality of $X$ from $p$ to $r$ where

$$W_{bx}^T S_{bx} W_{bx} = I \tag{12}$$

$$X'_{(r\times n)} = W^T_{bx_{(r\times p)}} X_{(p\times n)}. \tag{13}$$

In the same way as the second feature set $Y$, the transformation matrix is computed $W_{by}$, which unitizes $S_{by}$ and reduces the dimensionality of $Y$ from $q$ to

$$W_{by}^T S_{by} W_{by} = I \tag{14}$$

or

$$Y'_{(r\times n)} = W^T_{by_{(r\times q)}} Y_{(q\times n)} \tag{15}$$

where $r$ is the feature length of the transformed features

$$r \le \min\left(c - 1, \text{rank}(X), \text{rank}(Y)\right). \tag{16}$$

Now, we need to maximize the correlation between features across the two feature sets. To overcome this problem, we need to diagonalize the between-set covariance matrix of transformed feature sets, $S'_{xy} = X' Y'^T$. The research in [40] proposed singular value decomposition to achieve this problem

$$S'_{xy_{(r\times r)}} = U \sum V^T U^T S'_{xy} V = \sum. \tag{17}$$

Letting $W_{cx} = U \sum^{-1/2}$ and $W_{cy} = V \sum^{-1/2}$, we have

$$\left(U \sum^{-1/2}\right)^T S'_{xy} (V \sum^{-1/2}) = I \tag{18}$$

which unitizes matrix $S'_{xy}$. The transformation of the feature set is

$$X^* = W_{cx}^T X' = W_{cx}^T W_{bx}^T X = W_x X \tag{19}$$

$$Y^* = W_{cy}^T Y' = W_{cy}^T W_{by}^T X = W_y X. \tag{20}$$

In the same way, as DCA method, the feature fusion can be performed by summation or concatenation, where the summation method allows a lower dimension of the fused features with less change in the accuracy.

Fig. 3 describes the procedure followed for deep features fusion based on DCA method, where we use VGG-net to extract deep features from the input VHR image. Then,

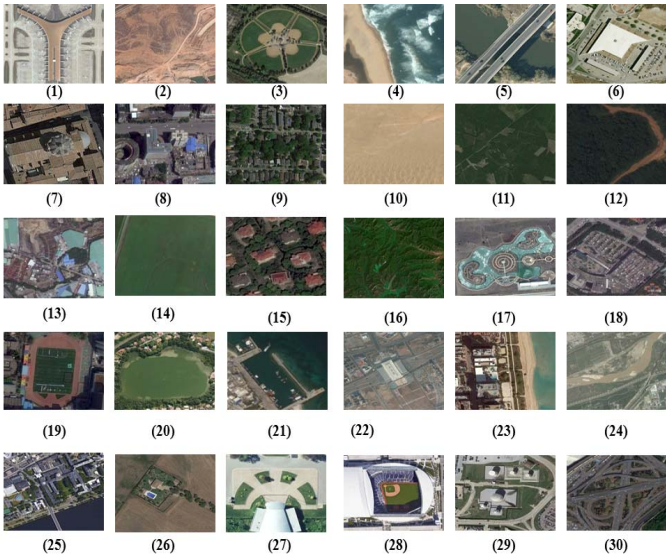


Fig. 4. Example images associated with 30 land-use categories in AID: (1) Airport, (2) Bare land, (3) Baseball field, (4) Beach, (5) Bridge, (6) Center, (7) Church, (8) Commercial, (9) Dense residential, (10) Desert, (11) Farmland, (12) Forest, (13) Industrial, (14) Meadow, (15) Medium residential, (16) Mountain, (17) Park, (18) Parking, (19) Playground, (20) Pond, (21) Port, (22) Railway station, (23) Resort, (24) River, (25) School, (26) Sparse residential, (27) Square, (28) Stadium, (29) Storage tanks, and (30) Viaduct.高架桥


Fig. 5. Example images associated with 21 land-use categories in the UC Merced data set: (1) Agricultural, (2) Airplane, (3) Baseball diamond, (4) Beach, (5) Building, (6) Chaparral, (7) Dense residential, (8) Forest, (9) Freeway, (10) Golf course, (11) Harbor, (12) Intersection, (13) Medium residential, (14) Mobile-homepark, (15) Overpass, (16) Parking lot, (17) River, (18) Runway, (19) Sparse residential, (20) Storage tanks, and (21) Tennis court.


Fig. 6. Example images associated with 19 land-use categories in the RS data set: (1) Airport, (2) Beach, (3) Bridge, (4) Commercial, (5) Desert, (6) Farmland, (7) Football field, (8) Forest, (9) Industrial, (10) Meadow, (11) Mountain, (12) Park, (13) Parking, (14) Pond, (15) Port, (16) Railway station, (17) Residential, (18) River, and (19) Viaduct.

we select two feature sets from two fully connected layers. We calculate the new transformation of those features based on DCA. After that, we combine (concatenation/addition) the transformed features to represent the input images by single informative features.

## III. Experimental Result and Setup

The proposed method is tested on three different data sets. The description of the data sets has been given in the following section, where we also analyze the parameters of the proposed method. And then the results for each data set are discussed.

### A. Data Sets

Three different data sets are used in this paper. The first data set was acquired from a large Google Earth satellite, which has 10 000 images distributed on 30 aerial scene categories [43]. The spatial resolution of this image is varied from 8 m to about half a meter. The large size of each image scene is 600 × 600 pixels as shown in Fig. 4. There are 30 classes of training images: Airport, Bare land, Baseball field, Beach, Bridge, Center, Church, Commercial, Dense residential, Desert, Farmland, Forest, Industrial, Meadow, Medium residential, Mountain, Park, Parking, Playground, Pond, Port, Railway station, Resort, River, School, Sparse residential, Square, Stadium, Storage tanks, and Viaduct. As cited in [43], the most recent works of VHR image classification on the UC-Merced and WHU-RS data sets are achieved a very high accuracy due to the small size of these data sets. Xia *et al.* [43] construct the Aerial Image data set (AID) in the aim to renew the challenge of VHR remote sensing image classification. The second data set chosen for evaluation is the public UC Merced data set. Fig. 5 shows some images representing different aerial scenes taken from this data set. Each image contains 256 × 256 pixels and 1 ft/pixel. The data set contains 21 challenging scene categories with 100 samples for each class, which is available to be downloaded from the U.S. Geological Survey national map [44]. The data set represents highly overlapping classes such as sparse residential, medium residential, and dense residential, which mainly differ in the density of the structures.

The third data set named WHU-RS was exported from Google Earth [45]. It contains 19 challenging scene categories with various resolutions. There are 950 samples with a size of 600 × 600 pixels. Fig. 6 shows some different images scene taken from this data set.

第一个数据集30个分类600*600像素空间分辨率达到0.5--8米之间

第二个数据集来自于UC Merced

第三个数据集来自于WHU-RS

### B. Experimental Setup

To analyze the scene classification performance on the mentioned data sets, we use VGG-Net to extract descriptive features from each image scene, which can provide many informative features. And then we select two layers as two different feature descriptors. After that, we use two techniques of fusion: 1) normal fusion and 2) the DCA fusion, to combine among the extracted features. For DCA fusion, the size of transformed features is determined by (16). In the classification

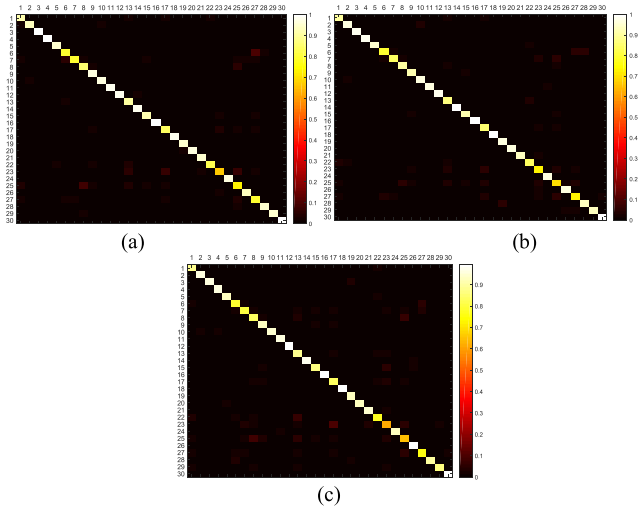

Fig. 7. Average confusion matrices on the AID (a) fusion by addition, (b) fusion by concatenation, (c) DCA with concatenation. The rows and columns of the matrix denote the actual and predicted classes, respectively. The class labels range in 1:30 as shown in Fig. 4. The vertical color bar indicates the proportion of samples over the actual total class samples.

TABLE I
OVERALL ACCURACY ON AID

| Method | Accuracy |
|---|---|
| CaffeNet [43] | 89.53±0.31 % |
| VGG-VD-16 [43] | 89.64±0.36 % |
| GoogLeNet [43] | 86.39±0.55 % |
| (our results) | |
| Fusion by addition | **91.87±0.36%** |
| Fusion by concatenation | **91.86±0.28%** |
| DCA with concatenation | **89.71±0.33%** |

task, we use LIBSVM library [46], for the linear SVM. The regularization parameters are selected by fivefold cross validation with the arrangement of $[1^{-2}, 1^{-1} \ldots 1^{9}, 1^{10}]$.

### C. AID Data Set

The performances of VHR scene classification on the UC Merced and WHU-RS data sets are recently sutured and achieved the highest accuracy [21]. Therefore, Xia *et al.* [43] constructed a new challenging data set named AID as illustrated above, to advance the state-of-the-art methods in the context of VHR image scene understanding. To evaluate the performance of our method on the AID data set, we follow the same experimental setup used in [43], which select 50% of the labeled images from each category for training and the remaining (50%) for test. In terms of classification accuracies, the strategies of fusion by concatenation, fusion by addition, and DCA with concatenation are compared with methods used in [43], which extract high-level feature extractors based on CaffeNet, VGG-VD-16, and GoogLeNet.

The experimental results shown in Table I clearly demonstrate that our method produced the best performance based on fusion by addition strategy with features' size equal to 4096. Moreover, using feature fusion strategy based on DCA achieved an important accuracy with a very small size of feature descriptor, which equals 58. The comparison of confusion matrices for different fusion strategies is shown in Fig. 7.

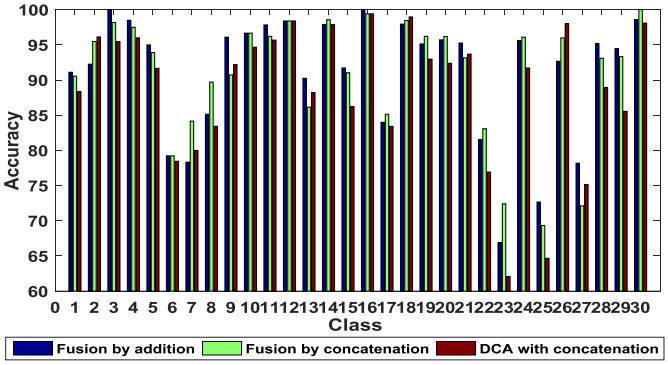


Fig. 8. Per-class classification performance of the proposed method on the AID data set using fusion by addition, concatenation, and DCA by concatenation as a feature descriptor.

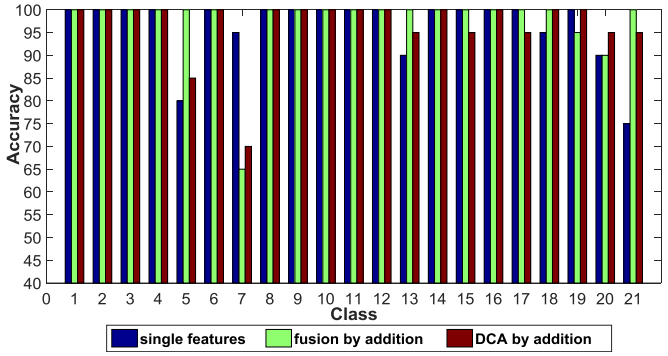


Fig. 9. Per-class classification performance of the proposed method on the UC Merced data set using single features, fusion by addition, and DCA by addition as a feature descriptor.

TABLE II
COMPARISON WITH DIFFERENT METHODS OF FUSION ON THE UC MERCED DATA SET

| Method | Features size | accuracy |
|---|---|---|
| Without fusion | 4096 | 95.99±0.21% |
| Fusion by concatenation | 8192 | 96.37±0.73% |
| Fusion by addition | 4096 | **97.42±1.79%** |
| DCA by concatenation | 40 | 96.90±0.56% |
| DCA by addition | 20 | **96.90±0.09**% |

The comparisons of per-class classification accuracies on AID fusion by addition, concatenation, and DCA by concatenation are shown in Fig. 8.

### D. UC Merced Data Set

To analyze the scene classification performance on the UC Merced data set, we follow the experiment setup used in [8], which selects 80 images per category for training and the remaining for test. In terms of classification accuracy, the strategy without fusion is compared with different ways of fusion (by concatenation, by addition, DCA by concatenation, and DCA by addition). The comparisons of per-class performance with different fusion strategies are shown in Fig. 9.

In order to study the sensitivity of the feature size, we compare the different fusion strategies and the way without feature fusion as shown in Table II. The comparisons of confusion

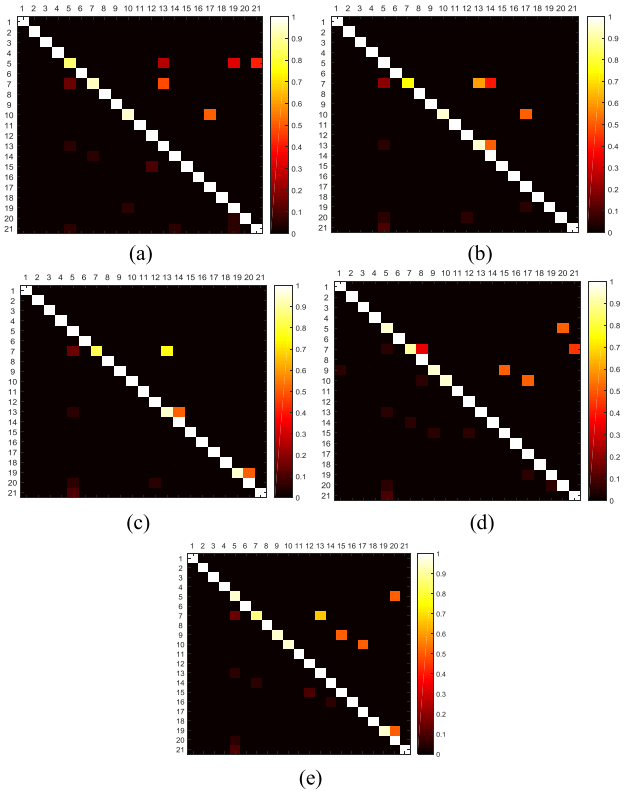


Fig. 10. Average confusion matrices on the UC Merced data set (a) without fusion, (b) fusion by concatenation, (c) fusion by addition, (d) DCA by concatenation, and (e) DCA by addition. The rows and columns of the matrix denote the actual and predicted classes, respectively. The class labels range in 1:21 as shown in Fig. 5. The vertical color bar indicates the proportion of samples over the actually total class samples.

matrices on the UC Merced data set for different fusion strategies are shown in Fig. 10: without fusion [Fig. 10(a)], fusion by concatenation [Fig. 10(b)], fusion by addition [Fig. 10(c)], DCA by concatenation [Fig. 10(d)], and DCA by addition [Fig. 10(e)].

We then compare the scene classification performance of the proposed method with 1) different approaches based on handcrafted features including SIFT and SC approach cited in [6], saliency guided sparse autoencoder noted in [8], recent work based on part detection [9], and land-use classification with compressive sensing multifeature fusion reported in [45] and 2) recent work based on deep features including the proposed approach developed in [17], which combines different deep neural networks via gradient boosting random convolutional network (GBRCN) for VHR scene classification, the fine-tuned GoogLe-Net followed by linear SVM introduced in [18], the pretrained Overfeat model followed by CNN classifier proposed in [24], convolutional features and sparse autoencoder for VHR scene classification cited in [19], land-use classification in remote sensing images by CNN referenced in [23], the approach described in [20], and the two scenarios developed in [21]. In [21], in the first scenario, the second fully connected layer is considered as a final feature descriptor of the images scene, and in the second scenario, they extract dense features from the last convolutional layer at multiscales, and then dense features are encoded into global representation by feature coding approaches such as BoVW. The comparison of different methods is shown in Table IV.

TABLE III

AVERAGE TRAINING AND TESTING TIME (SECONDS) ON THE UC MERCED DATA SET

| Feature size | Training | Testing |
|---|---|---|
| 20 (DCA by addition) | 0.084122 | 0.001502 |
| 40 (DCA by concatenation) | 0.270912 | 0.001811 |
| 4096 (first and second scenario[21] ) | 16.115644 | 0.110222 |

TABLE IV

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE UC MERCED DATA SET

| Method | Accuracy | Method | Accuracy |
|---|---|---|---|
| SIFT+SC [6] | 81.67 ±1.23% | scenarios (II) [21] | 96.90 ± 0.77 % |
| SSEA [8] | 82.72±1.18% | scenarios (I)+ (II) [21] | 98.49% |
| With CODP [9] | 91.33±1.11% | CaffeNet [20] | 93.42± 1.00 % |
| CS fusion [45] | 94.33 % | OverFeats [20] | 90.91± 1.91 % |
| GBRCN [17] | 94.53% | CaffeNet +fine-tuning [23] | 95.48 % |
| CNN-NN [19] | 97.19% | GoogLeNet +fine-tuning [23] | 97.10 % |
| CNN-AE [19] | 95.05% | CNN with OverFeat [24] | 92.4 % |
| scenarios (I) [21] | 96.88 ± 0.72 % | Fusion by addition (ours) | **97.42±1.79 %** |
| fine-tuned Goo-gLeNet [18] | 97.78±0.97% | DCA fusion(ours) | **96.90±0.09** % |

TABLE V

SCENE CLASSIFICATION PERFORMANCE UNDER DIFFERENT TRAINING AND TEST SIZES ON THE UC MERCED DATA SET

| Train data | Test data | Fusion by addition | DCA fusion |
|---|---|---|---|
| 20% | 80% | 92.96±0.58 % | 92.32±1.02 % |
| 40% | 60% | 96.01±0.51% | 94.06±0.59% |
| 60% | 40% | 96.82±0.59% | 95.89±0.45% |
| 80% | 20% | 97.42±1.79 % | 96.90±0.09 % |

We compare the training and the testing time of different feature sizes (20 represents the output feature size of our proposed framework and 4096 represents the output feature size of concatenation between the two scenarios introduced in [21]) on an Intel Core i3-4130 (3.4 GHz) personal computer. As shown in Tables III and IV, the proposed method achieves an important accuracy and consumes a shorter learning time than the fusion between scenario I and scenario II introduced in [21].

We remark that fusion by addition based on DCA did not outperform single features and fusion by addition method for some classes; it is mainly because the UC Merced data set contains 21 classes and among of them have highly overlapping classes, where a large size of features is needed to describe the spatial and structural image scenes. Table V shows the recognition accuracies of the proposed method with different rates of the training and test data. It can be noted that as the size of training samples increases, the overall accuracy of the proposed method increases, too.

TABLE VI

COMPARISON WITH DIFFERENT METHODS OF FUSION ON THE WHU-RS DATA SET

| Method | Features size | Accuracy |
|---|---|---|
| Without fusion | 4096 | 95.99±0.20% |
| Fusion by concatenation | 8192 | 97.46±0.59% |
| Fusion by addition | 4096 | **98.65±0.43%** |
| DCA by concatenation | 36 | 98.70±0.23% |
| DCA by addition | 18 | **98.70±0.22**% |

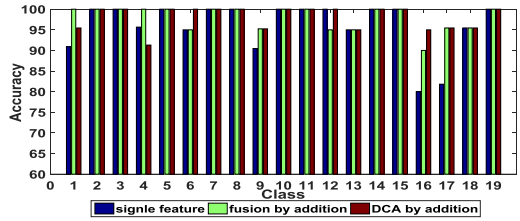


Fig. 11. Per-class classification performance of the proposed method on WHU-RS data set using single features, fusion by addition, and DCA by addition as a feature descriptor.

TABLE VII

OVERALL ACCURACY ON WHU-RS DATA SET

| Method | L1R-LR[27] | AlexNet [21] | CaffeNet [21] |
|---|---|---|---|
| Accuracy | 94.53±1.01 % | 93.81 % | 94.54 % |
| Method | VGG-VD16 [21] | BoVW-VGG-VD16 [21] | IFK-VGG-VD16 [21] |
| Accuracy | 94.35 % | 98.10 % | 97.79 % |
| Method | VLAD-VGG-VD16 [21] | Fusion by addition | DCA by addition |
| Accuracy | 98.64 % | **98.65±0.43%** | **98.70±0.22**% |

### *E. WHU-Data Set*

The same as the experiments discussed above for UC Merced data set, to evaluate the classification accuracy on the WHU-RS data set, we follow the experimental setup used in [27]. It selects 60% of the labeled images for each class as a training set for each class and the remaining images for test. To study the sensitivity of the feature size, we compare the strategy without feature fusion and different ways of fusion strategies by concatenation, by addition, and DCA by concatenation and by addition. The results are shown in Table VI.

The comparisons of classification accuracies on WHU-RS data set without fusion (i.e., single features), with fusion by addition, and DCA by addition are shown in Fig. 11. We note that in terms of the accuracy, the way of DCA by addition is superior to the ways of the single feature and fusion by addition in most classes, except airport and bridge classes. It may be because these two classes are more susceptible to the changes of resolution and orientation in images scene, so they need a large size of features. The comparisons of confusion matrices for different fusion strategies are shown in Fig. 12.

Table VII gives the comparison of the classification accuracies between our proposed method and the other two comparison methods [21], [27]. In our proposed method, it gives the final classification accuracies with two different fusion methods based on the addition. One comparison method is

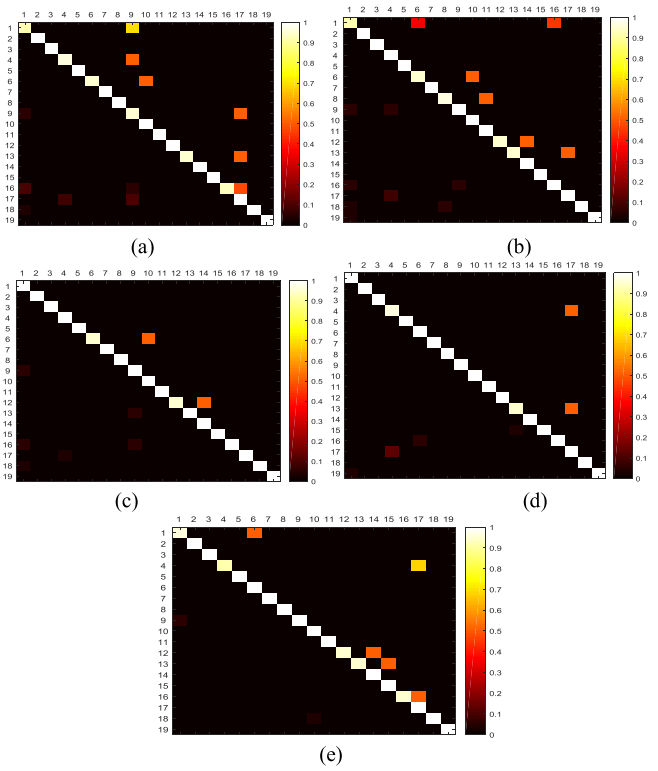


Fig. 12. Average confusion matrices on WHU-RS data set (a) without fusion, (b) fusion by concatenation, (c) fusion by addition, (d) DCA by concatenation, and (e) DCA by addition. The rows and columns of the matrix denote the actual and predicted classes, respectively. The class labels range in 1:19 as shown in Fig. 6. The vertical color bar indicates the proportion of samples over the actually total class samples.

TABLE VIII

AVERAGE TRAINING AND TESTING TIME (SECONDS) ON THE WHU-RS DATA SET

| Feature size | Training | Testing |
|---|---|---|
| 18 (DCA by addition) | 0.036715 | 0.001523 |
| 36 (DCA by concatenation) | 0.062131 | 0.001674 |
| 4096 (first and second scenario[21] ) | 3.931285 | 0.107598 |

a hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization cited in [27]. The other is the approach described in [21], which introduced the performance of different pretrained models on the WHU-RS data set and the different convolutional feature coding methods such as BoVW, improved Fisher kernel, and vector of locally aggregated descriptors (VLAD). The experimental results clearly demonstrate that our method produced the best performance based on DCA fusion for addition strategy, where the size of features is equal to 18. Moreover, the proposed method is based on DCA as feature fusion method. It achieves an important accuracy with a very small size of feature descriptor. Table VIII gives the average training and testing time (seconds) on the WHU-RS data set.

## IV. CONCLUSION

In this paper, a new approach for VHR images scene classification is presented. The proposed method is based on VGG-Net model as a feature extractor. Feature fusion strategies are introduced to combine among features extracted from

the images using VGG-Net. The traditional fusion methods based on addition and concatenation are used. And an efficient method developed for features fusion called DCA is introduced. We tested by addition and concatenation between the new transformation of the extracted features. The experimental results on three different data sets show that the feature fusion technique is efficient to represent the VHR images scene. Moreover, DCA fusion method gives a good representation of images scene with low dimension. The proposed method still achieves better performances than the current state-of-the-art methods.

## REFERENCES


[1] 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.

[2] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, 2003, pp. 1470–1478.

[3] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2006, pp. 2169–2178.

[4] I. T. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer, 2002.

[5] J. Xia, J. Chanussot, P. Du, and X. He, "(Semi-) supervised probabilistic principal component analysis for hyperspectral remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 7, no. 6, pp. 2224–2236, Jun. 2014.

[6] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.

[7] X. Lu, X. Li, and L. Mou, "Semi-supervised multitask learning for scene recognition," *IEEE Trans. on*, vol. 45, no. 9, pp. 1967–1976, Sep. 2015.

[8] 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.

[9] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.

[10] 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.

[11] C. Vaduva, I. Gavat, and M. Datcu, "Deep learning in very high resolution remote sensing image information mining communication concept," in *Proc. Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2012, pp. 2506–2510.

[12] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.

[13] C. Tao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2438–2442, Dec. 2015.

[14] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[15] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[16] A. Romero, P. Radeva, and C. Gatta, "Meta-parameter free unsupervised sparse feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1716–1722, Aug. 2015.

[17] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[18] K. Nogueira, O. A. B. 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. [Online]. Available: http://www.sciencedirect.com/science/article/pii/-S0031320316301509

[19] E. Othman, Y. Bazi, N. Alajlan, H. Alhichri, and F. Melgani, "Using convolutional features and a sparse autoencoder for land-use scene classification," *Int. J. Remote Sens.*, vol. 37, no. 10, pp. 2149–2167, 2016.

[20] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 44–51.

[21] 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. 14680–14707, 2015. [Online]. Available: http://-www.mdpi.com/2072-4292/7/11/14680

[22] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[23] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva. "Land use classification in remote sensing images by convolutional neural networks." Unpublished paper, 2015. [Online]. Available: https://arxiv.org/abs/1508.00092

[24] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using imagenet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[25] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

[26] X. Zheng, X. Sun, K. Fu, and H. Wang, "Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 652–656, Jul. 2013.

[27] W. Shao, W. Yang, G. S. Xia, and G. Liu, "A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization," in *Proc. 9th Int. Conf. Comput. Vis. Syst.-(ICVS)*, St. Petersburg, Russia, Jul. 2013, pp. 324–333.

[28] 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. Photogram. Remote Sens.*, vol. 116, pp. 73–85, Jun. 2016.

[29] 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.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran & Assoc. Inc, 2012, pp. 1097–1105.

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

[32] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. CoRR*, 2014, pp. 675–678.

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

[34] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. British Mach. Vis. Conf. (BMVC)*, 2014.

[35] C. Liu and H. Wechsler, "A shape- and texture-based enhanced Fisher classifier for face recognition," *IEEE Trans. Image Process.*, vol. 10, no. 4, pp. 598–608, Apr. 2001.

[36] J. Yang and J. Y. Yang, "Generalized K–L transform based combined feature extraction," *Pattern Recognit.*, vol. 35, no. 1, pp. 295–297, 2002.

[37] J. Yang, J.-Y. Yang, D. Zhang, and J.-F. Lu, "Feature fusion: Parallel strategy vs. serial strategy," *Pattern Recognit.*, vol. 36, no. 6, pp. 1369–1381, Jun. 2003.

[38] Q.-S. Sun, S.-G. Zeng, Y. Liu, P.-A. Heng, and D.-S. Xia, "A new method of feature fusion and its application in image recognition," *Pattern Recognit.*, vol. 38, no. 12, pp. 2437–2448, Dec. 2005.

[39] J. Yang and X. Zhang, "Feature-level fusion of fingerprint and finger-vein for personal identification," *Pattern Recognit. Lett.*, vol. 33, no. 5, pp. 623–628, Apr. 2012.

[40] Y. Bi, M. Lv, Y. Wei, N. Guan, and W. Yi, "Multi-feature fusion for thermal face recognition," *Infr. Phys. Technol.*, vol. 77, pp. 366–374, Jul. 2016.

[41] J. R. Schott, "Principles of multivariate analysis: A user's perspective," *J. Amer. Statist. Assoc.*, vol. 97, no. 458, 2002.

[42] M. Haghighat, M. Abdel-Mottaleb, and W. Alhalabi, "Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 9, pp. 1984–1996, Sep. 2016.
[43] G. Xia *et al.* "AID: A benchmark dataset for performance evaluation of aerial scene classification." Unpublished paper, 2016. [Online]. Available: https://arxiv.org/abs/1608.05167
[44] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, New York, NY, USA, 2010, pp. 270–279.
[45] 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.
[46] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, Apr. 2011, Art. no. 27.

**Souleyman Chaib** (S'06) was born in Mostaganem, Algeria, in 1988. He received the B.S. and M.S. degrees in computer science from the University of Science and Technology of Oran—Mohamed Boudiaf, Bir El Djir, Algeria, in 2009 and 2011, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.

His research interests include very high resolution image classification and scene classification.

**Huan Liu** (S'15) was born in Chaoyang, China, in 1992. She received the B.E. degree in electronics and information engineering from the Harbin Institute of Technology, Harbin, China, in 2015, where she is currently pursuing the Ph.D. degree in information and communication engineering.

Her research interests include deep learning, kernel methods, classification of hyperspectral image, and scene classification.

**Yanfeng Gu** (M'06–SM'16) received the Ph.D. degree in information and communication engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2005.

He joined as a Lecturer with the School of Electronics and Information Engineering, HIT, where he was an Associate Professor in 2006; meanwhile, he was enrolled in first Outstanding Young Teacher Training Program of HIT. From 2011 to 2012, he was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA, USA. He is currently a Professor with the Department of Information Engineering, HIT. He has authored more than 90 peer-reviewed papers, 4 book chapters, and is the inventor or co-inventor of 7 patents. His research interests include image processing in remote sensing, machine learning and pattern analysis, and multiscale geometric analysis.

Dr. Gu is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the *Neurocomputing*. He is also a Peer Reviewer for several international journals, such as the IEEE TRANSACTION ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTION ON INSTRUMENTATION AND MEASUREMENT, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and *IET Electronics Letters*.

**Hongxun Yao** (M'11) received the B.S. and M.S. degrees in computer science from the Harbin Shipbuilding Engineering Institute, Harbin, China, in 1987 and 1990, respectively, and the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, in 2003.

She is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. Her research interests include computer vision, pattern recognition, multimedia computing, and human computer interaction technology. She has authored six books and over 200 scientific papers.

Dr. Yao was a recipient of the honorary titles of The New Century Excellent Talent and Enjoy Special Government Allowances Expert in China.