IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 18, 2017, accepted January 16, 2018, date of publication January 26, 2018, date of current version March 16, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2798799

Exploiting Convolutional Neural Networks With Deeply Local Description for Remote Sensing Image Classification

NA LIU[1], LIHONG WAN[1], YU ZHANG[2], TAO ZHOU[1], HONG HUO[1], AND TAO FANG[1]

[1]Department of Automation, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Department of Psychiatry and Behavior Sciences, Stanford University, Stanford, CA 94305 USA

Corresponding author: Tao Fang (tfang@sjtu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 41571402, in part by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China under Grant 61221003, and in part by the Shanghai Jiao Tong University Agri-X Fund under Grant Agri-X2015004.

ABSTRACT The extraction of features from the fully connected layer of a convolutional neural network (CNN) model is widely used for image representation. However, the features obtained by the convolutional layers are seldom investigated due to their high dimensionality and lack of global representation. In this study, we explore the uses of local description and feature encoding for deeply convolutional features. Given an input image, the image pyramid is constructed, and different pretrained CNNs are applied to each image scale to extract convolutional features. Deeply local descriptors can be obtained by concatenating the convolutional features in each spatial position. Hellinger kernel and principal component analysis (PCA) are introduced to improve the distinguishable capabilities of the deeply local descriptors. The Hellinger kernel causes the distance measure to be sensitive to small feature values, and the PCA helps reduce feature redundancy. In addition, two aggregate strategies are proposed to form global image representations from the deeply local descriptors. The first strategy aggregates the descriptors of different CNNs by Fisher encoding, and the second strategy concatenates the Fisher vectors of different CNNs. Experiments on two remote sensing image datasets illustrate that the Hellinger kernel, PCA, and two aggregate strategies improve classification performance. Moreover, the deeply local descriptors outperform the features extracted from fully connected layers.

INDEX TERMS Convolutional neural networks (CNN), image classification, local description, remote sensing.

I. INTRODUCTION

As a fundamental task, remote sensing image (RSI) classification plays an important role in many remote sensing applications [1], [2], such as hazard detection, determination of land use and land cover, geospatial object detection, geographic image retrieval, environment monitoring, urban planning [3], spatial-temporal data analysis [4], [5], and smart city [6]. In this study, we focus on the stage of feature representation in RSI classification, considering robust feature extraction is a critical step for obtaining a high classification performance.

Numerous methods for RSI classification have been developed during the past years. Existing methods can be divided into three main categories according to feature type. They are low-level, mid-level, and deep features. Early works on RSI classification were mainly based on low-level features, which aim to construct various handcrafted features, such as color features, shape features, texture features, spectral information, or a combination of multiple feature cues. Typical handcrafted features are color histograms [7], scale-invariant feature transform (SIFT) [7], Gabor texture [8], GIST [8], local binary patterns [9], and histograms of oriented gradients (HOG) [10]. However, handcrafted features should be redesigned for different data types, which causes feature extraction to heavily depend on the experiences of researchers.

For mid-level features, bag-of-visual-words (BOW) [7], [11] is a popular topic in image representation. Early works on BOW were applied to text analysis [12]. Since then, BOW has been widely used for image representation, such as

counting the frequency of visual words emerged in images. The visual words can be obtained by low-level features (e.g., dense SIFT [13]). The quantization of the low-level features is implemented by k-means or Gaussian mixture model (GMM) clustering to divide a set of descriptors into clusters, thereby causing within-cluster samples to be similar and between-cluster samples to be dissimilar. The classification pipeline of mid-level-based image representation generally consists of three stages [14]: (1) extracting low-level features, such as dense SIFT, HOG, or other types of local descriptors; (2) aggregating these low-level features into a global image representation by feature encoding; and (3) classifying the global image features using a classifier, such as support vector machine (SVM). In the past years, researchers have been dedicated in developing the second stage, and numerous feature encoding methods, such as locality-constrained linear encoding [15], improved Fisher encoding [16], super vector encoding [17], and kernel codebook encoding [18], have been proposed. The abovementioned feature encoding methods can be divided into two types. First, local descriptors are represented by combinations of visual words (e.g., kernel codebook encoding, local linear encoding). Second, the differences between the local descriptors and visual words are computed (e.g., improved Fisher encoding, super vector encoding). In the field of RSI classification, Sheng *et al.* [19] proposed a multiple feature combination method using sparse coding framework. Zheng *et al.* [20] introduced multi-feature joint sparse coding with spatial relation constraint. Cheriyadat [21] explored unsupervised feature learning by extracting dense low-level feature descriptors, followed by sparse encoding with learned basis functions. Kobayashi [22] proposed Dirichlet Fisher kernel to transform histogram-based features (e.g., dense SIFT) for improving the distinguishability of feature representation without increasing dimensionality. Chen and Tian [23] proposed pyramid-of-spatial-relation to investigate the absolute and relative spatial relationships of dense SIFT. Wan *et al.* [24] proposed a combination of multiple local descriptors by improved Fisher encoding. Overall, SIFT, HOG, and other local descriptors are the basis of existing mid-level methods, and BOW representations heavily depend on the extraction of low-level features.

Low-level or mid-level features are shallow representations. By contrast, deep-learning-based methods [25] can obtain different levels of data abstractions, which can significantly improve the description capability of image representation. In 2012, a breakthrough for image classification using deep convolutional neural networks (CNNs) was made by Krizhevsky *et al.* [26]. Since then, CNNs have gained great success for a wide range of image recognition applications. Numerous CNN architectures, such as AlexNet [26], VGG-VD [27], and GoogLeNet [28], have been proposed for ImageNet [29] classification or other visual recognition applications. Several interesting works also exist in the field of RSI classification. For example, Makantasis *et al.* [30] proposed to use CNN for encoding spectral and spatial information of hyperspectral images. Zhang *et al.* [31] proposed a gradient-boosting random convolutional network that can be used to combine multiple CNNs. In most cases, using existing pretrained CNN models (trained on ImageNet) to extract feature representations for RSIs is a more suitable choice than designing and training a new CNN architecture because pretrained CNNs have good generality and a large-scale RSI dataset is unusual [32]. Penatti *et al.* [33] investigated the generalization power of pretrained CaffeNet [34] and OverFeat by using activation vectors extracted from fully connected layers. Nogueira *et al.* [32] conducted an extensive comparison analysis of three possible CNN strategies to explore the description capabilities of existing CNN architectures. Wan *et al.* [35] proposed selective CNNs and cascade classifiers to combine multiple pretrained CNNs.

In addition to the use of activation vectors from the fully connected layer of a CNN, the features in the convolutional layer also contain abundant information, such as local description. However, the features extracted from the convolutional layer are seldom investigated in RSI classification. To the best of our knowledge, Hu *et al.* [36] investigated dense descriptors from the convolutional layer of a single CNN, and the dense descriptors were directly aggregated into a global representation by feature coding. In this study, we explore local description and feature encoding for the features of the convolutional layer in three aspects: (1) How to extract deeply local descriptors? (2) How to make the deeply local descriptors distinguishable? and (3) How to aggregate these deeply local descriptors?

The main contributions of this study are as follows: (1) With the use of image pyramid, two types of CNN, CaffeNet and VGG-VD16, are introduced to extract deeply local descriptors and complement each other. The combination of CaffeNet and VGG-VD16 indicates significant performance advantages over a single CNN. (2) Principal component analysis (PCA) [37] and Hellinger kernel [38] are introduced to the linear and nonlinear transformations of deeply local descriptors. The distinguishability of deeply local descriptors can be significantly improved by PCA and Hellinger kernel, thereby improving RSI classification performances. and (3) Two aggregate strategies are proposed to form a global image representation from the deeply local descriptors. The first strategy aggregates the local descriptors of different CNNs by Fisher encoding [16], [40], and the second strategy concatenates the Fisher vectors (FV) of different CNNs. Both aggregate strategies indicate significant performance improvements.

The remainder of this paper is organized as follows. Section II elucidates the proposed method. Section III presents the experiments on two RSI datasets. Section IV concludes this paper.

II. PROPOSED METHOD

The proposed method consists of three parts, as shown in Fig. 1. The first part extracts deeply local descriptors from convolutional layers, given the input image and

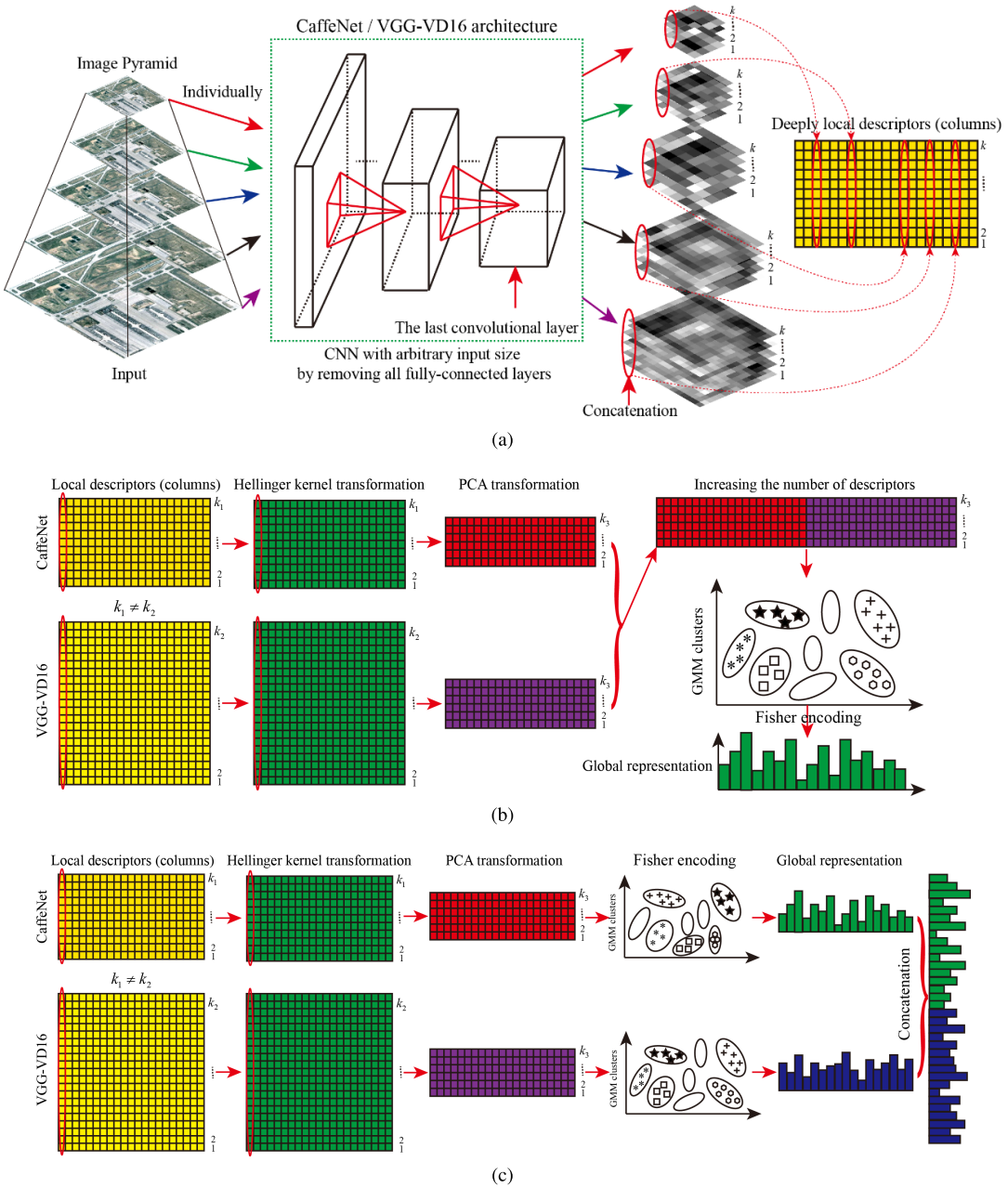


FIGURE 1. **Framework of the proposed method. (a) Given an image from the image pyramid, numerous local descriptors are extracted from the last convolutional layer using a CNN model. (b) Aggregate strategy 1: combination of local descriptors from different CNNs, followed by feature encoding. (c) Aggregate strategy 2: concatenation of feature encoding representations from different CNNs.**

CNN model. Two types of pretrained CNN, namely, CaffeNet [34] and VGG-VD16 [27], are used to extract deep convolutional features and complement each other. The second and third parts are the proposed aggregate strategies for the deeply local descriptors.

In the first part (Fig. 1 (a)), an image pyramid is constructed, and each image in the pyramid is individually used as the input of a CNN model to extract the convolutional features from the last convolutional layer. A local descriptor can be formed by concatenating the convolutional features that are at the same spatial position from all feature maps. Numerous local descriptors can be obtained for the image pyramid (multiple scales and positions).

In the second part (Fig. 1 (b)), the local descriptors are processed by two steps: (1) Hellinger kernel, which helps improve distinguishability, is applied to the nonlinear transformation of these deeply local descriptors. and (2) Considering that the dimensions of the deeply local descriptors extracted from different CNNs are different, PCA is introduced to reduce the dimensions of different descriptors to a

fixed length, thereby allowing the local descriptors extracted from different CNNs to be aggregated. Feature encoding, such as Fisher encoding with GMM clustering, is selected to aggregate the deeply local descriptors into a global representation, followed by linear SVM classification [41].

In the third part (Fig. 1(c)), the deeply local descriptors obtained by the first part are also processed by Hellinger kernel and PCA, similar to the second part. Then, the two types of local descriptor (CaffeNet-based and VGG-VD16-based) are individually aggregated by Fisher encoding with GMM clustering to form the global representations. Finally, the two types of global representations are concatenated to enhance the mid-level representation, to which linear SVM classification is performed.

A. DEEPLY LOCAL DESCRIPTION WITH CONVOLUTIONAL LAYER

Existing CNN architectures (e.g., AlexNet [26], CaffeNet [34], VGG-VD [27], GoogLeNet [28]) trained on ImageNet can obtain different classification performances on an ImageNet testing set. For example, VGG-VD and GoogLeNet can obtain significant accuracy advantages over AlexNet or CaffeNet. However, when transferring these pretrained CNNs (trained on ImageNet) to other types of datasets, such as RSIs, the abovementioned conclusion is unnecessarily reasonable. The experimental results from Xia *et al.* [2] indicated that pretrained CaffeNet shows a better performance than that of GoogLeNet and performs similar to VGG-VD in RSI classification. The performance (well or poor) of a pretrained CNN on the original training and testing dataset is generally independent of its performance on another dataset. Therefore, both pretrained CaffeNet and VGG-VD are used to extract deep convolutional features for enhancing the description capability of the deeply local descriptors.

1) CaffeNet

As a reference model in Caffe open source framework, CaffeNet [34] is nearly a replication of AlexNet [26]. Unlike AlexNet, CaffeNet has no data argumentation in the training stage, and the order of normalization and pooling is exchanged. CaffeNet contains five convolutional layers and three fully connected layers. The input size of CaffeNet is 227×227 pixels with three channels (red–green–blue). Each convolutional layer includes a linear convolution with one or more nonlinear operations (e.g., local response normalization, rectified linear units, and max pooling). The first convolutional layer includes 96 filters (the size of each filter is $11 \times 11 \times 3$). The second convolutional layer includes 256 filters (the size of each filter is $5 \times 5 \times 48$). The third convolutional layer includes 384 filters (the size of each filter is $3 \times 3 \times 256$). The fourth convolutional layer includes 384 filters (the size of each filter is $3 \times 3 \times 192$). The fifth convolutional layer includes 256 filters (the size of each filter is $3 \times 3 \times 192$). The sixth to eighth fully connected layers contain 4096, 4096, and 1000 neurons, respectively.

2) VGG-VD16

VGG-VD [27] won the localization and classification tasks in ILSVRC 2014. As one of the best performing CNNs in VGG-VD architectures, VGG-VD16 contains 13 convolutional layers, 5 pooling layers, and 3 fully connected layers. The input size of VGG-VD16 is 224×224 pixels with three channels (red–green–blue). The filter size of all convolutional layers is selected to be 3×3 pixels uniformly, and the stride of convolutional operation is set to 1 pixel. After the convolutional operation, the size of the feature maps can be preserved by spatial padding. Several convolutional layers are followed by a max pooling layer. The size of a max pooling region is 2×2 pixels, and the stride is set to 2. The last three fully connected layers contain 4096, 4096, and 1000 neurons.

3) EXTRACTING DEEPLY LOCAL DESCRIPTORS

A pretrained CNN can be considered a feature extractor for image representation, including RSI, because the learned convolutional kernels in the CNN model have minimal data dependence. When we take a pretrained CNN model as a feature extractor, a popular feature extraction strategy is extracting an activation vector from the last fully connected layer (except for the classification layer) [32], [33]. Although the activation vector extracted from fully connected layers can capture the global structure of an input image, it remains sensitive to object changes, such as rotations and scales. Thus, we investigate the deep features extracted from the convolutional layer (rather than the fully connected layer) to form local descriptions. The local descriptions can then be aggregated through feature encoding to enhance the robustness of image representation with changes in rotations, scales, and other local variations.

Given a pretrained CNN, the size of the input image must be fixed because the weight connections between the last convolutional layer and the first fully connected layer are predefined. All of the fully connected layers are removed in this study to overcome the limitation of input image size, thereby enabling the pretrained CNN to accept an input image with arbitrary size. An image pyramid is also constructed by resizing the original input image to different sizes to obtain abundant and multi-scale local descriptions. Each scale in the image pyramid is individually used as the input of the pretrained CNN to extract the feature maps of the last convolutional layer. Noting that these feature maps are not processed by ReLU and max pooling because the two CNN operations can greatly reduce the distinguishability of deeply local descriptor. For each scale, a deeply local descriptor with L2 normalization can be formed by concatenating the convolutional features that are at the same spatial position from all feature maps, as shown in Fig. 1(a).

B. FEATURE TRANSFORMATION WITH HELLINGER KERNEL AND PCA

Two types of processing techniques, Hellinger kernel [38] and PCA transformation [37], are introduced to make

the deeply local descriptors distinguishable, as shown in Figs. 1(b) and 1(c). The Hellinger kernel is a popular technique for the nonlinear transformation of local descriptors, such as histogram-based SIFT. Although the deeply local descriptors extracted from the convolutional layer are not strictly histograms, this study illustrates that applying Hellinger kernel to these descriptors does indeed improve their distinguish capabilities and the subsequent classification performance.

The dimensional lengths of the local descriptors extracted from CaffeNet and VGG-VD16 are inconsistent. For example, the lengths of CaffeNet-based and VGG-VD16-based local descriptors are 256 and 512 dimensions, respectively. PCA can project different local descriptors into the same length, thereby causing the ''aggregate strategy 1'' (Fig. 1(b)) for local descriptors with different dimensions to be feasible.

1) HELLINGER KERNEL

Distance measure exists in GMM clustering and feature encoding for the deeply local descriptors. A suitable measure for the descriptors should be selected. Euclidean distance is a popular choice for the measure of different descriptors (e.g., $\mathbf{x}_i, \mathbf{x}_j$) with L2 normalization. The Euclidean distance can be defined as

$$dis(\mathbf{x}_i, \mathbf{x}_j)^2 = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2 = \|\mathbf{x}_i\|_2^2 + \left\|\mathbf{x}_j\right\|_2^2 - 2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j, \quad (1)$$

where $\|\mathbf{x}_i\|_2^2 = 1$, $\left\|\mathbf{x}_j\right\|_2^2 = 1$. Equation (1) can be further represented by

$$dis(\mathbf{x}_i, \mathbf{x}_j)^2 = 2 - 2E, \quad (2)$$

where $E = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$.

For the measure of histogram-based descriptors, using χ^2 or Hellinger can often perform better than using Euclidean distance. Given arbitrary two local descriptors, the Hellinger kernel [38] is defined as follows:

$$H(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{d'} \sqrt{\mathbf{x}_i^m \mathbf{x}_j^m}, \quad (3)$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two descriptors with L1 normalization; d' is the dimension length of the local descriptor; and m is the index of dimension. The two descriptors with L2 normalization can be measured by a Hellinger kernel through replacing the Euclidean kernel E with H. To achieve Hellinger measure, each descriptor can be processed by two steps: (1) L1 normalization of each descriptor, which is originally normalized with L2 normalization; and (2) square rooting of each element.

2) PCA

PCA is a traditional subspace learning technique for dimensionality reduction to alleviate curse-of-dimensionality [39], and it seeks an optimal linear projection on the basis of least mean square reconstruction errors. Given a d'-dimensional descriptor $\mathbf{x}$ and a learned PCA subspace $\mathbf{W}$ with $d' \times d$ elements (d is the number of principal components), the projection of $\mathbf{x}$ is achieved by

$$\mathbf{x}' = \mathbf{W}^{\mathrm{T}}\mathbf{x}, \quad (4)$$

where $\mathbf{x}'$ is a d-dimensional vector. A large number of local descriptors, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, are collected from training images to learn $\mathbf{W}$. The projection $\mathbf{W}$ containing d top eigenvectors (corresponding to d largest eigenvalues) can be obtained by computing the covariance matrix of the local descriptors and performing eigenvalue decomposition.

C. ENCODING DEEPLY LOCAL DESCRIPTORS

Given a set of deeply local descriptors (processed by Hellinger kernel and PCA) for an image, Fisher encoding [40] is used to encode these local descriptors into a global representation with the use of a probability density distribution, such as GMM.

1) CLUSTERING WITH GMM

A GMM $p(\mathbf{x}|\phi)$ is the probability density on $\mathbb{R}^d$, i.e.,

$$p(\mathbf{x}|\phi) = \sum_{k=1}^{K} p(\mathbf{x}|\mu_k, \Sigma_k)\omega_k, \quad (5)$$

where $p(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{x}-\mu_k)}$, $\phi = (\omega_1, \mu_1, \Sigma_1, \ldots, \omega_K, \mu_K, \Sigma_K)$ represents the parameters, $\omega_k \in \mathbb{R}_+$ is the prior probability values, $\mu_k \in \mathbb{R}^d$ is the mean vectors, and $\Sigma_k \in \mathbb{R}^{d\times d}$ is the positive definite covariance matrices for a Gaussian component. The covariance matrices are assumed to be diagonal, and the variance vector is denoted by σ^2. Given a set of local descriptors $\mathbf{x}_1$, $\mathbf{x}_2, \ldots, \mathbf{x}_N$ collected from training images, the GMM parameters can be learned by using expectation-maximization algorithm. GMM defines the soft data-to-cluster assignments $q_{k,i}(k = 1, 2, \cdots, K, i = 1, 2, \cdots, N)$ from descriptors to the Gaussian components,

$$q_{k,i} = \frac{p(\mathbf{x}_i|\mu_k, \Sigma_k)\omega_k}{\sum_{j=1}^{K} p(\mathbf{x}_i|\mu_j, \Sigma_j)\omega_j}. \quad (6)$$

2) FISHER ENCODING

Fisher encoding, which can be considered a soft visual vocabulary, measures the average first- and second-order differences between local descriptors and the clusters of a GMM. The encoding of an FV starts by learning a GMM model ϕ. Given a set of local descriptors $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ extracted from an input image, we let $q_{k,i}$ be the soft assignments of the i-th $(i = 1, 2, \cdots, n)$ descriptor to the k-th $(k = 1, 2, \cdots, K)$ Gaussian component. For each k, we define the vectors as follows:

$$\begin{cases} \mathbf{u}_k = \dfrac{1}{N\sqrt{\omega_k}} \displaystyle\sum_{i=1}^{N} q_{k,i}\left(\frac{\mathbf{x}_i - \mu_k}{\sigma}\right) \\ \mathbf{v}_k = \dfrac{1}{N\sqrt{2\omega_k}} \displaystyle\sum_{i=1}^{N} q_{k,i}\left(\frac{(\mathbf{x}_i - \mu_k)^2}{\sigma} - 1\right), \end{cases} \quad (7)$$

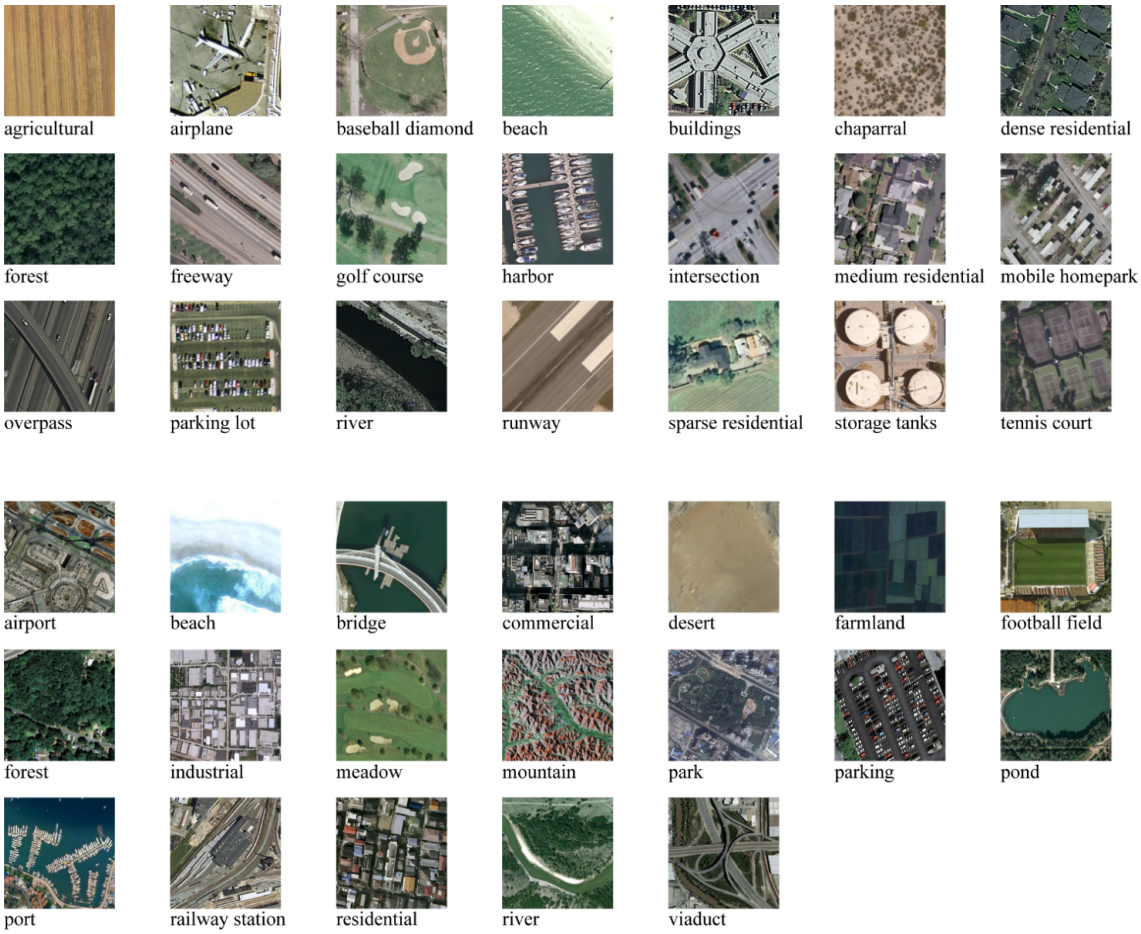


FIGURE 2. Datasets. Top: 21-class dataset. Bottom: 19-class dataset.

where the division between vectors is a term-by-term operation. The encoding result of these local descriptors is the concatenation of $\mathbf{u}_k$ and $\mathbf{v}_k$ for all Gaussian components, thereby resulting in a global vector with $2 \times d \times K$ dimensions.

III. EXPERIMENTS AND DISCUSSION

Experiments are conducted on two RSI datasets, as shown in Fig. 2. The first is a 21-class land use dataset (denoted by 21-class) [7] downloaded from the United States Geological Survey National Map. The second is a 19-class satellite scene dataset (denoted by 19-class) [19] collected by the Wuhan University from the Google Earth. The 21-class dataset is acquired from aerial orthoimagery with a pixel resolution of one foot and covers multiple regions of the United States. This dataset contains 2100 256 × 256-pixel images with red–green–blue channels (100 images per class) and includes different spatial structures, homogeneous texture and color, and some land cover and possibly object classes. The 19-class contains high-resolution satellite images up to half a meter. This dataset contains 950 600 × 600-pixel images with red–green–blue channels (50 images per class) and covers multiple regions around the world. Various objects with changes in scales, rotations, orientations, and illumination conditions exist in both datasets.

Each image in experiments is initially resized to 600×600 pixels, which indicates accuracy advantages than other sizes such as 500×500 or 400×400 pixels. With the use of a fixed subsampling ratio (e.g., 1.5), an image pyramid containing five scales (600×600 pixels, 400×400 pixels, 267×267 pixels, 178×178 pixels, and 119×119 pixels) is then constructed to obtain the desirable classification performances for both datasets. For the image pyramid, each image scale is individually used as the input of CaffeNet and VGG-VD16 to extract features from the last convolutional layer. Correspondingly, $37 \times 37 \times 256$, $24 \times 24 \times 256$, $16 \times 16 \times 256$, $10 \times 10 \times 256$, and $7 \times 7 \times 256$ features can be obtained by CaffeNet, and $38 \times 38 \times 512$, $25 \times 25 \times 512$, $17 \times 17 \times 512$, $12 \times 12 \times 512$, and $8 \times 8 \times 512$ features can be obtained by VGG-VD16, thereby resulting in a total of 2350 256-dimensional and 2566 512-dimensional descriptors for CaffeNet and VGG-VD16, respectively.

For both datasets, linear SVM is used for training and classification, and all results are repeated 10 times to report the average classification accuracy. In each round of testing, a certain percent (e.g., 10%, 50%, 80%) of images of each class are randomly selected to construct a subset for SVM training, and the remaining images are used for testing. One hundred thousand descriptors are randomly selected from

the training images for learning PCA subspace and GMM clusters. The amount of the selected descriptors is beneficial for obtaining desirable results and preserving computational cost.

In subsequent experiments, the effects of the two aggregate strategies, Hellinger kernel, numbers of PCA dimensions, and GMM clusters are analyzed by using 10% of the training images of each dataset; the confusion matrices for the two aggregate strategies are discussed by using 10% of the training images of each dataset; the performance comparisons of different methods are discussed with the use of different training ratios for both datasets.

A. ANALYSIS OF AGGREGATE STRATEGIES

Figure 3 shows the performance comparisons of three types of image representation. "Aggregate strategy 1" adopts both CaffeNet and VGG-VD16 to increase the total number of deeply local descriptors, followed by feature encoding; "aggregate strategy 2" encodes the CaffeNet-based and VGG-VD16-based local descriptors into two global representations through Fisher encoding, followed by feature concatenation; "single CNN-based description" encodes the CaffeNet-based or VGG-VD16-based local descriptors into a global representation. The number of GMM clusters is set to 64 for both datasets.

For "aggregate strategy 1," PCA transformation is necessary for the local descriptors extracted from CaffeNet (256 dimensions) and VGG-VD16 (512 dimensions) because the length of both descriptors is different. Thus, both AlexNet-based and VGG-VD16-based local descriptors are reduced to 255 dimensions to obtain the desirable classification performance. After both local descriptors are transformed to the same length, Fisher encoding can be applied to aggregate both local descriptors simultaneously. For "aggregate strategy 2," the lengths of CaffeNet-based and VGG-VD16-based local descriptors can be different because the encoding of both descriptors is independent. The CaffeNet-based and VGG-VD16-based local descriptors are therefore reduced to 255 and 511 dimensions, respectively.

The red and green bars shown in Fig. 3 represent single CaffeNet-based and single VGG-VD16-based local descriptions, respectively. The red bars shown in Fig. 3(a) or Fig. 3(b) are the same because the CaffeNet-based local descriptors are projected into 255 dimensions uniformly. The green bars shown in Fig. 3(a) or Fig. 3(b) present different PCA dimensions (255 dimensions in Fig. 3(a) and 511 dimensions in Fig. 3(b)).

The comparison of CaffeNet-based and VGG-VD16-based local descriptions indicates that VGG-VD16 indicates significant accuracy advantages over CaffeNet because the number of local descriptors extracted from VGG-VD16 is more than that from CaffeNet and VGG-VD16 is deeper than CaffeNet. The classification accuracy is improved for both datasets through the use of "aggregate strategy 1" to increase the total number of local descriptors compared with the use of a single CNN model. The concatenation of different Fisher encoding

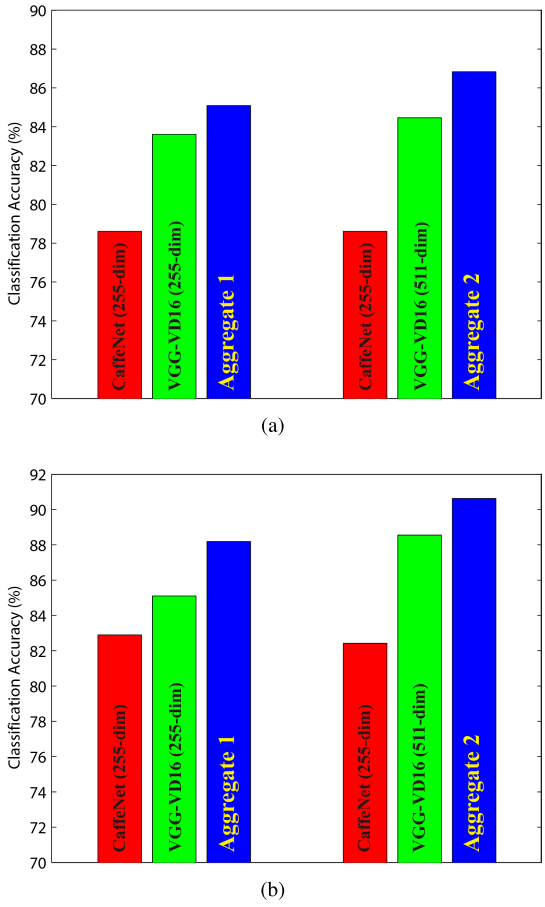


FIGURE 3. Comparisons of aggregate strategy 1, aggregate strategy 2, and single CNN-based description. (a) 21-class dataset. (b) 19-class dataset.

representations ("aggregate strategy 2") can also improve the classification accuracy compared with that of single CNN-based local description.

Unlike "aggregate strategy 1," "aggregate strategy 2" obtains a higher classification accuracy. The reason can be attributed to that "aggregate strategy 2" adopts 511-dimensional descriptors, whereas "aggregate strategy 1" selects 255-dimensional descriptors. A high-dimensional local descriptor helps preserve distinguishable image features.

B. ANALYSIS OF HELLINGER KERNEL

Figure 4 shows the effectiveness of Hellinger kernel on classification performance by using "aggregate strategy 1" and "aggregate strategy 2." The number of GMM clusters is set to 64 for both datasets. "Aggregate strategy 1" adopts 255-dimensional local descriptors for both CaffeNet and VGG-VD16, whereas "aggregate strategy 2" adopts 255-dimensional local descriptors for CaffeNet and 511-dimensional local descriptors for VGG-VD16.

The CNN-based local descriptor (without Hellinger kernel transformation) is dominated by its large feature values, which exert influences on the distance measure among

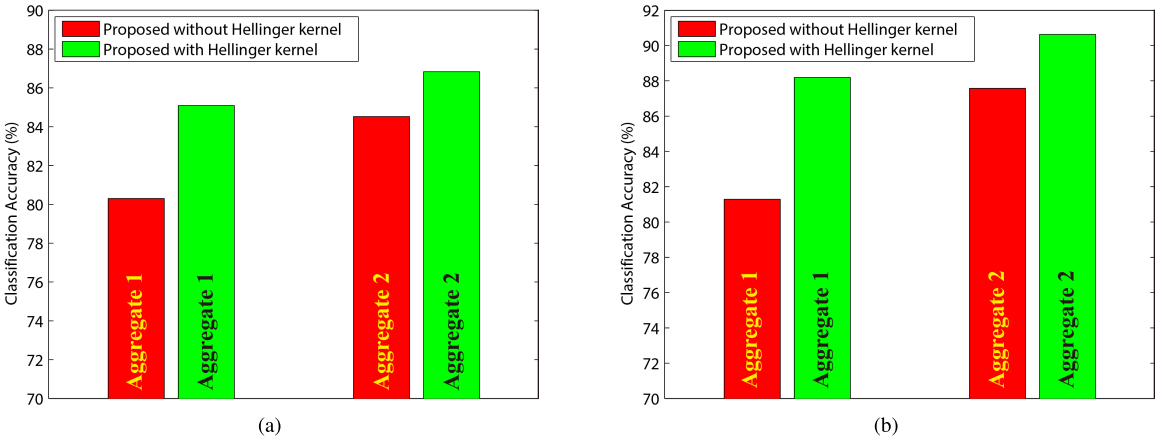


FIGURE 4. Effect of Hellinger kernel on "aggregate strategy 1" and "aggregate strategy 2." (a) 21-class dataset. (b) 19-class dataset.

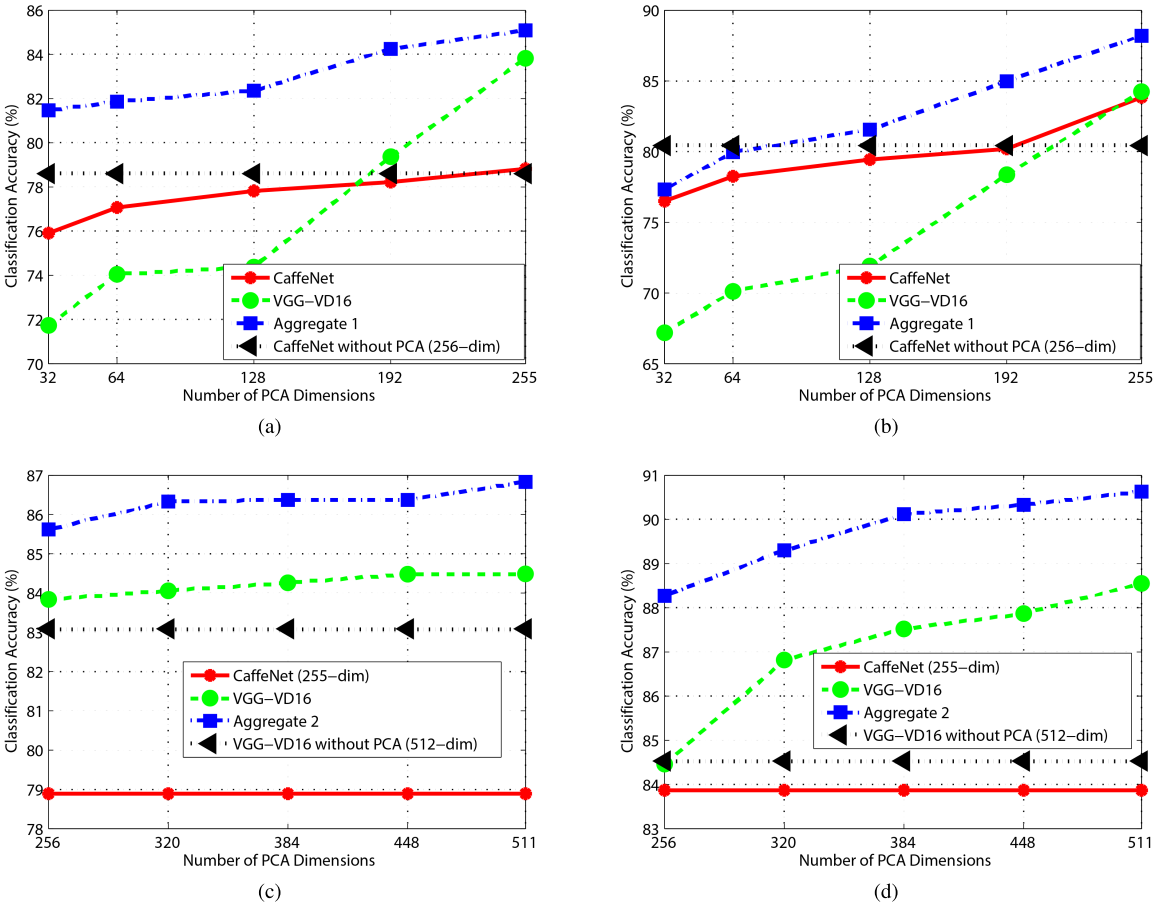


FIGURE 5. Effect of PCA transformation. (a) "aggregate strategy 1" on the 21-class dataset. (b) "aggregate strategy 1" on the 19-class dataset. (c) "aggregate strategy 2" on the 21-class dataset. (d) "aggregate strategy 2" on the 19-class dataset.

different local descriptors. The application of Hellinger kernel to each local descriptor is substantially a nonlinear transformation. Hellinger kernel can reduce the large feature values relative to small ones; thus, the distance measure is sensitive to the small feature values, thereby enhancing the distinguishability of local descriptors.

For "aggregate strategy 1," Hellinger kernel indicates approximately 5.5% and 6.5% accuracy improvements for the 21-class and 19-class datasets, respectively. For "aggregate strategy 2," Hellinger kernel indicates approximately 2% and 3% accuracy improvements for the 21-class and 19-class datasets, respectively. Both "aggregate strategy 1"

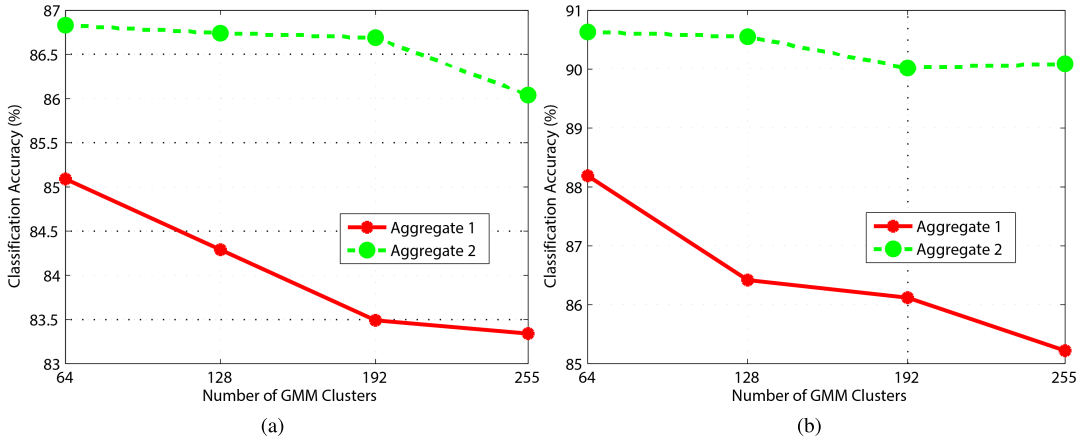


FIGURE 6. Effects of GMM clusters. (a) 21-class dataset. (b) 19-class dataset.

and "aggregate strategy 2" demonstrate that the description capabilities of the deeply local descriptors can be improved by Hellinger kernel.

C. ANALYSIS OF PCA

Figure 5 shows the performance comparisons of the proposed method with different numbers of PCA dimensions. The number of GMM clusters is set to 64 for both datasets.

Figures 5(a) and 5(b) show the performance of "aggregate strategy 1" with different PCA dimensions. Single CNN-based local description and feature encoding, such as "CaffeNet," "VGG-VD16," and "CaffeNet without PCA," are also given for comparison. With the increase in PCA dimensions, the classification accuracies for "CaffeNet," "VGG-VD16," and "aggregate strategy 1" increase. After PCA transformation, the use of 255-dimensional descriptors can obtain the best performances for all cases. The black lines represent the use of the original CaffeNet-based local descriptor (256 dimensions) for feature encoding. The comparison of "CaffeNet" and "CaffeNet without PCA" illustrates that retaining all principal components can obtain the best results.

Figures 5(c) and 5(d) demonstrate the performance of "aggregate strategy 2" with different PCA dimensions. In "aggregate strategy 2," CaffeNet-based local descriptors are transformed to a fixed length, whereas VGG-VD16-based local descriptors are transformed to different dimensions. Each CaffeNet-based local descriptor is transformed to 255 dimensions (the maximum length) by PCA, corresponding to the red straight lines shown in Figs. 5(c) and 5(d). With the increase in PCA dimensions, the VGG-VD16-based local descriptor combined with PCA transformation (green lines) indicates accuracy advantages over that without PCA (black lines), especially on the 19-class dataset. After PCA transformation, the use of 511 dimensions for the VGG-VD16-based local descriptor performs best. Correspondingly, "aggregate strategy 2" performs best when selecting 255-dimensional CaffeNet-based and 511-dimensional VGG-VD16-based local descriptors. Overall, the effect of PCA is accuracy improvement rather than dimensionality reduction.

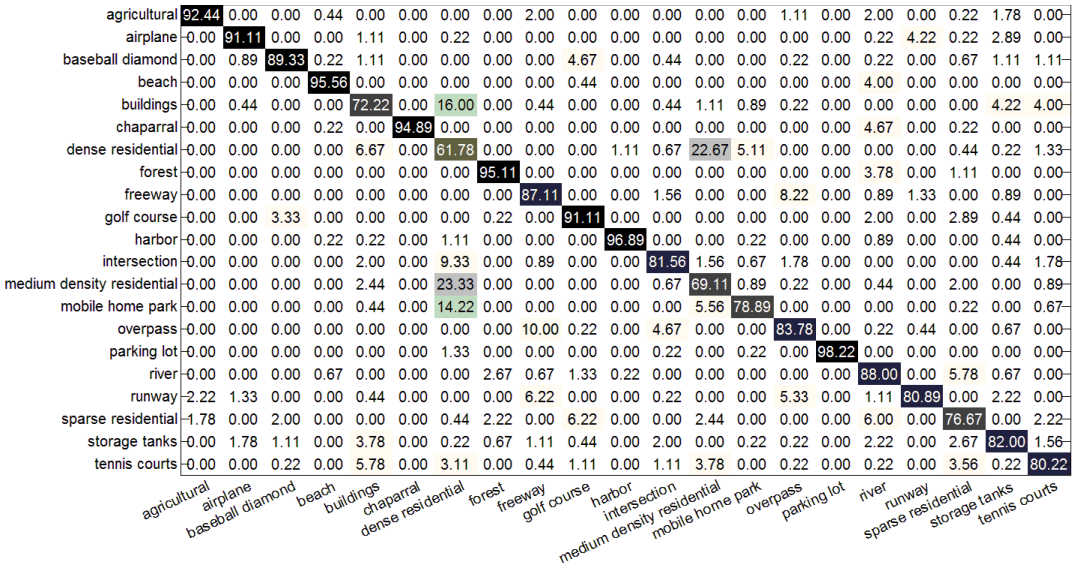
D. NUMBER OF GMM CLUSTERS

Figure 6 shows the effects of GMM clusters on the classification accuracies of both datasets. For "aggregate strategy 1," both CaffeNet-based and VGG-VD16-based local descriptors are reduced to 255 dimensions. For "aggregate strategy 2," the CaffeNet-based and VGG-VD16-based local descriptors are reduced to 255 and 511 dimensions, respectively. The reason for discussing GMM clustering in this section is that the number of GMM clusters decides the dimensionality of the final image representation. Given a GMM with 64 clusters, $2 \times 255 \times 64 = 32640$ and $2 \times (255 + 511) \times 64 = 98048$ dimensions can be obtained for "aggregate strategy 1" and "aggregate strategy 2," respectively. Therefore, although "aggregate strategy 2" performs better than "aggregate strategy 1," the dimensionality of the former is higher than that of the latter.

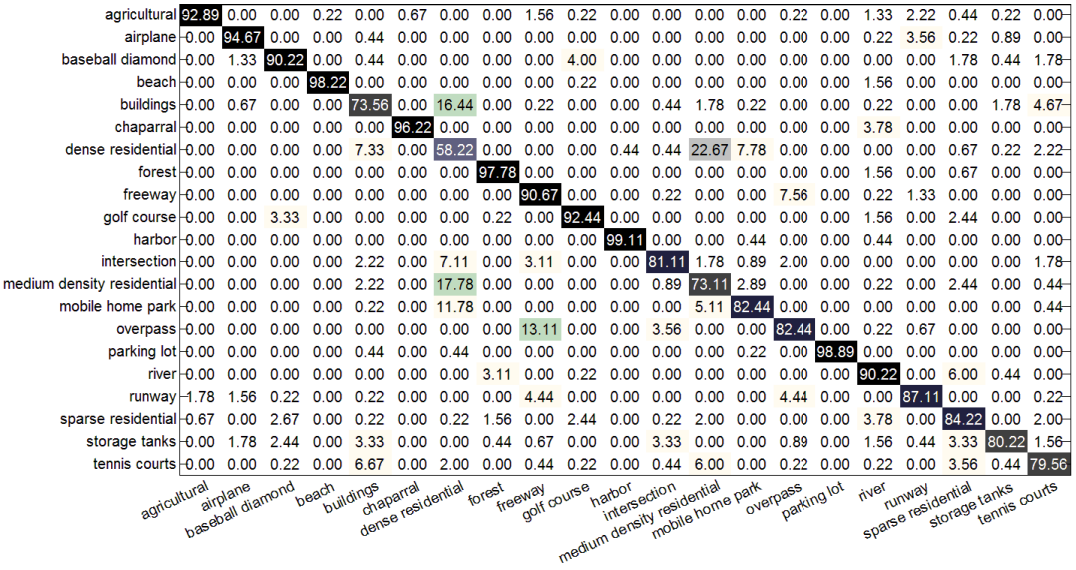
The selection of 64 clusters performs best for both datasets, and a large number of GMM clusters may generate overfitting because the classification accuracy decreases with the increase in the number of clusters. Few GMM clusters are generally beneficial for concatenating different feature encodings because they can avoid the explosive growth of dimension for the final image representation.

E. ANALYSIS OF CONFUSION MATRIX

We further analyze the confusion matrix of both datasets to evaluate the classification performance of each class. Figure 7 shows two confusion matrices of the 21-class dataset for "aggregate strategy 1" (Fig. 7(a)) and "aggregate strategy 2" with the use of 10% of training images (Fig. 7(b)). The rows in Fig. 7 represent the ground truth, and the columns represent the classification results, which are given as percentages. The average classification accuracies (corresponding to the average results of the diagonal elements) of "aggregate strategy 1" and "aggregate strategy 2" are



(a)



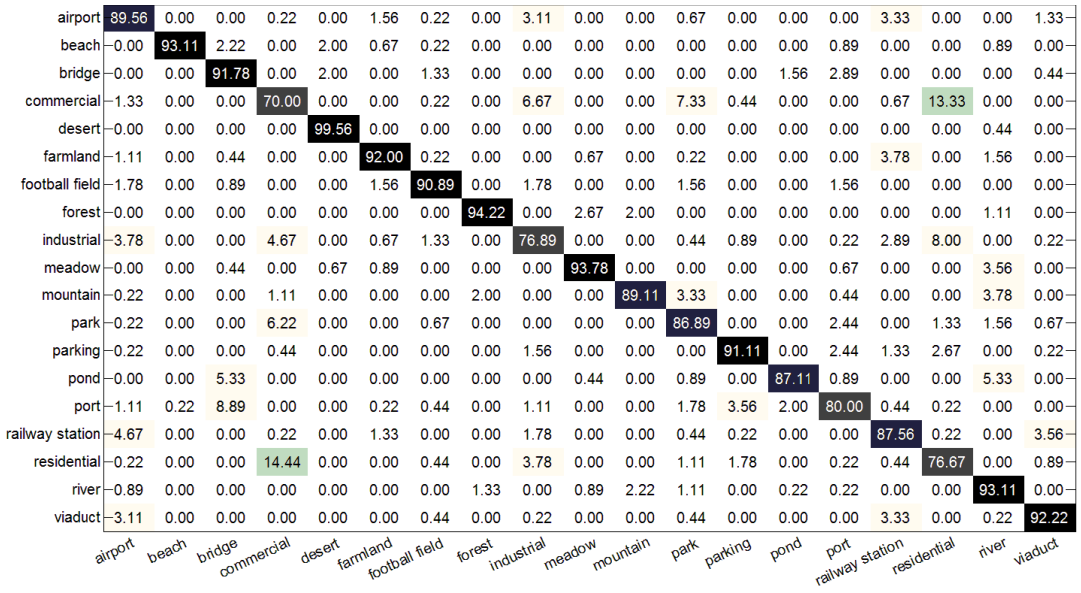
(b)

FIGURE 7. Confusion matrices for the 21-class dataset using 10% of training images. (a) "aggregate strategy 1" (85.09%). (b) "aggregate strategy 2" (86.83%).

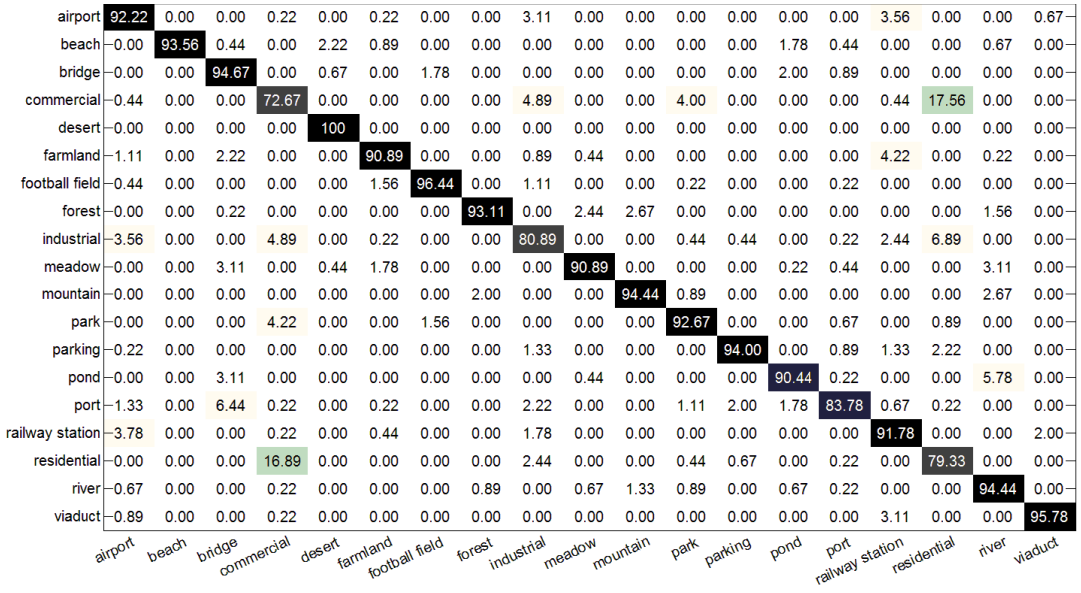
85.08% and 86.83%, respectively. Both "aggregate strategy 1" and "aggregate strategy 2" perform efficiently on the classes, such as "agricultural," "airplane," "beach," "chaparral," "forest," "golf course," "harbor," and "parking lot." These classes contain significant texture features or spatial structures. Although some classes contain objects with different scales, e.g., aircrafts with different sizes exist in the category of "airplane" (91.11%, 94.67%), the proposed method performs efficiently on these classes because the image pyramid significantly improves the local description.

The two confusion matrices shown in Fig. 7 perform poor on some classes, such as "buildings," "dense residential," and "medium density residential." The reason for these poorly performed classes is that numerous buildings exist in these classes and the buildings have high similarities across classes, thereby resulting in a confused classification among these classes.

Figure 8 shows two confusion matrices of the 19-class dataset for "aggregate strategy 1" (Fig. 8(a)) and "aggregate strategy 2" (Fig. 8(b)) with the use of 10% of training images. The meanings of the rows and columns shown in Fig. 8 are the same as those in Fig. 7. The average classification accuracies for "aggregate strategy 1" and "aggregate strategy 2" are 88.19% and 90.63%, respectively. Both "aggregate strategy 1" and "aggregate strategy 2" perform efficiently on the classes, such as "airport," "beach," "desert,"



(a)



(b)

FIGURE 8. **Confusion matrices for the 19-class dataset using 10% of training images. (a) "aggregate strategy 1" (88.19%). (b) "aggregate strategy 2" (90.63%).**

"farmland," "forest," "meadow," "mountain," "parking," "river," and "viaduct." The characteristics of these well-performed classes are close to the well-performed classes (e.g., significant textures, aircrafts) in the 21-class dataset. Both "aggregate strategy 1" and "aggregate strategy 2" perform poorly on "commercial," "industrial," and "residential." The reason can be attributed to that various buildings exist in "commercial," "industrial," and "residential," thereby resulting in a confused classification, similar to that in the 21-class dataset.

Overall, "aggregate strategy 2" indicates better performances than "aggregate strategy 1" on both datasets. The extraction of distinguishable local descriptions, which are robust to building regions, is an effective way to further improve the performance of the proposed method.

F. COMPARISONS BETWEEN CONVOLUTIONAL AND FULLY CONNECTED LAYERS

Given a CNN model, extracting the activation vectors from its fully connected layer as the global features (followed by linear SVM classification) can achieve state-of-the-art performances for RSI classification [32], [33], compared with shallow representations, such as low-level and mid-level features.

TABLE 1. Performance comparisons between fully-connected layer and convolutional layer on the 21-class and 19-class datasets. All results are given as percentages.

	21-class dataset			19-class dataset		
Number of training images per class	10	50	80	10	25	40
CaffeNet (fully connected layer)	82.25%	92.13%	93.33%	86.52%	94.86%	95.78%
VGG-VD16 (fully connected layer)	81.05%	91.14%	92.80%	85.02%	93.22%	95.36%
CaffeNet (255-dim convolutional layer) (Proposed)	78.86%	94.23%	96.00%	83.87%	95.83%	96.89%
VGG-VD16 (511-dim convolutional layer) (Proposed)	84.47%	94.35%	96.14%	88.24%	95.36%	96.21%
Aggregate strategy 1 (Proposed)	85.09%	95.84%	97.28%	88.19%	97.20%	97.89%
Aggregate strategy 2 (Proposed)	86.83%	96.25%	97.40%	90.63%	96.40%	97.52%

In this section, the features extracted from the convolutional and fully connected layers are compared, as shown in Table 1. For the features extracted from the fully connected layer, we can obtain CaffeNet-based and VGG-VD16-based activation vectors, namely, "CaffeNet (fully connected layer)" and "VGG-VD16 (fully connected layer)," respectively. Each input image is resized to the input size of the corresponding CNN model to extract the activation vectors, and 4096-dimensional activation vectors with L2 normalization can then be extracted from the last fully connected layer (except for the classification layer). Although each CNN contains two fully connected layers (except for the classification layer), both layers have similar classification performances. For the deeply local descriptors extracted from the convolutional layer, we can obtain two types of single CNN-based variants for Hellinger kernel and PCA transformation (255 dimensions for CaffeNet-based descriptors and 511 dimensions for VGG-VD16-based descriptors), namely, "CaffeNet (255-dim convolutional layer)" and "VGG-VD16 (511-dim convolutional layer)."

With the use of Hellinger kernel and PCA transformation, CNN-based local description outperforms the activation vector (global description) extracted from the fully connected layer in most cases. For example, "VGG-VD16 (511-dim convolutional layer)" indicates accuracy advantages over "VGG-VD16 (fully connected layer)" under 10%, 50%, and 80% of training images. "CaffeNet (255-dim convolutional layer)" indicates accuracy advantages over "CaffeNet (fully connected layer)" under 50% and 80% of training images. The proposed aggregate strategies, "aggregate strategy 1" and "aggregate strategy 2," indicate accuracy advantages over "CaffeNet (255-dim convolutional layer)" and "VGG-VD16 (511-dim convolutional layer)" under different numbers of training images.

IV. CONCLUSIONS

We investigate deeply local descriptions based on convolutional features extracted from the last convolutional layer of CNN models in this study. Two types of pretrained CNN, CaffeNet and VGG-VD16, are used to extract deeply local descriptors, followed by Hellinger kernel and PCA transformation. Two aggregate strategies are proposed to form a global representation from the deeply local descriptors. "Aggregate strategy 1" adopts both CaffeNet and VGG-VD16 to increase the number of local descriptors, followed by feature encoding. "Aggregate strategy 2" encodes CaffeNet-based and VGG-VD16-based local descriptors into two global representations through Fisher encoding, followed by feature concatenation. Experiments on two RSI datasets illustrate that Hellinger kernel, PCA transformation, and the two aggregate strategies can substantially improve classification accuracy.

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 *et al.*, "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] Z. Lv *et al.*, "Managing big city information based on WebVRGIS," *IEEE Access*, vol. 4, pp. 407–415, 2016.
[4] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-generation big data analytics: State of the art, challenges, and future research topics," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 1891–1899, Aug. 2017.
[5] T. Zhou, H. Bhaskar, F. Liu, J. Yang, and P. Cai, "Online learning and joint optimization of combined spatial-temporal models for robust visual tracking," *Neurocomputing*, vol. 226, pp. 221–237, Feb. 2016.
[6] Z. Lv, T. Yin, X. Zhang, H. Song, and G. Chen, "Virtual reality smart city based on WebVRGIS," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1015–1024, Dec. 2016.
[7] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM 18th SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, 2010, pp. 270–279.
[8] V. Risojević, S. Momić, and Z. Babić, "Gabor descriptors for aerial image classification," in *Proc. Int. Conf. Adapt. Natural Comput. Algorithms*, 2011, pp. 51–60.
[9] J. Ren, X. Jiang, and J. Yuan, "Learning LBP structure by maximizing the conditional mutual information," *Pattern Recognit*, vol. 48, no. 10, pp. 3180–3190, 2015.
[10] 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.
[11] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Stat. Learn. Comput. Vis. (ECCV)*, 2004, vol. 44. no. 247, pp. 1–22.
[12] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1470–1477.
[13] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun. 2005, pp. 524–531.
[14] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: An evaluation of recent feature encoding methods," in *Proc. Brit. Mach. Vis. Conf.*, 2011, vol. 2. no. 4, pp. 1–12.
[15] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.
[16] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 6314. 2010, pp. 143–156.

[17] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 6315. 2010, pp. 141–154.

[18] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 5304. 2008, pp. 696–709.

[19] 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.

[20] 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.

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

[22] T. Kobayashi, "Dirichlet-based histogram feature transform for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3278–3285.

[23] 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.

[24] L. Wan, N. Liu, Y. Guo, H. Huo, and T. Fang, "Local feature representation based on linear filtering with feature pooling and divisive normalization for remote sensing image classification," *J. Appl. Remote Sens.*, vol. 11, no. 1, p. 016017, 2017.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

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

[27] K. Simonyan and A. Zisserman. (Apr. 2015). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

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

[29] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[30] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IEEE Conf. Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 4959–4962.

[31] 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.

[32] 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.

[33] 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*, Jun. 2015, pp. 44–51.

[34] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[35] L. Wan, N. Liu, H. Huo, and T. Fang, "Selective convolutional neural networks and cascade classifiers for remote sensing image classification," *Remote Sens. Lett.*, vol. 8, no. 10, pp. 917–926, 2017.

[36] 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.

[37] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Jun./Jul. 2004, pp. 506–513.

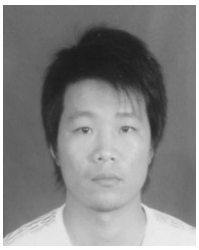
[38] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012 pp. 2911–2918.

[39] Y. Zhang, G. Zhou, J. Jin, Q. Zhao, X. Wang, and A. Cichocki, "Aggregation of sparse linear discriminant analyses for event-related potential classification in brain-computer interface," *Int. J. Neural Syst.*, vol. 24, no. 1, p. 1450003, 2014.

[40] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

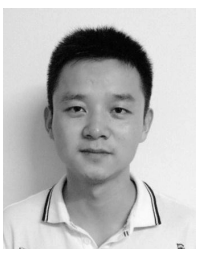
[41] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

NA LIU received the B.S. degree from the School of Electrical Engineering and Automation, Qilu University of Technology, Jinan, China, in 2008, and the M.S. degree from the School of Automation, Nanjing University of Science and Technology, Nanjing, China, in 2010. She is currently pursuing the Ph.D. degree with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. Her research interests include image processing, pattern recognition, machine learning, and object detection.



LIHONG WAN received the B.S. degree in computer application and the M.S. degree in computer application from China Jiliang University, Hangzhou, China, in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. His research interests include compute vision and pattern recognition, neural representation and brain-like computing, machine learning, and visual perception.

YU ZHANG received the Ph.D. degree (Hons.) in control science and engineering from the School of Information Science and Engineering, East China University of Science and Technology (ECUST), Shanghai, China, in 2013. In 2010, he was a Research Associate with the Laboratory for Advanced Brain Signal Processing (hosted by Prof. A. Cichocki, IEEE Fellow), RIKEN Brain Science Institute, Japan. From 2013 to 2016, he was an Assistant Professor with the Department of Automation, ECUST, China. In 2016, he was a Research Fellow with the Biomedical Research Imaging Center (hosted by Prof. D. Shen, IEEE Fellow), University of North Carolina at Chapel Hill, USA. Since 2017, he has been a Research Fellow with the Department of Psychiatry and Behavior Sciences, Stanford University, USA. He is also a Visiting Scientist with the RIKEN Brain Science Institute, Japan. He is the author of over 50 technical papers that have been published in the prestigious Journals, such as the PROCEEDINGS OF THE IEEE, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, and the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, where he is also serving as a Reviewer.



TAO ZHOU received the M.S. degree in computer application technology from Jiangnan University, in 2012, and the Ph.D. degree in pattern recognition and intelligent system from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, in 2016. He is currently a Post-Doctoral Research Fellow in pattern recognition and intelligent system with the same university. His current research interests include object detection, visual tracking and machine learning.

HONG HUO received the B.S. degree in computer application and the M.S. degree in computer application from the Jilin University of Technology (now merged into Jilin University), Changchun, China, in 1995 and 1998, respectively, and the Ph.D. degree in pattern recognition and intelligent system from Shanghai Jiao Tong University, Shanghai, China, in 2014. She has been an Instructor with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University since 2000. Her research interests include image analysis and interpretation, machine learning, and visual perception with applications to remote sensing imagery.

TAO FANG received the B.S. and M.S. degrees in geology and survey from the Xian University of Science and Technology, Xi'an, China, in 1988 and 1991, respectively, and the Ph.D. degree in remote sensing and geographical information system from the China University of Mining and Technology, Beijing, China, in 1996.

From 1996 to 1998, he was a Post-Doctoral Research Fellow in remote sensing and geographical information system with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China. From 1999 to 2000, he was an Assistant Professor with the Department of Electronic Engineering and a Post-Doctoral Research Fellow in electronics with the HDTV Laboratory, University of Electronic Science and Technology of China, Chengdu, China, respectively. He is currently a Professor with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. His research interests include neural representation and brain-like computing, deep learning with applications to remote sensing classification, and object recognition.