IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 3, 2021, accepted January 9, 2021, date of publication January 25, 2021, date of current version February 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052977

A Multi-Level Convolution Pyramid Semantic Fusion Framework for High-Resolution Remote Sensing Image Scene Classification and Annotation

XIONGLI SUN[1], QIQI ZHU[2,3], (Member, IEEE), AND QIANQING QIN[1]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430072, China
[2]School of Geography and Information Engineering, China University of Geoscience, Wuhan 430074, China
[3]State Key Laboratory of Resources and Environmental Information System, Beijing, China

Corresponding author: Qianqing Qin (00201541@whu.edu.cn)

This work was supported in part by the Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing under No. KLIGIP-2019A02, and in part by the State Key Laboratory of Resources and Environmental Information System.

ABSTRACT High spatial resolution (HSR) imagery scene classification has become a hot research topic in remote sensing. Scene classification method based on the handcrafted features, such as the bag-of-visual-words (BoVW) model, describes an image by extracting local features of the scene and mapping them to the dictionary space, but usually uses a shallow structure and loses the spatial distribution characteristics of the scene. The method based on deep learning extracts hierarchical features to describe the scene, which can maintain the spatial position information well. However, deep features in different levels have scale recognition restrictions for multi-scale ground objects, and cannot understand complex scenes well. In this paper, the multi-level convolutional pyramid semantic fusion (MCPSF) framework is proposed for HSR imagery scene classification. Differing from previous scene classification methods, which integrate the feature of different levels directly, of which the fusion features have large differences in both sparsity and eigenvalue magnitude, MCPSF integrates multi-level semantic features extracted by BoVW model and convolutional neural network (CNN) model. In MCPSF, two convolution pyramid feature expression strategies are proposed to enhance the ability of capturing multi-scale land objects, i.e., local and convolutional pyramid based BoVW (LCPB) model and local and convolutional pyramid based pooling-stretched (LCPP) model. The effectiveness of the proposed method is verified on 21-class UC Merced (UCM) dataset and 30-class Aerial Image Dataset (AID). The framework was also transferred toa case study of scene annotation in Wuhan. The proposed framework significantly improves the performance when compared with other state-of-the-art methods.

INDEX TERMS High spatial resolution image, scene classification, bag of visual words, feature pyramid, multi-level, remote sensing.

I. INTRODUCTION

With the development of remote sensing satellite technology, a large number of high-resolution remote sensing images with rich spectral and spatial information can be obtained. Diverse spatial structures form high-level scene semantic information, which can be available for a wide range of applications, e.g., digital city construction and environmental protection. However, as the resolution increases, it also brings about problems such as low inter-class disparity and high intra-class variability [1]. The pixel-oriented scene classification method has been transferred to the object-based information extraction method [2]–[4]. However, the object-based method usually focus on the categories of ground objects (such as buildings), and cannot obtain the scene categories composed of various objects with specific spatial relations (such as residential areas) [5]. In this way, the semantic gap between low-level features and high-level semantics is formed [6]. In order to bridge the semantic gap, scene classification method has become one of the most challenging topics in the field of remote sensing.

Compared to natural images, HSR remote sensing images have three distinct characteristics in the aspects of the distribution of ground objects, light condition and channel

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

numbers. In general, HSR remote sensing images are taken by looking down from above with sensors on aircraft or satellites, while natural images are taken with cameras by looking straight forward. In contrast to natural images, HSR images contain more ground objects and rarely shows significant spatial direction relations, such as up and down, or back and forth. In addition, HSR images are always taken in good light condition while natural images could be got in dawn or dusk. In other words, the spectral information is much more import for HSR images in scene classification than natural images. Thirdly, the channel number of natural color images is three, of which is much less than HSR images. Considering these three distinct characteristics of HSR imagery, the feature extractor and scene classification strategy should be carefully chosen instead of just using the methods which the natural imagery are always dealt with.

Different from pixels or simple objects in images, semantic scenes of HSR images [7], [8] refer to specific geographic regions composed of several ground objects. Different geospatial distributions of the same object may lead to different high-level semantic interpretation. Hence, scene classification of HSR images obtain the semantic information of geographical regions by automatically labeling an HSR image according to the geographical composition and spatial distribution of ground objects [7], [9]. The scene classification procedure is to extract the features and relate them to the high-level semantics of the whole scene. Therefore, it is required to propose an excellent feature extraction framework to meet the demands of complex scene recognition.

Scene classification based on low-level features usually utilized either local or global color, structure and texture features to describe the HSR images. For instance, [10] proposed a content-based soft annotation method, which extracts and uses SVM for training to define semantic labels for images. [11] classified the IKONOS satellite images using scale invariant feature transform (SIFT) and Gabor texture features. Due to the diversity of object classes and the complexity of spatial location in HSR image scenes, it is difficult to effectively describe the scene using low-level features. To enhance the scene classification ability, the bag of visual word (BoVW) model [12] became the basis of the mid-level feature scene classification method [13]. For instance, [14] used the BoVW model to extract the visual words in the scene and realized the feature extraction based on urban area segmentation. The probabilistic topic models (PTM), such as probabilistic latent semantic analysis (pLSA) [15] and latent Dirichlet allocation (LDA) [16], were developed to extend the BoVW model. The PTM were also introduced to the HSR image scene classification [17]. However, the model structure of the mid-level feature based method is usually shallow and the extracted features are local, thus lacking descriptions of global high-level semantics.

By making full use of the low-level features to understand the scene, the mid-level features can achieve better scene expression than the traditional low-level features. However, the mid-level features still have drawbacks in scene classification, such as ignoring the relationship between low-level features [18], [19] the lack of transferability between HSR images [20], [21], and low universality and efficiency. Based on the automatic feature learning and representation framework, deep learning is able to solve the problems existing in mid-level features. Deep learning technology has developed rapidly in recent years, mainly in video analysis [22], object detection [23], especially in image classification [24]. Convolutional neural network (CNN), as the most popular deep learning-based network, has shown amazing performance on different datasets. Castelluccio *et al.* [25] trained CaffeNet and GoogLeNet through complete training and fine-tuning, and verified the effectiveness of the method on two HSR datasets. Zhang *et al.* [26] proposed a mixed classification architecture, combined CNN and pixel-based shallow structure according to decision rules, and verified the method on urban and rural image scenes respectively. Lin *et al.* [27] used a Generative Adversarial Network (GAN) to implement an unsupervised training scene classification method for small samples of HSR images. However, CNN-based methods require a large number of sample data to train models, and remote sensing image samples tend to be small. Therefore, transfer learning-based scene classification methods have been proved to be an effective method for HSR image classification. The pre-trained CNN is usually utilized to extract deep features to describe the scenes, and can then be fed into the feature coding or classifier procedure [20], [28]. However, CNN may lack attention to local details of the scene, and single convolutional or fully-connected level cannot interpret the multi-scale objects in the scene. Methods combining the mid-level and deep features have been proposed and achieved satisfactory results. However, the magnitude and sparsity of the feature eigenvalues are significantly different, and the ability to describe scenes is limited when features are concatenated directly. Different features lead to different feature descriptors, and they usually differ greatly. When using K-Means clustering to quantize the vector concatenated by multiple feature descriptors, such as spectral, texture, structure feature, they interact on each other and the clustering is inadequate to fuse the complementary characteristics of different features. Integrating features that differ greatly in magnitude, such as 0.001 and 1020, will affect the performance of the classifier.

In this work, considering the existing shortcomings in current scene classification methods, a multi-level convolutional pyramid semantic fusion (MCPSF) framework is proposed. In MCPSF, BoVW and CNN are naturally integrated to build scene features. The combined features can capture comprehensive information for HSR scenes from local structure, spectral, and global deep perspectives. The handcrafted features are first obtained by sampling with the direction gradient and gray value statistics are calculated, representing the structure and spectral characteristics of the scene respectively. The BoVW is utilized to enhance the expression ability of low-level features, such as the SIFT and the mean and standard deviation (MSD)-based spectral features. To obtain

the deep features, a pretrained CNN is employed. Since pyramids constructed from multi-scale images can improve the accuracy of scene recognition, pretrained CNN is utilized to extract multi-level convolution features. However, because of the difference between the mid-level and deep features, directly integrating the two stage features is difficult and cannot improve the semantic description of scene. In MCPSF, the BoVW and Pooling method are utilized to implement feature representation of convolution features with pyramid characteristics, which can solve the problems of sparsity and magnitude respectively. Finally, the multi-level features implement the fusion based on two strategies and support vector machine (SVM) is utilized for scene classification.

The main contributions of this paper are as follows.

1) An efficient HSR image scene classification framework. The MCPSF framework is proposed to better distinguish complex scenes composed of diverse ground objects. Considering the characteristics of HSR image, MCPSF capture the local mean and standard deviation (MSD) of the image as spectral features, and the local directional gradients scale invariant feature transform (SIFT) as structural features. In addition, deep feature is utilized to obtain high-level semantic representations of HSR images. Features of different spaces are naturally fused by efficient transformation. Finally, automatic scene learning is implement based on comprehensive features.

2) Improved convolution pyramid feature expression method for scene understanding. The deep features extracted from the convolutional layer of CNN model are usually a 3-dimensional matrix, and cannot be used to describe the scene directly. This work is inspired by the image pyramid with the ability to recognize multi-scale objects, and the knowledge that convolutional layer of the CNN model has the similar characteristics. In this work, multi-level convolution features with pyramid characteristics are extracted, and the BoVW and global Pooling methods based on equalized sampling are used for dimension reduction expression.

3) Efficient feature fusion strategy for features dimensional and magnitude difference. In this paper, the feature value contrast stretching method is adopted to reduce the magnitude of different features, and the processed features even have the same magnitude. In addition, an improved visual words extraction method is adopted. All features at the same position in the convolutional feature map are regarded as one visual word, and the words for all scenes are extracted and rearranged to form a visual dictionary by clustering method.

Comprehensive evaluations on three distinct datasets, i.e., the 21-class UC Merced (UCM) dataset and the challenging 30-class Aerial Image data set (AID), confirm the effectiveness of the MCPSF framework. In addition, scene annotation of a large HSR image also confirms the effectiveness of MCPSF.

The remainder of this paper is organized as follows. Section 2 describes scene classification methods based on the BoVW model and CNNs in detail. In Section 3, the proposed framework MCPSF for HSR imagery scene classification is introduced. The experimental results and analysis are reported in the Section 4. Section 5 discussed and analyzed the sensitivity of the experimental parameters. Finally, the conclusions are provided in Section 6.

II. BACKGROUND

Scene classification methods based on mid-level features and deep learning are the main methods to bridge the semantic gap [6]. In this Section, the feature selected for HSR scene classification, the classic mid-level and deep feature based scene classification methods, i.e., BoVW and CNN, are briefly introduced.

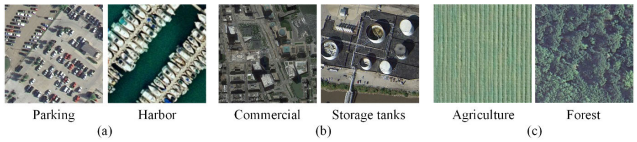

FIGURE 1. HSR scenes that cannot be accurately distinguished by a single feature: (a) importance of the spectral characteristics for HSR images; (b) importance of the structural characteristics for HSR images; (c) importance of the global characteristics for HSR images.

A. FEATURES SELECTED FOR SCENE CLASSIFICATION

Remote sensing scenes tend to be complicated due to the increasing resolution of remote sensing images. As shown in Fig. 1(a), it is difficult to distinguish parking lot and harbor directly by structural and textual features. Due to the different spectral characteristics of the ocean and the road, the spectral features play an important role. In Fig. 1(b), the spectral features of commercial and storage tanks are similar, the main difference lies in the structure level. In Fig. 1(c), the agriculture and forest scenes are similar in both spectral and structural characteristics, while global characteristics play a crucial role. Therefore, the features selected in this paper include Mean and Standard (MSD) features focusing on the local spectral characteristics, the SIFT features focusing on local structural characteristics, and the deep convolutional features focusing on the global features of the scene.

B. SCENE CLASSIFICATION BASED ON BOVW MODEL

Artificial design features [29] refer to the extraction of the low-level features of the scene, which can be roughly divided into three categories, spectral, texture, and structural features. Spectral features of scene classification are usually based on the mean and standard deviation of gray values [18]. Local Binary Pattern (LBP) [30] and Gray-level Co-occurrence Matrix (GLCM) [31] are usually used as the texture features of the scene. Structural features such as the SIFT [32] feature proposed by Professor David G. Lowe have been widely used in HSR scene classification.

The low-level features are dense with redundant information [33], and the complex scene cannot be effectively distinguished. Therefore, scene classification method based on mid-level features is introduced, establishing the relationship

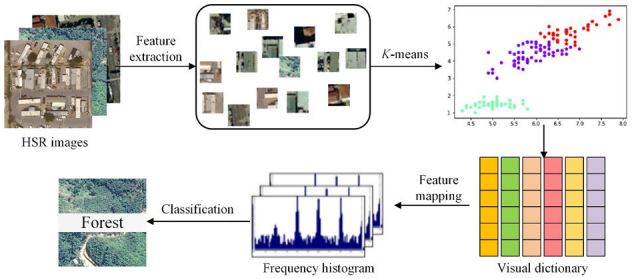

FIGURE 2. HSR image scene classification procedure based on BoVW mode.

between low-level features and high-level semantics. The methods for constructing mid-level features is developed on the basis of the BoVW model [11]. The procedure of BoVW model is shown in Fig 2.

The BoVW model was first proposed and applied to the field of text processing, and is widely used in scene classification of HSR images [34]. BoVW treats the two-dimensional HSR images as a collection of independent words without grammatical or lexical order, and obtain a set of representative words as the visual dictionary by k-means clustering. K cluster centers are formed after multiple iterations, and then a visual dictionary can be obtained. Given a dataset consisting of I images, each scene can be represented by K visual words in the dictionary. By counting the frequency of each visual word, the frequency histogram of each image is constructed. In this way, the image is converted into a one-dimensional vector, and the length of the vector is the number of words in the visual dictionary. The BoVW-based scene classification method obtains the mid-level feature by mapping the local low-level feature to the corresponding parameter space. However, the scene classification method based on the BoVW model ignores the spatial position relationship of the scene.

C. SCENE CLASSIFICATION BASED ON DEEP LEARNING

The strategies of HRS imagery scene classification based on deep CNN can be categorized as follows: (1) Training from scratch [35]–[37]. (2) Semi-training parameter fine-tuning [20]. (3) Deep feature vector extraction [21], [25].

Deep learning [38] technology has been outstanding in different fields, such as artificial intelligence, speech, image processing. Various CNN-based methods have dominated the field of remote sensing image scene classification [37], [39]–[41]. The neural network of CNN can automatically learn and update parameters in training iterations and fine-tune the parameters under supervised training to obtain a model with good performance. CNN is mainly composed of convolutional layers, pooling layers and fully connected layers, and it is a supervised learning network based on error back propagation (BP) framework. Scene classification methods based on classic CNN can be divided into three types: (a) full training of a CNN model from scratch [35], [36]; (b) parameter fine-tuning based on pre-trained CNN model [20]; (c) feature vector extraction based

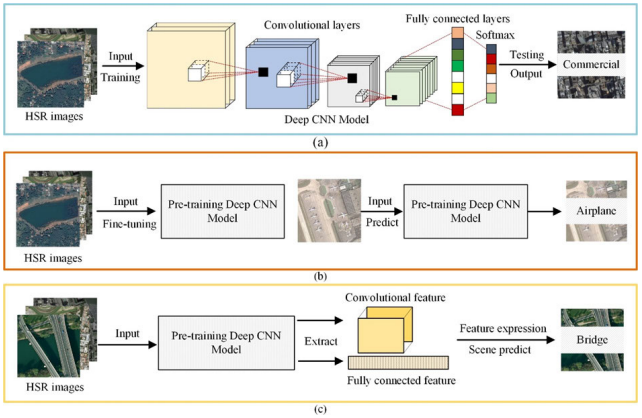

FIGURE 3. HSR image scene classification procedure based on CNN: (a) Full training. (b) fine-tuning. (c) feature vector.

on pre-trained CNN model [21], [25]. The scene classification procedure of the three types is shown in Fig. 3 (a), Fig. 3 (b), and Fig. 3 (c), respectively.

Training the CNN model from scratch needs to initialize all the weights randomly, and then train the model based on HSR images. Due to the depth and complexity of the CNN model, the network may contain thousands of parameters, and large amounts of scene images are required for training. On the one hand, the tasks often involve multiple iterations to get a more suitable model, which takes a lot of time and space [21]. On the other hand, training the CNN model is a supervised method that requires quantities of artificially labeled semantic scenes, which consumes huge manpower and resources. When the image dataset is large enough, this method can achieve the best results because the model training is performed by HSR images. But in fact, the cardinality of remote sensing images tends to be small, thusthe data size required to train CNN model cannot be achieved. Therefore, the scene classification task for remote sensing images can be performed by an efficient migration learning method.

Parameter fine-tuning based on pre-trained CNN model involves fine-tuning parameters and extracting feature vectors. The former keeps some of the parameters (usually the first few layers) of the model unchanged, and adjusts the parameters of the partial level to achieve the effect of the training model. The first few layers of the CNN model are often general-purpose features. The subsequent levels are often characterized by categories. Therefore, the method controls the first few levels to be unchanged, and the HSR images are appropriately trained for subsequent levels. Compared with the fully trained CNN model, this method requires a smaller amount of training samples, and the network is easier to converge, reducing training time and space consumption.

Feature vector mode treats the pre-trained CNN model as a feature extractor for arbitrary images. The features extracted from the pre-trained CNN model are divided into low-level and high-level, the low-level features can be re-extracted by mid-level feature extraction methods such as BoVW to reduce dimensions; The high-level features can be directly

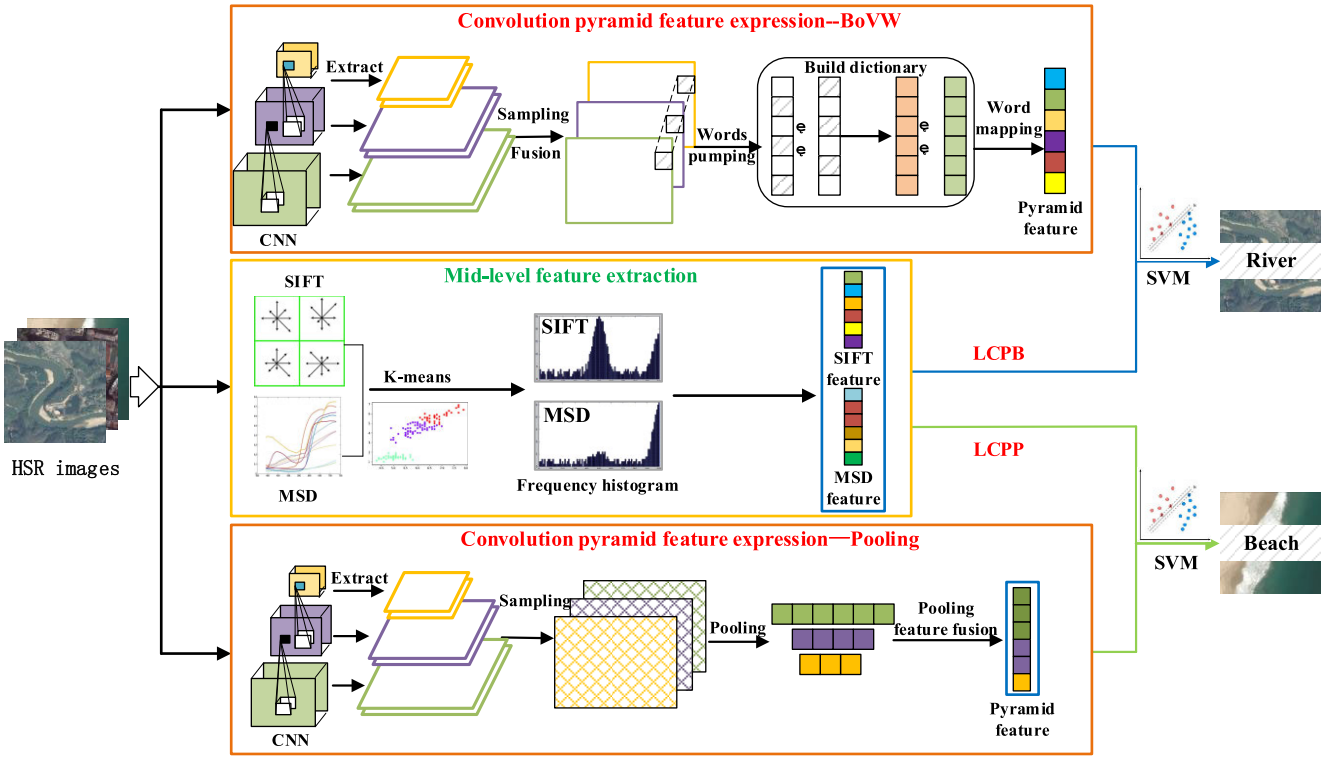

FIGURE 4. The flowchart of the proposed MCPSF.

input into the classifier for scene training and prediction. This method does not require training or fine-tuning parameters, and can be combined with traditional scene classification with the advantage of simple scalability [20]. In this paper, feature extraction method is employed and the VGG-19 network is regarded as the feature extractor, which is pre-trained by the ImageNet dataset. The convolution layer features with pyramid characteristics were extracted from the pre-trained CNN model.

III. METHODOLOGY

In this paper, the MCPSF framework is carefully designed to enhance the performance of scene classification for HSR images, which includes four main tasks. First, mid-level and deep features are extracted by BoVW and CNN-based method, with the deep feature enhanced by pyramid fusion. Continuously, the mid-level feature and convolution pyramid feature are fused based on LCPB (Local Convolutional Pyramid BoVW, LCPB) and LCPP (Local Convolutional Pyramid Pooling, LCPP) Strategy. Finally, scene labels are acquired by training fusion features with SVM [42] classifiers. The flowchart of the proposed MCPSF is shown in Fig. 4.

A. MID-LEVEL FEATURE GENERATION BASED ON LOCAL EXPRESSION

The low-level features extracted in MCPSF include SIFT based structural feature and MSD based spectral feature. The image is divided into uniform sampling patches, each of which is described by local features. The MSD feature is calculated using the mean and standard deviation of the gray value, and the SIFT feature is calculated based on the direction gradient of the key points. In addition, previous research by Fei-Fei and Perona [43] indicated that uniform grid sampling has better performance in image classification than random sampling when extracting features. As seen in Fig. 5, the patches are acquired with the patch spacing during the sampling process for the MSD and SIFT features. The local refinement level of the scene is determined by the patch size, and the sampling frequency is determined by the patch spacing.

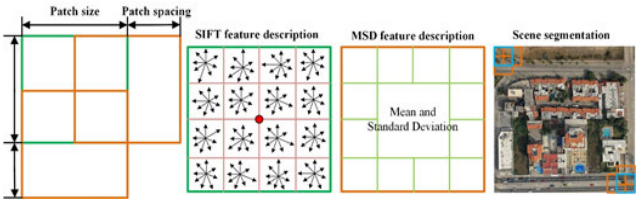

FIGURE 5. Patch sampling size and patch spacing for low-level feature extraction.

In MCPSF, SIFT features is utilized to describe the local structural properties of image scene. SIFT is invariant to image rotation and scale, and have a certain robustness to illumination and viewing angle transformation [32]. To acquire strong scene recognition ability, in this paper, dense SIFT is extracted based on uniform grid sampling, covering all

locations of the scene. The generation of SIFT features can be divided into four steps: (1) Searching the key points in each sampling space; (2) Positioning stable key points; (3) Determining one or more directions for key points; (4) Calculating the direction gradient in the neighborhood of key points. The sampling patch is divided into 4×4 neighborhoods, where the gradients in 8 directions are calculated. In this way, $4 \times 4 \times 8 = 128$ - dimension vectors is obtained to describe each image.

Similarly, MSD features is utilized to describe the local spectral properties of the scene, which reflects the brightness and variation of the image as well as the object composition. Based on the sampling method, the MSD features calculate the first-order and second-order statistics for each channel of the sampling patches, i.e. the mean and standard deviation values. Taking an image with three channels as an example, the 6-dimensional spectral vector can be obtained within each sampling patch. We let I be the number of pixels in the patch, and represents the value of the j-th band value of the i-th pixel. In this way, the mean (M_j) and standard deviation (SD_j) of the sample patch in the j-th band can be obtained as follows

$$M_j = \frac{\sum_{i=1}^{I} p_{ij}}{I} \tag{1}$$

$$SD_j = \sqrt{\frac{\sum_{i=1}^{I} \left(p_{ij} - M_j\right)^2}{I}} \tag{2}$$

After acquiring the MSD and SIFT features, K-means clustering method is usually utilized to construct the visual dictionary. The image patches are clustered to generate K cluster centers, and a visual dictionary with K visual words is obtained. The dictionary can be denoted as $D = \{V_1, V_2, \ldots\ldots V_k\}$, where $V_i (1 \leq i \leq k)$ represents the visual words in the visual dictionary. Each patch in the scene is mapped to the dictionary. The image patch with the smallest distance to one of the visual words can be allocated to it. The frequency of word occurrence is then calculated for all the visual words, and an image can be transformed to a 1-D histograms.

B. DEEP CONVOLUTION PYRAMID FEATURE GENERATION BY CNN

Mid-level feature based method is able to capture the local significant characteristics for the scene, but it ignores the spatial information. Constructed based on the hierarchical feature learning structure, CNN can effectively capture the spatial arrangements inside the scenes. To obtain deep features, a pre-trained VGG-19 based CNN model is employed in MCPSF, which is trained by ImageNet based on the Tensorflow framework. VGGNet [44], a network developed from AlexNet [45], mainly modifies the following two aspects from the AlexNet: 1) The entire network uses the filter size of 3×3 and maximum pool size of 2×2; 2) The network structure is deepened to improve the performance. VGG-19 [44] consists of 16 convolutional layers and 3 fully-connected layers. The convolutional layer is divided into 5 levels, and each level has a maximum pooling layer to reduce the feature map size. As a feature extractor, VGG-19 allows arbitrary level of features to be extracted from the network. Before inputting the images into the CNN model for convolution and pooling, the grayscale value of each image is 255-normailized. Then a convolution filter is used to slide on the image for convolution, and the convolution feature map is regarded as the global feature of the scene. Since the first two levels of feature maps are large in size and contain more redundant information, and previous research have proved that the last layer of each convolution level of VGG-19 is more adequate for feature description [20], convolution layers in 3-5 level are extracted in MCPSF.

The scene of HSR images tend to be complicated due to the scale and content diversity of the ground objects. Feature pyramids are crucial to the identification of different-scale objects [27]. Traditional methods are utilized to construct feature pyramids based on image pyramids, which are mainly applied to artificial designed feature extraction. CNN has currently become the mainstream in the field of image processing, the image features extracted by CNN have good robustness, and the hierarchical structure of CNN can fit the feature pyramid well. The convolution features of each level have the following characteristics. The low-level convolution layer focuses on describing the linear and edge features; the mid-level convolution layer focuses on the characteristics of the object; the high-level convolution layer focuses on describing the overall information of the scene [44]. Therefore, the convolution pyramid structure of VGG-19 is adopted to replace the image pyramid, and the layers can be divided into 5 levels. This paper uses the last layer of the 3rd to 5th convolutional levels to construct the feature pyramid.

The convolution features with pyramid characteristics extracted have different sizes. In order to improve the expression ability of this feature, the nearest neighbor interpolation method is adopted to sample the feature maps and resize them to the same size. Then, the feature maps are longitudinally fused to obtain a deeper convolution feature, which has both pyramid and deep convolution characteristics. The nearest method takes the pixel value closest to a pixel position in the image as the new value of the pixel. The advantages of this method are simple, efficient, and does not change the original image grid value. We let Org_x and Org_y be the coordinates of the original image, Obj_x and Obj_y be the coordinates of target image. The pixel value of the target image can be filled by the original image, and the coordinate relationship between two images is shown in Eq. (2), where Org_w and Org_h represents the width and height of the images.

$$\begin{aligned} Org_x &= \left[Obj_x \times \left(\frac{Org_w}{Obj_w} \right) \right], \\ Org_y &= \left[Obj_y \times \left(\frac{Org_h}{Obj_h} \right) \right], P_{(Obj_x, Obj_y)} = P_{(Org_x, Org_y)} \end{aligned} \tag{3}$$

C. CONVOLUTION PYRAMID FEATURE EXPRESSION BASED ON BOVW AND POOLING

The deep convolution feature extracted from the CNN model has a size of n × x × y × M, where n denotes the number of images, x, y denotes the size of feature map, and M denotes the number of feature map. The dimensions of the 5 convolution levels of the VGG-19 [44] network model are shown in Table 1.

TABLE 1. Convolution feature map size of different level for VGG-19.

No	Layer Level	Size
1	Conv1	64×224×224
2	Conv2	128×112×112
3	Conv3	256×56×56
4	Conv4	512×28×28
5	Conv5	512×14×14

Different from the feature matrix obtained by the low-level feature extraction, the convolution feature has 3 dimensions extracted from the CNN of each scene, which can be expressed as $N^l \times n^l \times n^l$, where N^l denotes the depth of the convolution feature map, and $n^l \times n^l$ denotes the size of single convolution feature map. According to the characteristics of the convolution operation, each position of the current feature map is a result of convolution calculation of the local region of the previous layer.

The BoVW model is used to re-express the feature, and the feature value is expressed as the frequency of the word, which makes the feature of the scene more abstract and loses the spatial position information of the pyramid [18], [46]. However, the Pooling method can maintain the essential characteristics of the scene well, which can be described as follows. The convolutional feature maps of different scales are sampled, and then the feature maps of the same size are longitudinally connected. After obtaining the feature maps with the same size and deeper dimensions, the average value of all the feature values of each feature map is calculated. Then all the mean values are connected into a one-dimensional vector, which is the convolution pyramid feature expression obtained by the pooling method. In this way, the feature is more streamlined, has lower dimensions, less redundant information, and can reflect the essential characteristics of the scene. The experimental results prove that the feature expression ability is stronger. By comparing the experimental results, the classification effect of the convolutional pyramid feature scene is better than that of single deep convolution feature. The flowchart is shown in Fig. 6 (b).

D. SCENE CLASSIFICATION BASED ON MULTI-LEVEL FEATURE FUSION

In MCPSF, mid-level features SIFT, MSD and deep convolutional pyramid features based on two enhancement strategies implement two fusion strategies respectively [1]. These features are denoted as F_s, F_m, F_{pb}, F_{pp}. Two scene understanding frameworks LCPB and LCPP have been established for scene classification of HSR images. In LCPB, the

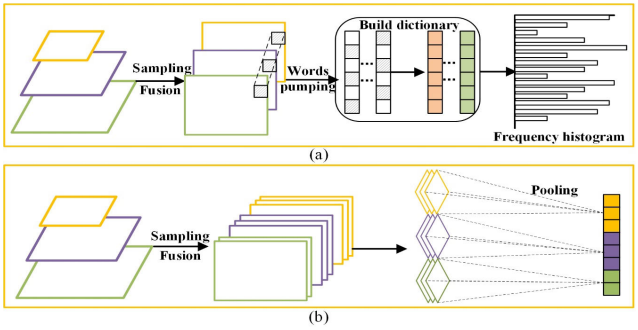

FIGURE 6. The procedure of deep convolution feature expression. (a) The building of deep feature dictionary. (b) The average values of all the feature maps with same size are connected into one vector.

SIFT, MSD, and convolution pyramid features re-extracted by BoVW model are combined. In LCPP, the SIFT, MSD, and improved convolution pyramid features by pooling processing model are combined. The dimension of mid-level features encoded by BoVW model is L, the convolutional pyramid enhanced feature dimension extracted by BoVW model is C, and the feature dimension obtained by Pooling is P. The mid-level features obtained by BoVW model represent the occurrence times of visual words, ranging from 0-K. K represents the size of the visual dictionary, the magnitude is large, and the convolution layer feature values tend to be small. Therefore, the convolutional pyramid enhancement features are normalized and stretched [1]. The feature value range is normalized to 0∼255, consistent with the level of feature retention processed by the BoVW model.

The low-level features can be expressed as Eq. (3), where S_i and M_i represent the number of occurrences of the visual word. s_i and m_i represent the normalized feature representation.

$$\begin{aligned} L &= (s_1, \cdots s_k, m_1 \cdots m_k) \\ &= \left(\frac{S_1}{K} \cdots \frac{S_k}{K}, \frac{M_1}{K} \cdots \frac{M_k}{K}\right) \times 255 \end{aligned} \tag{4}$$

The expression of the deep convolutional pyramid feature is shown in Eq. (4), where C_i represents the number of occurrences of the visual word, and c_i represents the normalized feature representation.

$$C = (c_1, c_2 \cdots\cdots c_k) = \left(\frac{c_1}{K} \frac{c_1}{K} \cdots\cdots \frac{c_k}{K}\right) \times 255 \tag{5}$$

In the pyramid feature enhancement expression, the convolution layer feature can be expressed as x × y × D. The convolution layer feature map size is x × y, and D represents the depth of convolution pyramid feature map. The feature map with D dimension is then input to pooling layer to obtain the feature expression, as shown in Eq. (5), where sum_i represents the sum of all the values in the feature map, x ∗ y represents the number of values in the feature map.

$$\begin{aligned} P &= (p_1, p_2 \cdots\cdots p_m) \\ &= \left(\frac{sum_1}{x*y} \frac{sum_2}{x*y} \cdots\cdots \frac{sum_m}{x*y}\right) \times 255 \end{aligned} \tag{6}$$

Based on the above description of the feature, the LCPB can be denoted as $\{F_s, F_m, F_{pb}\}$, and the LCPP can be denoted as $\{F_S, F_M, F_{PB}\}$. The fusion result is a one-dimensional feature vector of two different dimensions. In the task of LCPB and LCPP classification, LCPB and LCPP with discriminative semantics are classified by the SVM classifier with a linear kernel. Finally, the scene label of each image can be predicted by the two different approaches.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP AND DATASETS

In order to test the performance of the MCPSF framework, the commonly used 21-class UC Merced dataset [11] and the 30-class AID dataset [8] were evaluated in the experiments. In the migration experiment, large area satellite images of Hanyang district in Wuhan were used to annotate the scene. In this section, the proposed MCPSF is compared to the other scene classification methods [28], [37], [44], [46], [47]. The sensitivity analysis of the experimental parameters is also provided. The datasets used in this experiment are described as follows.

The UC Merced (UCM) dataset was downloaded from the USGS National Map Urban Area Imagery collection, which contain 20 regions of the United States. As shown in Fig. 7, the images in UCM dataset were cropped into small areas with a size of 256 × 256, which was divided into 21 land use categories. Each category contained 100 images with a 1-ft spatial resolution. In the scene classification experiment, 80 images of each category were used for training.

FIGURE 7. Example images from the UC Merced dataset.

The Aerial Image Dataset (AID) is a large-scale remote sensing image dataset composed from Google Earth. AID contains 10000 images divided into 30 scene classes. Each class contains hundreds of images (ranging from 220 to 420) with the size of 600× 600 pixels in the RGB space. The spatial resolution ranges from about 8 to 0.5 m. The training set ratio for each category is set to 20% and 50%, respectively [8]. Fig. 8 shows some examples of the AID dataset.

In the experiments with uniform-grid based region sampling, the patch size and spacing were optimally set to 8 × 8 pixels and 4 × 4 pixels, respectively, for the spectral and structural features of the two datasets. The mid-level features based on BoVW model are obtained by clustering method,

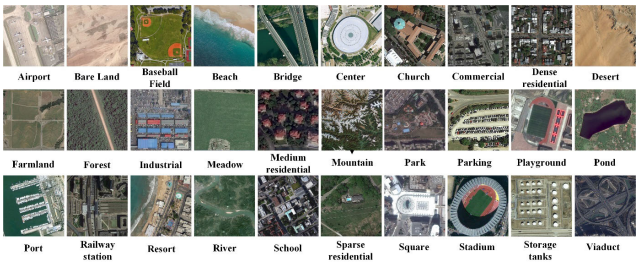

FIGURE 8. Example images from the AID dataset.

and the change of visual dictionary size K will affect the classification accuracy. In the experiments, the visual word number K was set from 100 to 2500 to test the sensitivity analysis. Too large or small dictionary size will have an impact on the predicted results, as well as the time and CPU cost. Hence, in MCPSF, the visual word number K was optimally set to 2000 for the MSD and SIFT features on the two datasets, respectively. For deep feature based CNN method, the images were both resized to 224 × 224 for the two dataset. The experiment in this paper was carried out on a personal device containing a NVIDIA GeForce GTX 950, Intel core i5-6300HQ CPU, RAM: 16GB. The deep feature based experiment environment was the windows-based GPU tensorflow framework. The experimental environment for mid-level feature based scene classification was undertaken using MATLAB 2018a.

To further validate the practical application for scene annotation, a large satellite image of Hanyang District of Wuhan in 2009, with a size of 6150 × 8250 is employed. This dataset was acquired from the IKONOS sensor with a spatial resolution of 1 m, and is named as Wuhan IKONOS dataset. 8 scene categories are defined for Wuhan IKONOS dataset, including commercial, industrial, dense residential, idle, medium residential, parking lot, vegetation and water. To achieve scene annotation, the size of small images used for large image segmentation is set to 150 × 150, and the patch spacing is 100. This leads to 50 overlapped pixels between adjacent images. In this way, 4000 scenes were selected to build the Wuhan IKONOS dataset, and the number of scenes in each category ranged from 100 to 800. The Wuhan IKONOS dataset uses 20% of the segmented scenes as the training data and the remaining 80% as the test data. For the BoVW based MSD and SIFT feature extraction, uniform patch with a size of 8 × 8 and patch spacing with a size of 4 × 4 is adopted. To cluster the SIFT and MSD feature has been extracted, a visual dictionary with the size of 1000 is built. All scenes obtained by uniform grid were resized to 224 × 224, and were input into the VGG-19 model to extract the multi-level convolution pyramid feature. The Wuhan IKONOS with 8 scene categories is shown in Fig. 9.

B. EXPERIMENT 1: THE UC MERCED IMAGE DATASET

The performance of single feature based BoVW method, BoVW-SIFT and BoVW-MSD, deep Conv and Fc features

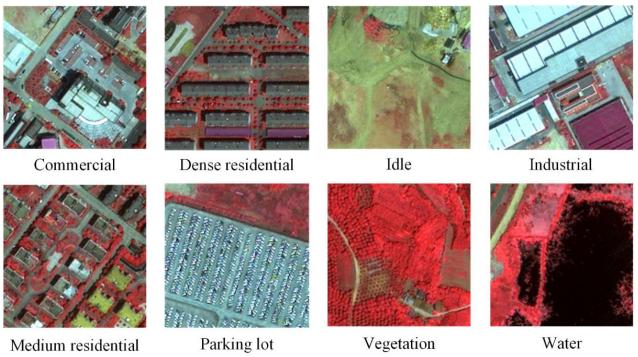

FIGURE 9. Example images from the Wuhan IKONOS dataset.

TABLE 2. Overall classification accuracy (%) comparison with the UCM dataset.

Method	Classification Accuracy(%)
BoVW-SIFT	80.62 ± 1.43
BoVW-MSD	83.03 ± 1.62
VGG-19-Conv4	92.57 ± 1.38
VGG-19-Fc	93.80 ± 1.07
BoVW-Conv pyramid	94.89 ± 0.84
Pooling-Conv pyramid	95.61 ± 0.68
SIFT+MSD+Conv4	93.57 ± 1.14
Yang and Newsam (2010)	81.19
Zhao et al. (2016)	91.67 ± 1.70
Nogueira et al. (2017)	97.10
Zheng et al. (2019)	96.90 ± 0.23
LCPB	96.66 ± 1.36
LCPP	97.54 ± 1.02

based methods, VGG-19-Conv4, VGG-19-Fc, BoVW-Conv-pyramid, Pooling-Conv pyramid, and SIFT+MSD+Conv4, as well as the proposed LCPB and LCPP are shown in Table 2. Here, BoVW-SIFT and BoVW-MSD refers to the methods mapping low-level features to mid-level using k-means clustering method. VGG-19-Conv4 refers that the features of the fourth convolutional layer are used for classified, and the VGG-19-Fc indicates that the fully connected layer features are used for classification directly. BoVW-Conv-pyramid and Pooling-Conv pyramid refers to multi-scale convolution pyramid features expressed by BoVW and Pooling models. SIFT+MSD+Conv4 refers to feature fusion with mid-level and Single-scale convolutional layer features. The performance of deep Conv and Fc features based methods are better than BoVW based methods, which proves that deep feature has stronger expression ability. The BoVW-Conv pyramid, Pooling-Conv pyramid are better than that of VGG-19-Conv4 and VGG-19-Fc, which confirms the effectiveness of combining BoVW and CNN. The classification accuracies of the proposed LCPB and LCPP are the best among all the different methods, which are 96.66% and 97.54%, respectively. This indicates that the multi-level convolution pyramid semantic expression can provide discriminative image representation for scene classification. The classification results of LCPP are slightly better than that of LCPB. This implies that encoded CNN feature using the mid-level feature based methods may lose crucial information. It also implies that LCPP which appropriately fused multi-level features, improves the feature expression ability. In addition, it can be seen that the MCPSF framework performs better than the other current methods, such as the mid-level based methods [11], [46] and deep learning based methods [21], [42].

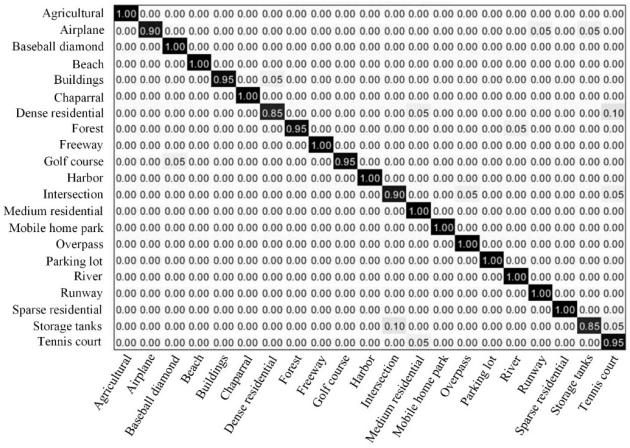

FIGURE 10. Confusion matrix of LCPB with UCM dataset.

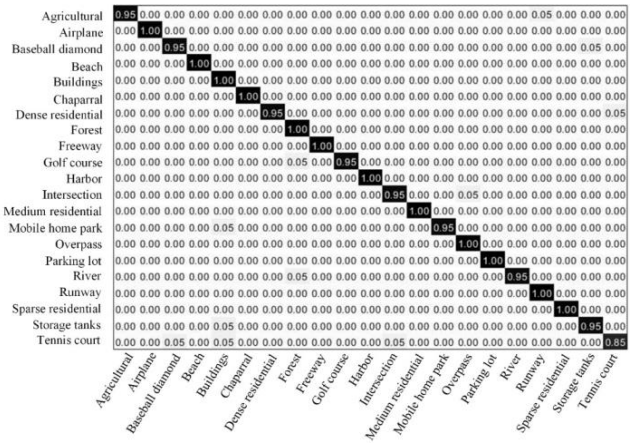

FIGURE 11. Confusion matrix of LCPP with UCM dataset.

The confusion matrices generated by LCPB and LCPP on the UCM dataset are shown in Figs. 10 and 11, respectively. The proposed LCPB and LCPP achieved the classification accuracy of 95% for at least 17 categories. Scene categories such as airplane, baseball diamond, beach, harbor, all with an accuracy of 100%, usually have representative objects. This indicates that the proposed MCPSF can acquire significant representation for complex scenes. Compared to the confusion matrix of LCPB, the scene categories in the confusion matrix of LCPP obtain a better performance. For example, the forest and building scenes are confused in LCPB, but are fully identified by LCPP. The tennis court and dense residential scenes are misclassified by both LCPB and LCPP, which may due to the high variability of these scenes.

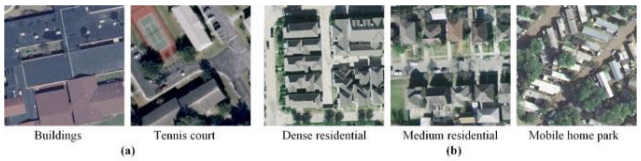

FIGURE 12. Similar target shapes of spectral features in categories within (a) and (b).

Some scenes categories are confused both in LCPB and LCPP for the UCM dataset. These categories tend to have similar target shapes or spectral features. As shown in Fig. 12 (a), the tennis court is often surrounded by buildings and roads, and is similar to the building in shape and spectral characteristics. Therefore, the tennis court and the building scene are easily confused in the classification process. As shown in Fig. 12 (b), the medium residential, dense residential and mobile home park classes are also confused due to the same object classes, i.e., roads, houses, and trees.

TABLE 3. Overall classification accuracy (%) comparison with the AID dataset.

Method	Classification Accuracy (%)	
	20%	50%
BoVW-SIFT	60.83 ± 0.32	65.74 ± 0.56
BoVW-MSD	61.27 ± 0.62	67.59 ± 0.70
VGG-19-Conv4	78.87 ± 0.34	83.46 ± 0.32
VGG-19-Fc	83.64 ± 0.41	87.18 ± 0.42
BoVW-Conv pyramid	85.38 ± 0.32	89.18 ± 0.56
Pooling-Conv pyramid	86.55 ± 0.48	90.87 ± 0.45
SIFT+MSD+Conv4	80.87 ± 0.45	86.50 ± 0.20
Xia et al. (2017)	86.59 ± 0.29	89.64 ± 0.36
Bian et al. (2017)	86.92 ±0.35	89.76 ±0.45
Anwer et al. (2018)	90.87 ± 0.11	92.96 ± 0.18
LCPB	87.68 ± 0.25	91.33 ± 0.36
LCPP	90.96 ± 0.33	93.12 ± 0.28

C. EXPERIMENT 2: THE AID IMAGE DATASET

The classification performance of single feature based BoVW method, BoVW-SIFT and BoVW-MSD, deep Conv and Fc features based methods, VGG-19-Conv4, VGG-19-Fc, BoVW-Conv-pyramid, Pooling-Conv pyramid, and SIFT+MSD+Conv4, the proposed LCPB and LCPP are shown in Table 3. The classification results for the proposed LCPB and LCPP, 91.33% and 93.12%, respectively, are better than the rest of the methods, i.e., BoVW-SIFT, VGG-19-Conv4, and SIFT+MSD+Conv4. This demonstrates that the MCPSF is an effective framework, and the fusion feature of multi-level is able to improve the classification performance. In addition, the classification results of methods including deep feature are better than those without deep feature, which indicates that the handcrafted features are slightly insufficient in the semantic expression of the scene. The results of BoVW-Conv-pyramid and Pooling-Conv pyramid are not only better than single Conv4 but also better than SIFT+MSD+Conv4. The results show that the pyramid character enhance the ability to recognize multi-scale objects and improve scene understanding. The proposed LCPP, which integrates the essential features of scene is better than that of LCPB, which integrates the frequency features. Previous studies have shown that traditional methods fail to achieve good results in AID dataset, while CNN-based methods can significantly improve classification accuracy, such as [8], [41], [48].

The confusion matrices generated by LCPB and LCPP on the AID are shown in Figs. 13 and 14, respectively. The proposed LCPB and LCPP achieved the classification accuracy of 90% for at least 18 categories. Scenes with classification accuracy up to 99% or completely correct include desert, farmland, mountain, viaduct, etc., which have obvious spectral or structural features or ground object targets with high identification. Compared to the confusion matrix of LCPB, the scene categories in the confusion matrix of LCPP obtain a better performance. For example, the resort and center scenes are confused in LCPB, but recognition accuracy has been greatly improved in LCPP. The school and park scenes are misclassified by both LCPB and LCPP. The park and the resort have a strong correlation because resorts usually include parks. The school scenes contain a lot of buildings, so it is easy to be confused with commercial and church scenes.

D. EXPERIMENT 2: THE AID IMAGE DATASET

In the scene annotation experiment, dataset is constructed from local scenes obtained from image segmentation, the training set ratio for each category is set to 20%. The test accuracies of MCPSF and other methods are shown in Table 4. The experimental results in Table 4 indicate that the proposed MCPSF framework can be well migrated to the task of land use classification in urban areas. The multi-level feature fusion method LCPB and LCPP obtain a higher accuracy than other single feature methods. Methods using deep CNN features perform better than those using mid-level features. These results are consistent with experiments on public datasets and prove the portability of this framework. For the Wuhan IKONOS dataset, the LCPB and LCPP methods achieved 92.28% and 93.78% accuracy, respectively. Compared with LCPB, LCPP designing appropriate multi-level feature fusion strategy improves scene classification performance. The result confusion matrix obtained by LCPP method on this dataset is shown in Fig. 15. It can be seen from the matrix that the identification accuracies of vegetation, water and industrial are more than 95%, which may indicate that the MCPSF framework captures the feature of these categories well. Commercial and parking lot have low recognition accuracies, and some scenes are incorrectly identified as industrial due to their similar structural and spectral characteristics.

The large satellite image of Wuhan IKONOS dataset and the results of labeling it are shown in Figs. 16 (a) and 16 (b),

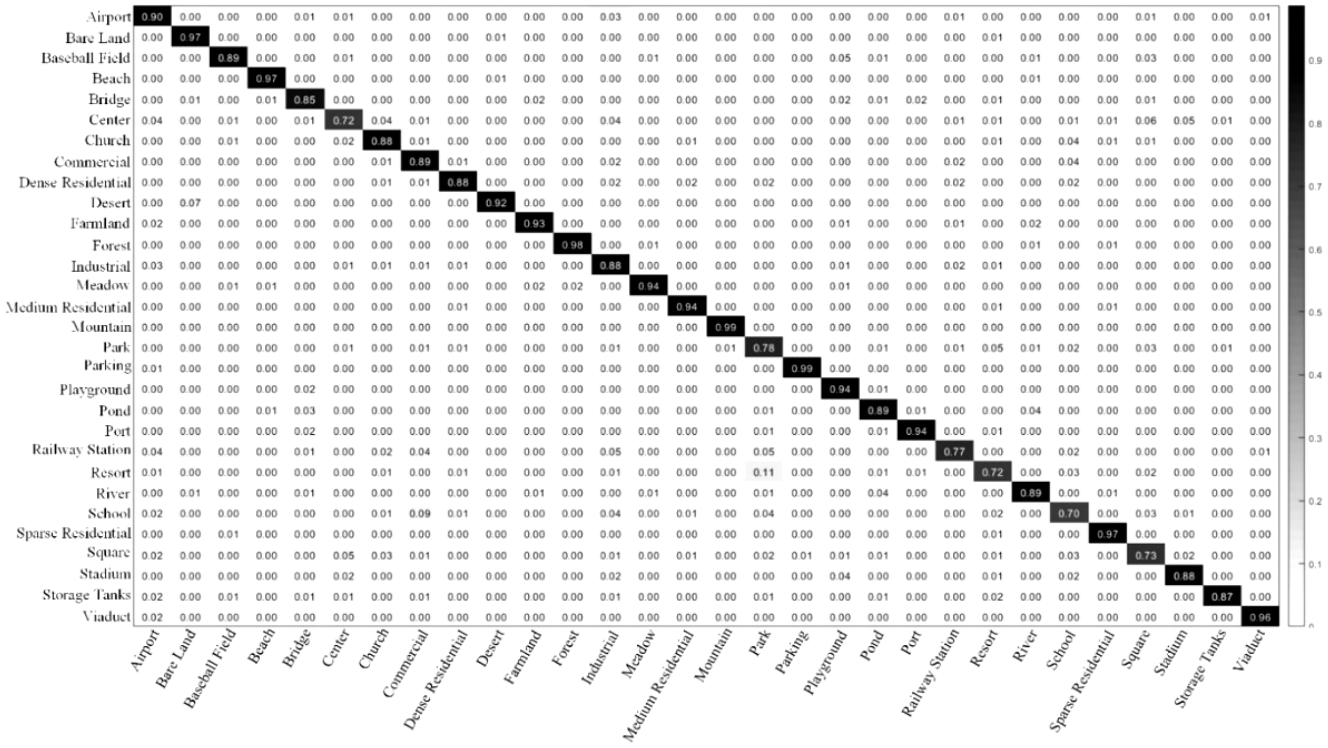

FIGURE 13. Confusion matrix of LCPB with AID.

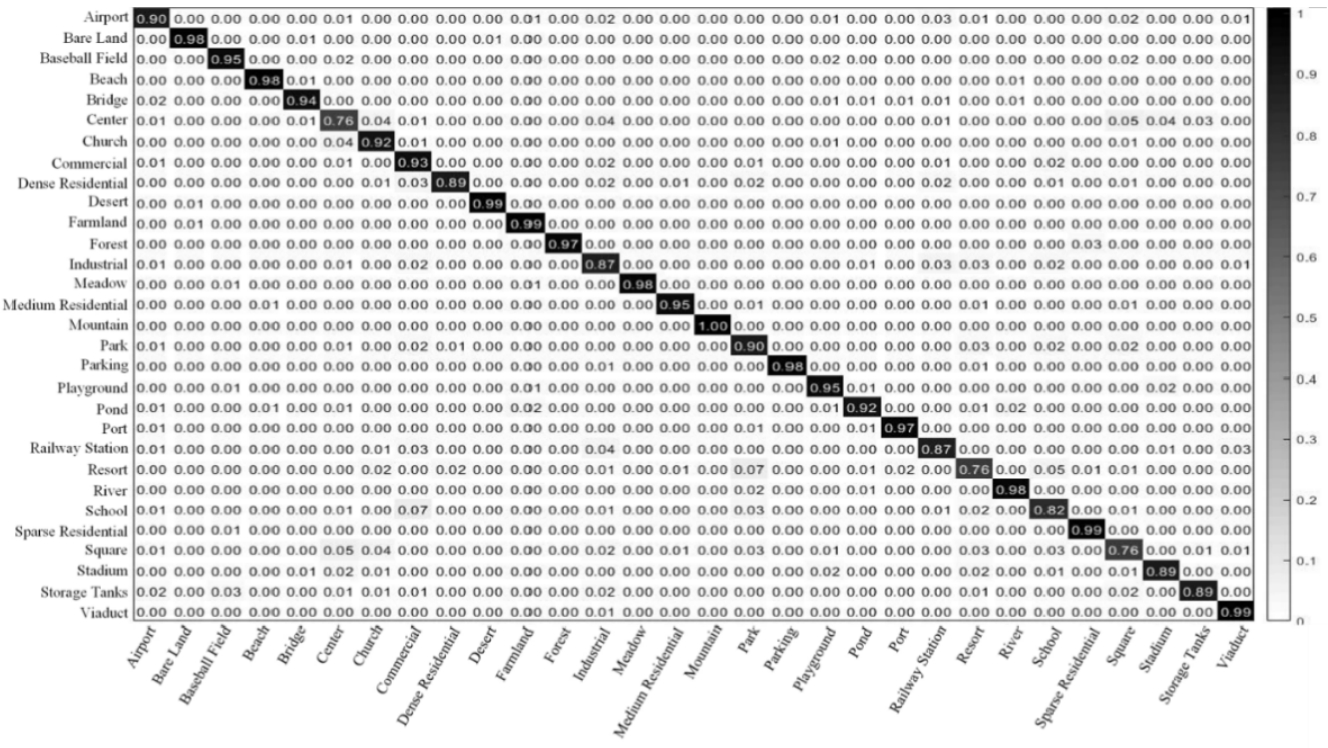

FIGURE 14. Confusion matrix of ADSSM with the Google dataset of SIRI-WHU.

respectively. It can be seen that using MCPSF framework to annotate local scenes is roughly correct, but few confused local patches also exist. Due to the similarity between medium and dense residential in the structure and spatial distribution of the buildings and their structure, there are some confusions between two scenes. Similarly, there are

TABLE 4. Overall classification accuracy (%) comparison with the Wuhan IKONOS dataset.

Method	Classification Accuracy (%)
BoVW-SIFT	86.26 ± 1.03
BoVW-MSD	88.07 ± 1.28
SSBFC	89.77 ± 3.75
VGG-19-Conv4	90.10 ± 0.56
LCPB	92.28 ± 0.32
LCPP	93.78 ± 0.28

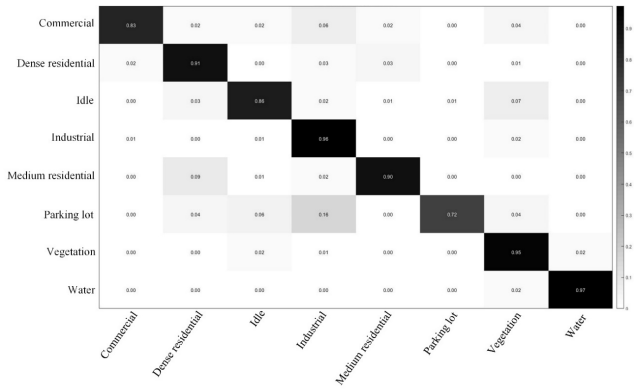

FIGURE 15. Confusion matrix of LCPB with the Wuhan IKONOS dataset.

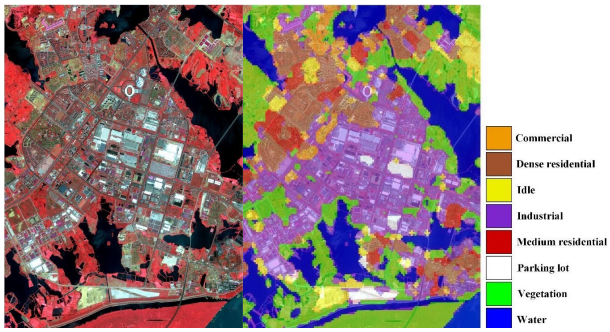

FIGURE 16. (a)Wuhan IKONOS image to be annotated. (b) Scene annotation results of Wuhan IKONOS image.

confusions between commercial and industrial areas due to the diversity of their objects and spatial distribution. In addition, the edge areas of different scenes are easily confused, and the mixed scenes generated by the uniform grid are difficult to be accurately identified. However, this paper has achieved satisfactory results by transferring the scene classification framework based on public datasets to large-scale urban image annotation experiments.

V. DISCUSSION

A. SENSITIVITY ANALYSIS

The low-level feature extraction method based on uniform grid sampling needs to control the patch size and spacing. Whether there are overlaps between grids will affect the integrity of feature extraction. The core of obtaining mid-level features by using BoVW model is K-means

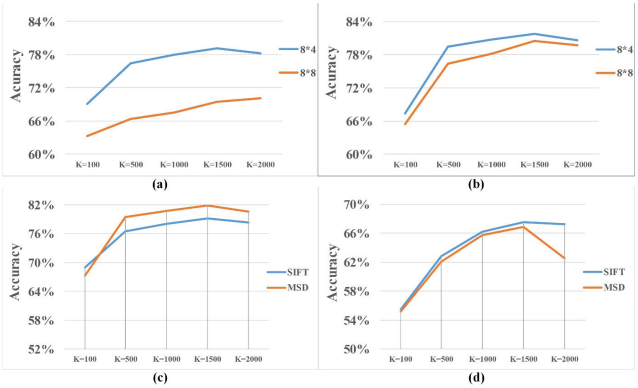

FIGURE 17. (a) Accuracy of SIFT with different patch spacing on UCM. (b) Accuracy of MSD with different patch spacing on UCM. (c) Accuracy of mid-level features on UCM with different dictionary size. (d) Accuracy of mid-level features on AID with different dictionary size.

clustering algorithm, which calculates the Euclidean distance between visual words and keeps iterating to find K clustering centers. Then a visual dictionary with K visual words can be acquired. According to experiments, the size of dictionary leads to slight fluctuation of classification accuracy. Too large visual dictionary scale will lead to a sharp increase for the time and space consumed, and too small scale will lead to insufficient understanding of the scene. In this paper, sensitivity experiments are utilized to explore the appropriate sampling spacing and visual dictionary size. The size of sampling patches with low-level feature extraction is 8×8. When the patch spacing is 4, 4 pixels will be overlapped between sampling patches; when the step size is 8, no overlapped area will be generated. As shown in Figs. 17 (a) and 17 (b), on the UCM data set, both SIFT and MSD have higher accuracy than step size 8 when step size is 4. In the experiments, the visual word number K was varied over the range of [100, 500, 1000, 1500, 2000] for the UC Merced dataset and the Aerial Image Dataset. From Figs. 17 (c) and 17 (d), it can be seen that SIFT and MSD have better expression ability when the scale of visual dictionary reaches 1500.

B. SENSITIVITY ANALYSIS

VGG-19 feature extractor consists of 5 levels with 16 convolutional layers, which can extract features of any layer in the network. The first two levels contain 4 convolution layers, and the size of the feature map is relatively large. In this paper, we focus on the analysis of the expression ability of 12 convolution layers at levels 3-5. By analyzing the expression ability of each convolution layer under different dictionary scales, the optimal dictionary size for expressing convolution layer features using BoVW model is obtained. Figs. 18 (a) and 18 (b) show the UCM classification results of the individual expressions of all convolutional layers at different dictionary scales and the average performance of each convolution level at different K values, respectively. The same analysis results on the AID dataset are shown in Figs. 18 (c) and 18 (d). According to Fig. 18, when K=100,

FIGURE 18. (a) Classification accuracy of all convolution layers in UCM dataset. (b) Average performance of each convolution level of UCM dataset. (c) Classification accuracy of all convolution layers in AID dataset. (d) Average performance of each convolution level of AID dataset.

each convolution layer performs poorly, because the scale of the dictionary is too small to fully describe the scene. When K=1500, the expression ability of each convolution layer begins to decline, because the scale of visual dictionary is too large, resulting in excessive expression of scenes and poor classification effect. In general, in UCM and AID datasets, the fourth convolution level performs best on average. For visual dictionary size, UCM dataset performs better on average when K=500, while AID data set performs better when K=1000.

VI. CONCLUSION

In this paper, the multi-level convolutional pyramid semantic fusion (MCPSF) framework has been proposed for high spatial resolution (HSR) imagery scene classification. In MCPSF, considering the special solar illumination condition of HSR images, the first and second-order statistics mean and standard (MSD) of pixels are used as the local spectral characteristics. Considering the special land coverage of HSR images, scale-invariant feature transform (SIFT) are used to describe the structural characteristics. Both handcrafted features are mapped to the mid-layer using the BoVW model. Experiments with three HSR image datasets indicate that the MSD works better than the SIFT features, and the combination of MSD and SIFT can perform better. Instead of the mid-level feature, MCPSF represents the images with high-level features extracted from the pre-trained CNN model. Features of multi-scale convolutional layer and fully connected layer are extracted from the CNN model. Experiments on three datasets show that compared to feature extracted from single convolutional layer, the feature of fully connected layer can describe the scene better, but features of multi-scale Convolutional layer perform best after fusion. Finally, the proposed LCPP and LCPB methods efficiently fuse mid-level features and multi-scale deep features, which acquires the best performance for HSR image scene classification.

However, there is scope for further improving the classification performance of the proposed MCPSF framework. The deep CNN model dedicated to HSR image training is limited by the number of samples, and remote sensing images tend to show multi-scale characteristics. In our future research, the data enhancement method for remote sensing images will be studied to meet the needs of training specific model. The CNN network will be appropriately improved, and the model will be trained through samples of different scales.

REFERENCES

[1] Q. Zhu, Y. Zhong, Y. Liu, L. Zhang, and D. Li, "A deep-local-global feature fusion framework for high spatial resolution imagery scene classification," *Remote Sens.*, vol. 10, no. 4, p. 568, Apr. 2018.

[2] G. J. Hay, T. Blaschke, D. J. Marceau, and A. Bouchard, "A comparison of three image-object methods for the multiscale analysis of landscape structure," *ISPRS J. Photogramm. Remote Sens.*, vol. 57, nos. 5–6, pp. 327–345, Apr. 2003.

[3] S. Aksoy, K. Koperski, C. Tusk, G. Marchisio, and J. C. Tilton, "Learning Bayesian classifiers for scene classification with a visual grammar," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 581–589, Mar. 2005.

[4] T. Blaschke, G. J. Hay, M. Kelly, S. Lang, P. Hofmann, E. Addink, R. Q. Feitosa, F. van der Meer, H. van der Werff, and F. van der Coillie, "Geographic object-based image analysis—Towards a new paradigm," *ISPRS J. Photogramm. Remote Sens.*, vol. 87, pp. 180–191, Jan. 2014.

[5] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[6] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 193–204, Mar. 2011.

[7] 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.

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

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

[10] E. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 26–38, Jan. 2003.

[11] 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.*, 2010, pp. 270–279.

[12] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. ECCV*, vol. 1, nos. 1–22, 2004, pp. 1–2.

[13] 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.

[14] L. Weizman and J. Goldberger, "Urban-area segmentation using visual words," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 3, pp. 388–392, Jul. 2009.

[15] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, nos. 1–2, pp. 177–196, Jan. 2001.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 5, pp. 993–1022, Jan. 2003.

[17] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.

[18] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[19] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.
[20] 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, Nov. 2015.
[21] 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.
[22] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[23] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
[24] X. Liu, Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng, "Siamese convolutional neural networks for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1200–1204, Aug. 2019.
[25] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," Aug. 2015, *arXiv:1508.00092*. [Online]. Available: http://arxiv.org/abs/1508.00092
[26] C. Zhang, X. Pan, H. Li, A. Gardiner, I. Sargent, J. Hare, and P. M. Atkinson, "A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 133–144, Jun. 2018.
[27] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "MARTA GANs: Unsupervised representation learning for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2092–2096, Nov. 2017.
[28] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Dec. 2018.
[29] C. Chen, B. Zhang, H. Su, W. Li, and L. Wang, "Land-use scene classification using multi-scale completed local binary patterns," *Signal, Image Video Process.*, vol. 10, no. 4, pp. 745–752, Apr. 2016.
[30] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[31] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.
[32] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[33] 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.
[34] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. Int. Workshop Workshop Multimedia Inf. Retr.*, 2007, pp. 197–206.
[35] Y. Zhong, F. Fei, and L. Zhang, "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery," *J. Appl. Remote Sens.*, vol. 10, no. 2, Apr. 2016, Art. no. 025006.
[36] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.
[37] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 197–209, Nov. 2018.
[38] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[39] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.
[40] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.
[41] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 74–85, Apr. 2018.
[42] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.
[43] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 524–531.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Systems.*, 2012, pp. 1097–1105.
[46] B. Zhao, Y. Zhong, and L. Zhang, "A spectral–structural bag-of-features scene classifier for very high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 73–85, Jun. 2016.
[47] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.
[48] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2889–2901, Jun. 2017.

XIONGLI SUN received the B.S. degree from the Wuhan University of Technology, Wuhan, China, in 2013. She is currently pursuing the Ph.D. degree in photogrammetry and remote sensing with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, China. Her major research interests include scene analysis for high-spatial resolution remote sensing imagery and object tracking.

QIQI ZHU (Member, IEEE) received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2013, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, China, in 2018. She has been an Associate Professor with the School of Geography and Information Engineering, China University of Geosciences, since 2018. Her research interests include high-resolution remote sensing image understanding, geoscience interpretation for multi-source remote sensing data, and applications. She has published more than 20 research articles, including peer-reviewed articles in international journals, such as *Remote Sensing of Environment*, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. She was also a Referee of more than 20 international journals.

QIANQING QIN received the Ph.D. degree in probability and statistics from Nankai University, Tianjin, China, in 1989.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. His research interests include photogrammetry and remote sensing, researching field involves in remote sensing image processing, and wavelet and its application and computer vision.