AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification

Gui-Song Xia, *Senior Member, IEEE*, Jingwen Hu, Fan Hu, *Student Member, IEEE*, Baoguang Shi, Xiang Bai, *Senior Member, IEEE*, Yanfei Zhong, *Senior Member, IEEE*, Liangpei Zhang, *Senior Member, IEEE*, and Xiaoqiang Lu, *Senior Member, IEEE*

***Abstract*— Aerial scene classification, which aims to automatically label an aerial image with a specific semantic category, is a fundamental problem for understanding high-resolution remote sensing imagery. In recent years, it has become an active task in the remote sensing area, and numerous algorithms have been proposed for this task, including many machine learning and data-driven approaches. However, the existing data sets for aerial scene classification, such as UC-Merced data set and WHU-RS19, contain relatively small sizes, and the results on them are already saturated. This largely limits the development of scene classification algorithms. This paper describes the Aerial Image data set (AID): a large-scale data set for aerial scene classification. The goal of AID is to advance the state of the arts in scene classification of remote sensing images. For creating AID, we collect and annotate more than 10 000 aerial scene images. In addition, a comprehensive review of the existing aerial scene classification techniques as well as recent widely used deep learning methods is given. Finally, we provide a performance analysis of typical aerial scene classification and deep learning approaches on AID, which can be served as the baseline results on this benchmark.**

***Index Terms*— Aerial images, benchmark, scene classification.**

I. Introduction

NOWADAYS, aerial images enable us to measure the Earth's surface with detailed structures and are a kind of data source of great significance for earth observation [1]–[6]. Due to the drastically increasing number of aerial images and the highly complex geometrical structures and spatial patterns, to effectively understand the semantic content of them is particularly important, driven by many real-world applications in the remote sensing community. In this paper, we focus on aerial scene classification, a key problem in aerial image understanding, which aims to automatically assign a semantic label to each aerial image in order to know which category it belongs to.

The problem of aerial scene classification has received growing attention in recent years [2]–[4], [7]–[42]. In the literature, primary studies have devoted to classifying aerial images at pixel level, by assigning each pixel in an aerial image with a thematic class [43], [44]. However, with the increment of the spatial resolutions, it turns to be infeasible to interpret aerial images at pixel level [43], [44], mainly due to the fact that single pixels quickly lose their thematic meanings and discriminative efficiency to separate different types of land covers. Specifically, in 2001, Blaschke and Strobl [45] raised the question "*What is wrong with pixels?*" and argued that it is more efficient to analyze aerial images at object level, where the "objects" refer to local regions of pixels sharing spectral or texture homogeneities, e.g., superpixels [46]. This kind of approaches then has dominated the analysis of high-resolution remote sensing images for decades [47]–[53]. It is worth noticing that both pixel- and object-level classification methods attempt to model an aerial scene in a bottom–up manner by aggregating extracted spectral, texture, and geometrical features for training a strong classifier.

However, due to the growing of image spatial resolutions, aerial scenes may often consist of different and distinct thematic classes [20] and it is of great interest to reveal the context of these thematic classes, i.e., semantic information, of aerial scenes. Aerial scene classification aims to classify an aerial image into different semantic categories by directly modeling the scenes by exploiting the variations in the spatial arrangements and structural patterns. In contrast with pixel-/object-oriented classification, scene classification provides a relatively high-level interpretation of aerial images. More precisely, the item "scene" hereby usually refers to a local area in large-scale aerial images that contain clear semantic information on the surface [2]–[4], [7]–[42], [54]–[62].

Though much exciting progress on aerial scene classification has been extensively reported in recent years, e.g., [2]–[4], [19]–[42], there are two major issues that seriously limit the development of aerial scene classification.

1) *Lacking a Comprehensive Review of Existing Methods:* Although many methods have been presented to advance the aerial scene classification, most of them were

Manuscript received August 14, 2016; revised December 24, 2016; accepted February 28, 2017. Date of publication April 24, 2017; date of current version June 22, 2017. This work was supported by the National Natural Science Foundation of China under Contract 41501462 and Contract 91338113. *(Corresponding author: Gui-Song Xia.)*

G.-S. Xia, Y. Zhong, and L. Zhang are with the State Key Laboratory of Information Engineering, Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: guisong.xia@whu.edu.cn; zlp62@whu.edu.cn).

J. Hu and F. Hu are with the State Key Laboratory of Information Engineering, Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China, and also with the Signal Processing Laboratory, School of Electronics Information, Wuhan University, Wuhan 430072, China (e-mail: hujingwen@whu.edu.cn).

B. Shi and X. Bai are with the School of Electronics Information, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: xbai@hust.edu.cn).

X. Lu is with the State Key Laboratory of Transient Optics and Photonics, Center for OPTical IMagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China.

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

Digital Object Identifier 10.1109/TGRS.2017.2685945

evaluated on different data sets under different experimental settings. This somewhat makes the progress confused and may mislead the development of the problem. Moreover, the codes of these algorithms have not been released, which brings difficulties to reproduce the works for fair comparisons. Therefore, the state of the art of aerial scene classification is not absolutely clear.

2) *Lacking Proper Benchmark Data Sets for Performance Evaluation:* In order to develop robust methods for aerial scene classification, it is highly expected that the data sets for evaluation demonstrate all the challenging aspects of the problem, for instance, the high diversities in the geometrical and spatial patterns of aerial scenes. Currently, the evaluations of aerial scene classification algorithms are typically done on data sets containing up to 2000 images at best, e.g., the UC-Merced data set [58] and the WHU-RS19 data set [57]. Such a limited number of images are critically insufficient to approximate the real applications, where the images are with high intraclass diversity and low interclass variation. Recently, the saturated results on these data sets demonstrated that more challenging data sets are badly required.

Due to the above-mentioned issues, in this paper, we present a comprehensive review of up-to-date algorithms as well as a new large-scale benchmark data set of aerial images (AID), in order to fully advance the task of aerial scene classification. AID provides the research community a benchmark resource for the development of the state-of-the-art algorithms in aerial scene classification or other related tasks, such as aerial image search, aerial image annotation, and aerial image segmentation. In addition, our experiments on AID demonstrate that it is quite helpful to reflect the shortcomings of existing methods. In summary, the major contributions of this paper are as follows.

1) We provide a comprehensive review of aerial scene classification, by giving a clear summary of the development of scene classification approaches.
2) We construct a new large-scale data set, i.e., AID, for aerial scene classification. The data set is, to the best of our knowledge, of the largest size and the images are with high intraclass diversity and low interclass dissimilarity, which can provide the research community a better data resource to evaluate and advance the state-of-the-art algorithms in an aerial image analysis.
3) We evaluate a set of representative aerial scene classification approaches with various experimental protocols on the new data set. These can serve as the baseline results for future works.
4) The source codes of our implementation for all the baseline algorithms are released, and will be general tools for other researchers.

The rest of this paper is organized as follows. The details of AID data set are first described in Section II. We then provide a comprehensive review of related methods in Section III. Subsequently, we provide a description of baseline algorithms for benchmark evaluation in Section IV. In Section V, the evaluation and the comparison of baseline algorithms on AID under different experimental settings are given. Finally, some conclusion remarks are drawn in Section VI.

The AID and the codes for reproducing all the results in this paper are downloadable at the project Web page www.lmars.whu.edu.cn/xia/AID-project.html.

II. Aerial Image Data Sets for Aerial Scene Classification

This section first reviews several data sets commonly used for aerial scene classification and then described the proposed AID.[1]

A. Existing Data Sets for Aerial Scene Classification

1) UC-Merced Data Set [58]: It consists of 21 classes of land-use images selected from aerial orthoimagery with the pixel resolution of 1 ft. The original images were downloaded from the United States Geological Survey National Map of the following U.S. regions: *Birmingham, Boston, Buffalo, Columbus, Dallas, Harrisburg, Houston, Jacksonville, Las Vegas, Los Angeles, Miami, Napa, New York, Reno, San Diego, Santa Barbara, Seattle, Tampa, Tucson, and Ventura*. They are then cropped into small regions of 256 × 256 pixels. There are totally 2100 images manually selected and uniformly labeled into 21 classes: *agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts*.

It is worth noticing that UCM data set contains a variety of spatial land-use patterns, which make the data set more challenging. Moreover, some highly overlapped classes, e.g., dense residential, medium residential, and sparse residential that mainly differ in the density of structures, make the data set difficult for classification. This data set is widely used for the task of aerial image classification [2]–[4], [7], [8], [10], [13], [15], [16], [18]–[20], [22]–[27], [30], [32]–[39], [58].

2) WHU-RS Data Set: This data set is collected from Google Earth imagery.[2] The images are with a fixed size of 600 × 600 pixels with various pixel resolutions up to half a meter. This data set has been updated to the third versions until now. In its original version [57], there are 12 classes of aerial scenes, including *airport, bridge, river, forest, meadow, pond, parking, port, viaduct, residential area, industrial area*, and *commercial area*. For each class, there were 50 samples. Later, Sheng *et al.* [9] expanded the data set to 19 classes with 7 new ones, i.e., *beach, desert, farmland, football field, mountain, park*, and *railway station*. Thus, the data set is composed of a total number of 950 aerial images, which is widely used as the *WHU-RS19 data set* [4], [9], [13], [30]–[32], [36], [59]. However, the size of this data set is relatively small compared with UC-Merced data set [58]. Thus, we reorganized and expanded WHU-RS19 to form its third version [31], by adding a new aerial scene type "bare land" and increasing the number of samples in each class. In the newest version of the

[1] The AID is downloadable at www.lmars.whu.edu.cn/xia/AID-project.html.
[2] https://www.google.com/earth/

WHU-RS data set, it thus has 5000 aerial images with each class containing more than 200 sample images. It is worth noticing that the sample images of the same class in WHU-RS data set are collected from different regions all around the world and the aerial scenes might appear at different scales, orientations, and with different lighting conditions.

3) RSSCN7 Data Set [41]: This data set is also collected from Google Earth,[3] which contains 2800 aerial scene images labeled into 7 typical scene categories, i.e., the grassland, forest, farmland, parking lot, residential region, industrial region, and river and lake. There are 400 images in each scene type, and each image has a size of 400×400 pixels. It is worth noticing that the sample images in each class are sampled on 4 different scales with 100 images per scale with different imaging angles, which is the main challenge of the data set.

4) Other Small Data Sets: Besides the three public data sets mentioned earlier, there are also several nonpublic data sets, e.g., the *IKONOS Satellite Image data set* [54], the In-House data set [7], [19], [61], the SPOT Image data set [12], and the ORNL data set [20]. Note that the numbers of scene types in all these data sets are less than 10, which thus results in small intraclass diversity. Moreover, the commonly used midlevel scene classification methods get saturated and have reported overall accuracies nearly 100% on these data sets. Therefore, these less challenging data sets will severely restrict the development of aerial scene classification algorithms.

B. AID: A New Data Set for Aerial Scene Classification

To advance the state of the arts in scene classification of remote sensing images, we construct AID, a new large-scale aerial image data set, by collecting sample images from Google Earth imagery. Note that although the Google Earth images are postprocessed using RGB renderings from the original optical aerial images, Hu *et al.* [1] have proven that there is no significant difference between the Google Earth images with the real optical aerial images even in the pixel-level land use/cover mapping. Thus, the Google Earth images can also be used as aerial images for evaluating scene classification algorithms.

The new data set is made up of the following 30 aerial scene types: *airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct*. All the images are labeled by the specialists in the field of remote sensing image interpretation, and some samples of each class are shown in Fig. 1. The numbers of sample images vary a lot with different aerial scene types (see Table I), from 220 up to 420. In all, the AID data set has a number of 10 000 images within 30 classes.

The images in AID are actually multisource, as Google Earth images are from different remote imaging sensors. This brings more challenges for scene classification than the single source images, such as UC-Merced data set [58]. Moreover, all the sample images per each class in AID are carefully chosen from different countries and regions around the world, mainly in *China, the United States, England, France, Italy, Japan, and Germany*, and they are extracted at different time and seasons under different imaging conditions, which increases the intraclass diversities of the data.

Note that another main difference between AID and UC-Merced data sets is that AID has multiresolutions: the pixel resolution changes from about 8 m to about half a meter, and thus, the size of each aerial image is fixed to be 600×600 pixels to cover a scene with various resolutions.

C. Why Is AID Proper for Aerial Image Classification?

In contrast with the existing remote sensing image data sets, e.g., UC-Merced data set and WHU-RS19 data set, AID has the following properties.

1) *Higher Intraclass Variations:* In aerial images, due to the high spatial resolutions, the geometrical structures of scenes become more clear and bring more challenges to image classification. First, thanks to the high complexity of the earth surface, objects in the same type of scene may appear at different sizes and orientations. Second, the different imaging conditions, e.g., the flying altitude and direction and the solar elevation angles, may also vary a lot the appearance of the scene. Thus, in order to develop robust aerial image classification algorithms with stronger generalized capability, it is accepted that the data set contains high intraclass diversity. The increasing numbers of sample images per each class in AID allow us to collect images from different regions all over the world accompanied with different scales, orientations, and imaging conditions, which can increase the intraclass diversities of the data set (see Fig. 2). In Fig. 2(a), we illustrate two examples of the same scene with different scales. In Fig. 2(b), we display the examples of the same type of scene with different building styles, as the sample images are collected in different regions and countries and the appearances of the same scene vary a lot due to the cultural differences. In Fig. 2(c), the shadow direction of the buildings varies from west to north at different imaging time, and a mountain varies from green to white along with the seasonal variation.
2) *Smaller Interclass Dissimilarity:* In real cases of aerial image classification, the dissimilarities between different scene classes are often small. The construction of AID well considered this point, by adding more scene categories. As shown in Fig. 3, AID contains scenes sharing the similar objects, e.g., both *stadium* and *playground* may contain sports field [see Fig. 3(a)], but the main difference lies in whether there are stands around. Both *bare land* and *desert* are fine-grained textures and share similar colors [see Fig. 3(b)], but *bare land* usually has more artificial traces. Some scene classes have similar structural distributions, such as *resort* and *park* [see Fig. 3(c)], which may contain a lake and some buildings; however, a park is generally equipped with

[3]https://www.google.com/earth/

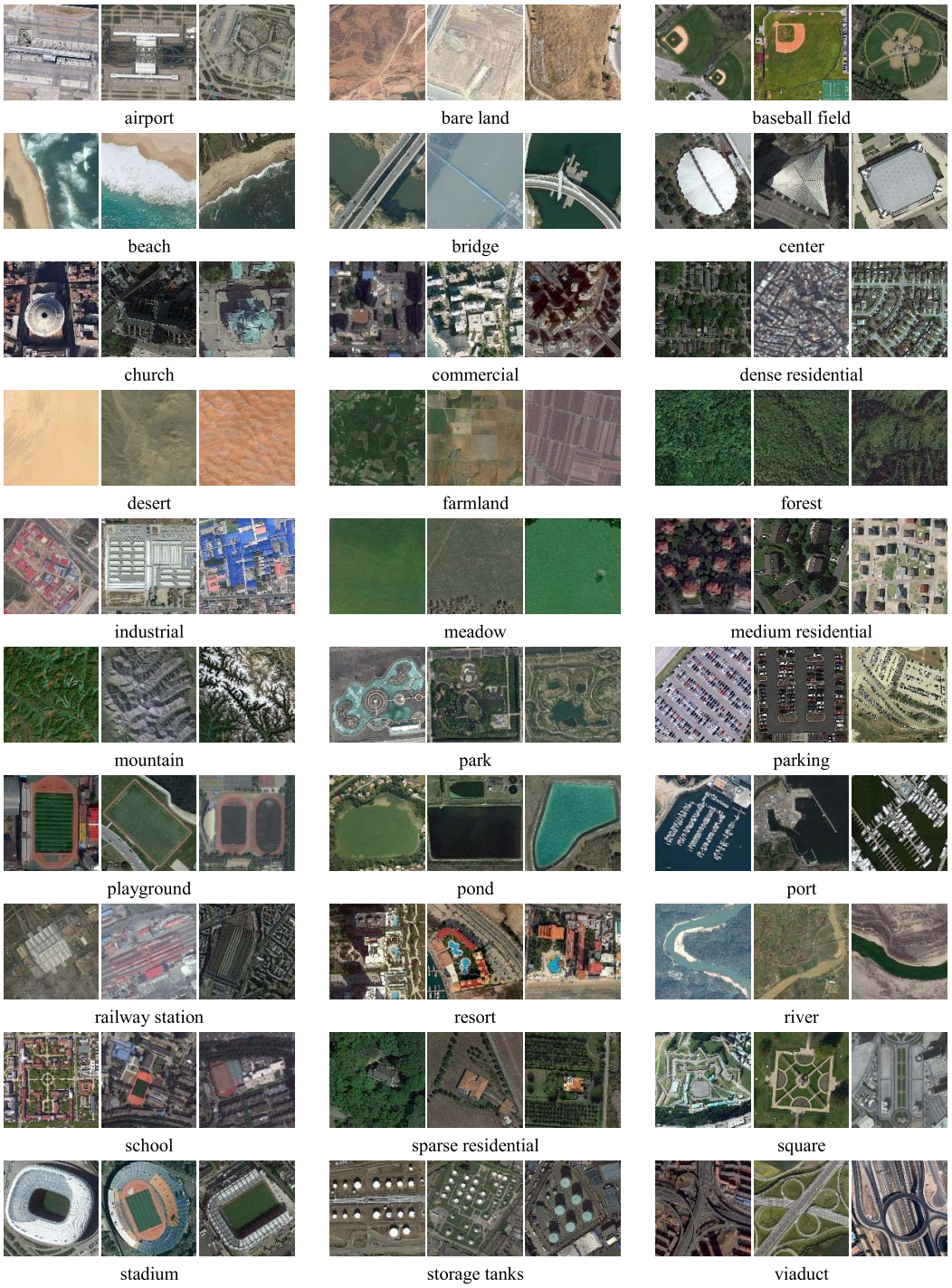


Fig. 1. Samples of AID: three examples of each semantic scene class are shown. There are 10 000 images within 30 classes.

amusement and leisure facilities, while a resort is usually composed of villas for vacations. The AID has taken into account many these kinds of scene classes with small interclass dissimilarity and makes it closer to real aerial image classification tasks.

3) *Relative Large-Scale Data Set:* For validating the classification algorithm, large-scale labeled data are often expected. However, the manual annotation of aerial images requires expertise and is extremely time-consuming. AID has a total number of 10 000 images which is, to the best of our knowledge, the largest annotated aerial image data sets. It can cover a much wider range of aerial images and better approximate the real aerial image classification problem than the existing data set. In contrast with our AID, both the UC-Merced

TABLE I
DIFFERENT SEMANTIC SCENE CLASSES AND THE NUMBER OF IMAGES IN EACH CLASS OF THE NEW DATA SET

Types	#images	Types	#images	Types	#images
airport	360	farmland	370	port	380
bare land	310	forest	250	railway station	260
baseball field	220	industrial	390	resort	290
beach	400	meadow	280	river	410
bridge	360	medium residential	290	school	300
center	260	mountain	340	sparse residential	300
church	240	park	350	square	330
commercial	350	parking	390	stadium	290
dense residential	410	playground	370	storage tanks	360
desert	300	pond	420	viaduct	420

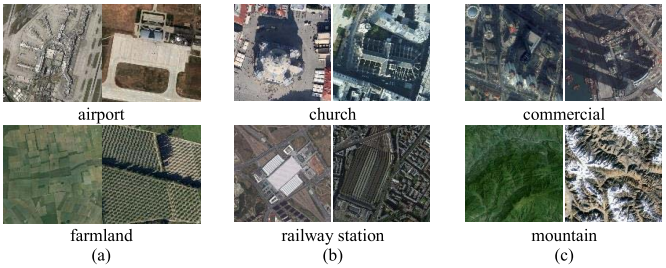


Fig. 2. Large intraclass diversity. (a) Multiscale images of the same scene. (b) Different building styles of the same scene. (c) Different imaging conditions of the same scene.

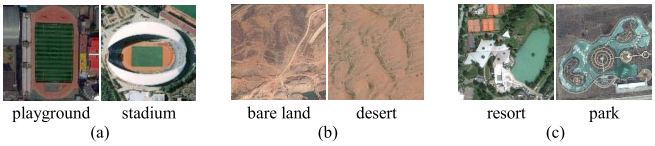


Fig. 3. Small interclass distance. (a) Similar objects between different scenes. (b) Similar textures between different scenes. (c) Similar structural distributions between different scenes.

data set [58] and WHU-RS19 data set [9] contain 100 images per class, and only 20 images in each class were usually used for testing the algorithms [4], [8], [20], [24]–[26], [30], [32]–[34], [36]–[39], [58]. In such case, the classification accuracy that will be seriously affected by even one image is predicted correctly or not and results in big standard deviation, especially when analyzing the classification results on each class. Therefore, our AID with relatively large-scale data can provide a better benchmark to evaluate image classification methods.

III. REVIEW ON AERIAL SCENE CLASSIFICATION

This section comprehensively reviews the existing scene classification methods for aerial images. Distinguished from pixel-/object-level image classifications which interpret aerial images with a bottom–up manner, scene classification is apt to directly model an aerial scene by developing a holistic representation of the aerial image [2]–[4], [7]–[42], [54]–[61]. One should observe that, actually, the underlying assumption of scene classification is that the same type of scene should share certain statistically holistic visual characteristics. This point has been verified on natural scenes [63] and demonstrates its efficiency on classifying aerial scenes. Thus, most of the works on aerial scene classification focus on computing such holistic and statistical visual attributes for classification. In this sense, scene classification methods can be divided into three main categories: methods using low-level visual features, methods relying on midlevel visual representations, and the methods based on high-level vision information. In what follows, we review each category of the methods in detail.

A. Methods Using Low-Level Visual Features

With this kind of methods, it is supposed that the aerial scenes can be distinguished by low-level visual features, e.g., spectral, texture, and structure. Consequently, an aerial scene image is usually described by a feature vector extracted from such low-level visual attributes, either locally [54] or globally [55], [57]. On the one hand, in order to describe the complex structures, local structure descriptors [64], [65] have been widely used for modeling the local variations of structures in aerial images. The classification feature vector is usually formed by concatenating or pooling the local descriptors of the subregions of an image. On the other hand, for depicting the spatial arrangements of aerial scenes, statistical and global distributions of certain spatial cues, such as color [66] and texture information [7], [67]–[69], have also been well investigated. For instance, Yang and Newsam [54] compared scale invariant feature transform (SIFT) [64] and Gabor texture features for classifying IKONOS satellite images by using maximum *a posteriori* classifier and found that SIFT performs better. dos Santos *et al.* [55] evaluated various global color descriptors and texture descriptors, e.g., color histogram (CH) [66] and local binary pattern (LBP) [70], for scene classification.

Although single type of features works well for classification [54], [55], [61], the combinations of complementary features can often improve the results [12], [19], [27]. In particular, Luo *et al.* [12] extracted six different kinds of feature descriptors, i.e., simple radiometric features, Gaussian wavelet features [71], gray-level co-occurrence matrix (GLCM), Gabor filters, shape features [72], and SIFT, and combined them to form a multiple-feature representation for indexing remote sensing images with different spatial resolutions, which reported that multiple features can describe aerial scenes better. Avramović and Risojević [19] integrated gist [63] and SIFT descriptors for aerial scene classification.

Some class-specific feature selection methods were also developed to select a good subset of low-level visual features for aerial image classification [27].

In order to encode the global spatial arrangements and the geometrical diversities of aerial scenes, Xia *et al.* [57] proposed an invariant and robust shape-based scene descriptor [68] to describe the structure distributions of aerial images, while Risojević and Babić [7], [10], [15] focused on the texture information of scenes and they successively proposed a local structural texture descriptor, an orientation difference descriptor, and an enhanced Gabor texture descriptor based on the Gabor filters [67] to further improve the performance. In [40], a multiscale completed LBPs was proposed for land-use scene classification and achieved the state-of-the-art performance among low-level methods.

It is worth noticing that scene classification methods with low-level visual features perform well on some aerial scenes with uniform structures and spatial arrangements, but it is difficult for them to depict the high-diversity and the nonhomogeneous spatial distributions in aerial scenes [16].

B. Methods Relying on Midlevel Visual Representations

In contrast with methods relying on low-level visual attributes, the midlevel aerial scene analysis approaches attempt to develop a holistic scene representation through representing the high-order statistical patterns formed by the extracted local visual attributes. A general pipeline is to first extract local image attributes, e.g., SIFT, LBP, and color histograms of local image patches, and then encode these local cues for building a holistic midlevel representation for aerial scenes.

One of the most popular midlevel approaches is the *bag-of-visual-words* (BoVW) model [58]. More precisely, this method [58] first described local image patches by SIFT [64] descriptors and then learned a vocabulary of visual words (also known as dictionary or codebook) for instance by k-means clustering. Subsequently, the local descriptors were encoded against the vocabulary by hard assignment, i.e., vector quantization, and a global feature vector of the image could be obtained by the histogram of visual words, which is actually counting the occurrence frequencies of each visual word in the image. Thanks to its simplicity and efficiency, the BoVW model and its variants have been widely adopted for computing midlevel representation for aerial scenes [8], [13], [14], [22]–[24], [26], [29], [31], [58], [59].

In order to improve the discriminative power of the BoVW model, multiple complemented low-level visual features were combined under the framework. For instance, in [59], various local descriptors, including SIFT, GIST, CH, and LBP, were evaluated with the standard BoVW model for aerial scene classification. The experiments of the concatenation of BoVW representations from different local descriptors proved that combining complemented features can significantly improve the classification accuracy. Similarly, in [14], a highly discriminative texture descriptor, i.e., combined scattering feature [73], was incorporated with SIFT and CH to extract structure and spectral information under a multifeature extraction scheme. Unlike the simple concatenation in [59], in this paper, a hierarchical classification method incorporating *extreme value theory*-based normalization [74] was used to calibrate multiple features. In [9], multiple features, such as structure features, spectral features, and texture features, were extracted and encoded in a sparse coding scheme [75] besides BoVW. The sparse coding scheme was developed based on BoVW but adding a sparsity constraint to the feature distributions to reduce the complexity of *support vector machine* meanwhile maintaining good performance. In [22], various feature coding methods developed from the BoVW model were evaluated for scene classification using multifeatures. By applying principal component analysis for dimension reduction before concatenating multifeatures, the *improved Fisher kernel* (IFK) [76] and *vectors of locally aggregated tensors* [22] methods reported to achieve the state-of-the-art performances.

Note that in BoVW models, they count the frequencies of the visual words in an image, with regardless of the spatial distribution of the visual words. However, the spatial arrangements of visual words, e.g., co-occurrence, convey important information of aerial scenes. Therefore, some methods were proposed to incorporate the spatial distribution of visual words beyond the BoVW models. For instance, Yang and Newsam [8] developed the *spatial pyramid co-occurrence kernel* to integrate the absolute and relative spatial information ignored in the standard BoVW model setting, by relying on the idea of *spatial pyramid matching kernel* (SPM) [77] and *spatial co-occurrence kernel* [58]. Later, Zhao *et al.* [23] proposed another way to incorporate the spatial information, where wavelet decomposition was utilized in the BoVW model to combine not only the spatial information but also the texture information. Moreover, Zhao *et al.* [24] proposed a concentric circle-based spatial-rotation-invariant representation to encode the spatial information. In [26], a *pyramid-of-spatial-relations* (PSR) model was developed to capture both the absolute and relative spatial relationships of local low-level features. Unlike the conventional co-occurrence approaches [8], [24] that describe pairwise spatial relationships between local features, the PSR model employed a novel concept of spatial relation to describe the relative spatial relationship between a group of local features and reported a better performance.

In addition, to encode higher order spatial information between low-level local visual words for scene modeling, topic models along with the BoVW scheme are developed to take into account the semantic relationship among the visual words [11], [21], [28], [39], [56]. Among them, the *latent Dirichlet allocation* (LDA) [78] model defines an intermediate variable named "topic," which serves as a connection between the visual words and the image. The probability distribution of the topics is estimated by the Dirichlet distribution and was used to describe an image instead of the marginal distribution of visual words with much lower dimensional features. To combine different features, Kusumaningrum *et al.* [21] used CIELab color moments [79], GLCM [80], and *edge orientation histogram* [21] to extract spectral, texture, and structure information, respectively, in the LDA model. In [28],

the *probabilistic latent semantic analysis* (pLSA) model [81] was adopted for scene classification in a multifeature fusion manner and achieved better results than a single feature. In [39], pLSA [81] and LDA [78] were compared using a multifeature fusion strategy to combine three complementary features in a semantic allocation level, and the LDA demonstrated slightly better performances.

Observe that, in all the aforementioned methods, various hand-craft local image descriptors are used to represent aerial scenes. One main difficulty of such methods lies in the fact that they may lack the flexibility and adaptivity to different scenes. In this sense, unsupervised feature learning approaches have been developed to automatically learn adaptive feature representations from images [20], [32], [37]. In [20], a sparse coding-based method was proposed to learn a holistic scene representation from raw pixel values and other low-level features for aerial images. Hu *et al.* [32] discovered the intrinsic space of local image patches by applying different manifold learning techniques and made the dictionary learning and feature encoding more effective. Also with the unsupervised feature learning scheme, Zhang *et al.* [37] extracted the features of image patches by the *sparse autoencoder* [82] and exploited the local spatial and structural information of complex aerial scenes.

C. Methods Based on High-Level Vision Information

Currently, deep learning methods achieve impressive results on many computer vision tasks, such as image classification [83]–[85], and object and scene recognition [86], [87]. This type of methods also achieves the state-of-the-art performance on the aerial scene analysis [4], [33]–[35], [38], [41], [42], [88], [89]. In general, deep learning methods use a multistage global feature learning architecture to adaptively learn image features and often cast the aerial scene classification as an end-to-end problem. Compared with low-level and midlevel methods, deep learning methods can learn more abstract and discriminative semantic features and achieve far better classification performance [4], [33]–[35], [38], [42].

It has been reported that by directly using the pretrained deep neural network architectures on the natural images [90], the extracted global features showed the impressive performance on aerial scene classification [34]. The two freely available pretrained deep convolution neural network (CNN) architectures are OverFeat [91] and CaffeNet [92]. In [33], another promising architecture, i.e., GoogLeNet [84], was considered and evaluated. This architecture also showed astounding performance for aerial images. In [35], it demonstrated that a multiscale input strategy for multiview deep learning can improve the performance of aerial scene classification.

In contrast with directly using the features from the fully connected layer of the pretrained CNN architectures as the final representation [33]–[35], others use the deep-CNN as a local feature extractor and combine it with feature coding techniques. For instance, Hu *et al.* [4] extracted multiscale dense CNN activations from the last convolutional layer as local features descriptors and further coded them using feature encoding methods, such as BoVW [93], *vector of locally aggregated descriptors* (VLAD) [94], and IFK [76] to generate the final image representation. For all the deep-CNN architectures used earlier, either the global or local features were obtained from the networks pretrained on natural image data sets and were directly used for classification aerial images.

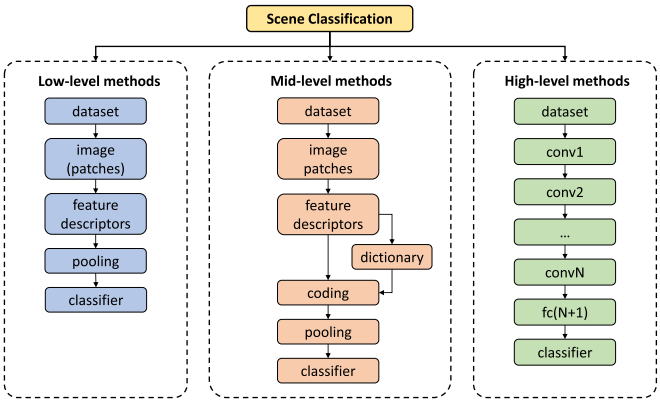


Fig. 4. General pipeline of three types of scene classification methods.

In addition to the above two ways using deep-learning methods, another choice is to train a new deep network. However, as reported in [42], using the existing aerial scene data sets (e.g., UC-Merced data set [58] and the WHU-RS19 data set [57]) to fully train the networks, such as CaffeNet [92] or GoogLeNet [84], showed a drop in accuracies compared with using the networks as global feature extractors. This can be explained by the fact that such large-scale networks with multiple layers usually contain millions of parameters to be learned from the training data, thus, to fully train them usually need millions of training images. However, the existing aerial data sets only contain hundreds or thousands of images, it will easily trap in local optimum and over-fitting problems. Thus, to better fit the data set, smaller networks for classification were trained [38], [41]. In [38], a *gradient boosting random convolutional network* was proposed for classifying aerial images with only two convolutional layers. In [41], a *deep belief network* (DBN) [95] was trained on aerial images, and the feature selection problem was formulated as a feature reconstruction problem in the DBN scheme. By minimizing the reconstruction error over the whole feature set, the features with smaller reconstruction errors can hold more feature intrinsics for image representation. However, the generalization ability of a small network is often lower than that of large-scale networks. It is highly demanded to train a large-scale network with a large number of annotated aerial images.

IV. Baseline Methods

In this section, we evaluate different aerial scene classification methods with low-level, midlevel, and high-level scene descriptions reviewed previously.[4] The general classification pipeline is shown in Fig. 4.

[4]The codes of the baseline methods are downloadable at www.lmars.whu.edu.cn/xia/AID-project.html.

A. Methods With Low-Level Scene Features

Aerial image classification methods using low-level scene features often first partition an aerial image into small patches, then use low-level visual features, e.g., spectral information, texture information, or structure information, to characterize patches, and finally output the distribution of the patch features as the scene descriptor. In our tests, we choose four commonly used low-level methods in the experiment, i.e., SIFT [64], LBP [70], CH [66], and GIST [63].

1) *SIFT [64]:* It describes a patch by the histograms of gradients computed over a 4×4 spatial grid. The gradients are then quantized into eight bins so the final feature vector has a dimension of 128 ($4 \times 4 \times 8$).
2) *LBP [70]:* Some works adopt LBP to extract texture information from aerial images [55], [59]. For a patch, it first compares the pixel to its eight neighbors: when the neighbor's value is less than the center pixel's, output "1," otherwise, output "0." This gives an 8-b binary number to describe the center pixel. The LBP descriptor is obtained by computing the histogram of the binary numbers over the patch and results in a feature vector with 256 dimensions.
3) *Color Histogram [66]:* CHs are used for extracting the spectral information of aerial scenes [16], [55], [59]. In our experiments, CH descriptors are computed separately in three channels of the RGB color space. Each channel is quantized into 32 bins to form a total histogram feature length of 96 by simply concatenation of the three channels.
4) *GIST [63]:* Unlike aforementioned descriptors that focus on local information, GIST represents the dominant spatial structure of a scene by a set of perceptual dimensions (naturalness, openness, roughness, expansion, and ruggedness) based on the spatial envelope model [63] and thus widely used for describing scenes [61]. This descriptor is implemented by convolving the gray image with multiscale (with the number of S) and multidirection (with the number of D) Gabor filters on a 4×4 spatial grid. By concatenating the mean vector of each grid, we get the GIST descriptor of an image with $16 \times S \times D$ dimensions.

B. Methods With Midlevel Scene Features

In contrast with low-level methods, midlevel methods often build a scene representation by coding low-level local feature descriptors [76], [93], [96]. In this paper, we evaluated 21 commonly used midlevel features obtained by combining three local feature descriptors (i.e., SIFT [64], LBP [70], and CH [66]) with seven midlevel feature coding approaches.

1) BoVW [93] models an image by leaving out the spatial information and represents it with the frequencies of local visual words [58]. The BoVW model and its variants are widely used in scene classification [13], [14], [23], [24], [26], [29]–[31], [59], [86]. The visual words are often produced by clustering local image descriptors to form a dictionary (with a given size K), e.g., using the k-means algorithm.
2) *SPM* [77] uses a sequence of increasingly coarser grids to build a spatial pyramid (with L levels) coding of local image descriptors. By concatenating the weighted local image features in each subregion at different scales, one can get a $(((4^L - 1) \times K)/3)$ dimension global feature vector, which is much longer than BoVW with the same size of dictionary (K).
3) *Locality-constrained linear coding* (LLC) [96] is an effective coding scheme adapted from sparse coding methods [9], [18], [60]. It utilizes the locality constraints to code each local descriptor into its local-coordinate system by modifying the sparsity constraints [75], [97]. The final feature can be generated by max pooling of the projected coordinates with the same size of dictionary.
4) The pLSA [81] is a way to improve the BoVW model by topic models. A latent variable called topic is introduced and defined as the conditional probability distribution of visual words in the dictionary. It can serve as a connection between the visual words and images. By describing an image with the distribution of topics (the number of topics is set to be T), one can solve the influence of synonym and polysemy meanwhile reducing the feature dimension to be T.
5) LDA [78] is a generative topic model evolved from the pLSA with the main difference that it adds a Dirichlet prior to describe the latent variable topic instead of the fixed Gaussian distribution, and is also widely used for scene classification [21], [28], [39]. As a result, it can handle the problem of overfitting and also increase the robustness. The dimension of final feature vector is the same as the number of topics T.
6) IFK [76] uses the Gaussian mixture model (GMM) to encode local image features [98] and achieves a good performance in scene classification [22], [30], [31]. In essence, the feature of an image got by the Fisher vector encoding method is a gradient vector of the log likelihood. By computing and concatenating the partial derivatives of the mean and variance of the Gaussian functions, the final feature vector is obtained with the dimension of $2 \times K \times F$ (where F indicates the dimension of the local feature descriptors and K denotes the size of the dictionary).
7) VLAD [94] can be seen as a simplification of the IFK method, which aggregates descriptors based on a locality criterion in a feature space [22]. It uses the nonprobabilistic k-means clustering to generate the dictionary by taking the place of GMM model in IFK. When coding each local patch descriptor to its nearest neighbor in the dictionary, the differences between them in each dimension are accumulated and resulting in an image feature vector with a dimension of $K \times F$.

C. Methods With High-Level Scene Features

In recent years, the learned high-level deep features have been reported to achieve impressive results on aerial image classification [4], [33]–[35], [38], [41], [42]. In this paper,

we also compare three representative high-level deep-learned scene classification methods in our benchmark.

1) *CaffeNet:* Convolutional architecture for fast feature embedding [92] is one of the most commonly used open-source frameworks for deep learning (deep convolutional neural networks in particular). The reference model, CaffeNet, which is almost a replication of ALexNet [83], is proposed for the ILSVRC 2012 competition [90]. The main differences are: 1) there is no data argumentation during training and 2) the order of normalization and pooling is switched. Therefore, it has quite similar performances to the AlexNet [4], [42]. For this reason, we only test CaffeNet in our experiment. The architecture of CaffeNet comprises five convolutional layers, each followed by a pooling layer, and three fully connected layers at the end. In this paper, we directly use the pretrained model obtained using the ILSVRC 2012 data set [90], and extract the activations from the first fully connected layer, which results in a vector of 4096 dimensions for an image.
2) *VGG-VD-16:* To investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting, [85] gives a thorough evaluation of networks by increasing depth using an architecture with very small (3×3) convolution filters, which shows a significant improvement on the accuracies, and can be generalized well to a wide range of tasks and data sets. In this paper, we use one of its best-performing models, named VGG-VD-16, because of its simpler architecture and slightly better results. It is composed of 13 convolutional layers and followed by 3 fully connected layers, and thus results in 16 layers. Similarly, we extract the activations from the first fully connected layer as the feature vectors of the images.
3) *GoogLeNet:* This model [84] won the ILSVRC-2014 competition [90]. Its main novelty lies in the design of the "inception modules," which is based on the idea of "network in network" [99]. By using the inception modules, GoogLeNet has two main advantages: 1) the utilization of filters of different sizes at the same layer can maintain multiscale spatial information and 2) the reduction of the number of parameters of the network makes it less prone to overfitting and allows it to be deeper and wider. Specifically, GoogLeNet is a 22-layer architecture with more than 50 convolutional layers distributed inside the inception modules. Different from the above CNN models, GoogLeNet has only one fully connected layer at last, and therefore, we extract the features of the fully connected layer for testing.

V. Experimental Studies

We evaluate all the three kinds of scene classification methods mentioned earlier: methods with low-level, midlevel, and high-level scene features. For each type, we choose some representative ones as baseline for evaluation: SIFT [64], LBP [70], CH [66], and GIST [63] for low-level methods, BoVW [93], SPM [77], LLC [96], pLSA [81], LDA [78], IFK [76], and VLAD [94] combined with three local feature descriptors (i.e., SIFT [64], LBP [70], and CH [66]) for midlevel methods, and three representative high-level deep-learning methods (i.e., CaffeNet [92], VGG-VD-16 [85], and GoogLeNet [84]) are adopted.

A. Parameter Settings

In our experiment, we first test four kinds of low-level methods for classification: SIFT, LBP, CH, and GIST. For the local patch descriptor, such as SIFT, LBP, and CH, we use a fixed size grid (16×16 pixels) with the spacing step to be 8 pixels to extract all the descriptors in the image plain and adopt the average pooling method for each dimension of the descriptor so as to get the final image features. As for the GIST descriptor, we use it as global descriptors that can extract the feature vectors on the whole image very efficiently. We set the same parameters as in its original work [63]: the number of scales is set to be 4, the orientations are quantized into 8 bins, and a 4×4 spatial grid is utilized for pooling, and thus, it results in 512 dimensions ($4 \times 8 \times 4 \times 4$).

For midlevel methods, we test aforementioned seven different feature coding methods: BoVW, SPM, LLC, pLSA, LDA, IFK, and VLAD. Three local patch descriptors, such as SIFT, LBP, and CH, have been utilized for extracting the local structure, texture, and spectral features, respectively. In the patch sampling procedure, we use the grid sampling as our previous work [30] has proven that grid sampling has a better performance for scene classification of remote sensing imagery. Therefore, we set the patch size to be 16×16 pixels and the grid spacing to be 8 pixels for all the local descriptors to balance the speed/accuracy tradeoff. By combining the three local feature descriptors and seven global feature coding methods, we can get 21 different midlevel features in all. As for the size of the dictionary, we set it from 16 to 8192 for the 7 coding methods when using SIFT as local feature descriptors, and select the optimal one when using LBP and CH for describing local patches. For some special parameters defined in each coding methods, we empirically set the spatial pyramid level to be 2 in SPM, and both the numbers of topics in pLSA and LDA are set to be 64.

For high-level methods, we just use the CNN models pretrained on the ILSVRC-2012 data set [90] and extract the features from the first fully connected layer in each CNN model as the global features, which is unlike many CNN feature-based applications in computer vision which generally adopt the output of the second fully connected layer as the final CNN feature. This setting is according to the observations in the previous work [4], which reported that using the output of the second fully connected layer to describe aerial scenes usually resulted in worse performances.[5] Finally, CaffeNet and VGG-VD-16 result in a vector of 4096 dimensions while GoogLeNet a 1024-D results in feature vector owing to the fact that GoogLeNet has only one fully connected layer, and all the features are $L-2$ normalized for better performance.

[5]To the best of our knowledge, this is mainly because the used CNN models are pretrained on the ILSVRC-2012 data set, but there are big differences between aerial images and natural images, especially in the view angles and imaging conditions. Thus, it is of interest to train our own CNN models by directly relying on the AID data set in the future work.

TABLE II
OA OF DIFFERENT LOW-LEVEL METHODS ON THE UC-MERCED DATA SET, THE WHU-RS19 DATA SET, THE RSSCN7 DATA SET, AND OUR AID DATA SET

Methods	UC-Merced (50%)	UC-Merced (80%)	WHU-RS19 (40%)	WHU-RS19 (60%)	RSSCN7 (20%)	RSSCN7 (50%)	AID (20%)	AID (50%)
SIFT	28.92±0.95	32.10±1.95	25.37±1.32	27.21±1.77	28.45±1.03	32.76±1.25	13.50±0.67	16.76±0.65
LBP	34.57±1.38	36.29±1.90	40.11±1.46	44.08±2.02	57.55±1.18	60.38±1.03	26.26±0.52	29.99±0.49
CH	42.09±1.14	46.21±1.05	48.79±2.37	51.87±3.40	57.20±1.23	60.54±1.01	34.29±0.40	37.28±0.46
GIST	44.36±1.58	46.90±1.76	45.65±1.06	48.82±3.12	49.20±0.63	52.59±0.71	30.61±0.63	35.07±0.41

After getting the global features using various methods, we use the Liblinear [100] for supervised classification, because it can quickly train a linear classifier on large-scale data sets. More specifically, we split the images in the data set into training set and testing set. The features of the training set are used to training a linear classification model by Liblinear, and the features of the testing set are used for estimating the performance of the trained model.

B. Evaluation Protocols

To compare the classification quantitatively, we compute the commonly used measures: overall accuracy (OA) and confusion matrix. OA is defined as the number of correctly predicted images divided by the total number of predicted images. It is a direct measure to reveal the classification performance on the whole data set. A confusion matrix is a specific table layout that allows direct visualization of the performance on each class. Each column of the matrix represents the instances in a predicted class, and each row represents the instances in an actual class, and thus, each item x_{ij} in the matrix computes the proportion of images that predicted to be the ith type meanwhile truly belonging to the jth type.

To compute OA, we adopt two different settings for each tested data set in the supervised classification process. For the RSSCN7 data set and our AID data set, we fix the ratio of the number of training set to be 20% and 50%, respectively, and the left for testing, while for UC-Merced data set, the ratios are set to be 50% and 80%, respectively. For the WHU-RS19 data set, the ratios are fixed at 40% and 60%, respectively. To compute the OA, we randomly split the data sets into training sets and testing sets for evaluation, and repeat it ten times to reduce the influence of the randomness and obtain reliable results. The OA is computed for each run, and the final result is reported as the mean and standard deviation of OA from the individual run.

To compute the confusion matrix, we fix the training set by choosing the same images for fair comparison on each data sets and fix the ratio of the number of training set of the UC-Merced data set, the WHU-RS19 data set, the RSSCN7 data set, and our AID data set to be the commonly used ones, i.e., 80%, 60%, 50%, and 50%, respectively.

C. Experimental Results

In this section, we evaluate different methods on the commonly used UC-Merced data set, WHU-RS19 data set, RSSCN7 data set, as well as our AID data set, and give the corresponding results and analysis, which are divided into four phases: the results of low-level methods, the results of midlevel methods, and the results of high-level methods and confusion matrix.

1) Results With Low-Level Methods: Table II illustrates the means and standard variances of OA using the four kinds of low-level methods (e.g., SIFT, LBP, CH, and GIST) with randomly choosing the fixed percent of images to construct the training set by repeating ten times on the UC-Merced data set, the WHU-RS19 data set, the RSSCN7 data set, and our AID data set. Although different features give different performances on different data sets, we can observe the consistent phenomenon on all data sets that SIFT descriptor performs far less than others with about 20% lower OA than the highest ones, which indicates that the SIFT descriptor is not suitable to be as low-level feature for direct classification. For the other three low-level features, GIST performs the best on the UC-Merced data set, and CH gives the best performances on both WHU-RS19 data set and our AID data set, while LBP and CH give comparable results on the RSSCN7 data set. The different performances can be explained by the characteristics of the data sets, for example, both the UC-Merced data set and our AID data set contain various artificial scene types, which are mainly made up of various buildings, and therefore, GIST, which can extract the dominant spatial structure of a scene, performs well on these data sets. For the RSSCN7 data sets, which contain much natural scene types, thus the texture feature descriptor LBP works the best. In addition, CH gives the most robust performances on all the data sets, because most scene types are color consistent, e.g., grass is mostly green and desert is dark yellow.

2) Results With Midlevel Methods: For the seven kinds of feature coding methods, the size of the dictionary has a great influence on the classification results, and therefore, we need to first find the optimal dictionary size for each coding method. To do so, we fix the local patch descriptor using SIFT, and gradually double increase the dictionary size from 16 to 8192 for each coding methods on the four data sets. The corresponding OA is shown in Fig. 5.

It is worth noticing that the suitable dictionary sizes vary for different types of methods, meaning that too large dictionary size makes no sense for IFK and VLAD, while it will not work to use too small dictionary size for BoVW, LLC, LDA, pLSA, and SPM methods. Thus, in our cases, we tested the dictionary size used for IFK and VLAD on the interval [16, 1024], while the interval [128, 8192] for that of BoVW, LLC, LDA, pLSA, and SPM.

From the results, we can observe that, on the general trend, the algorithms are robust to the dictionary size when it changes in a reasonable interval, but either oversized dictio-

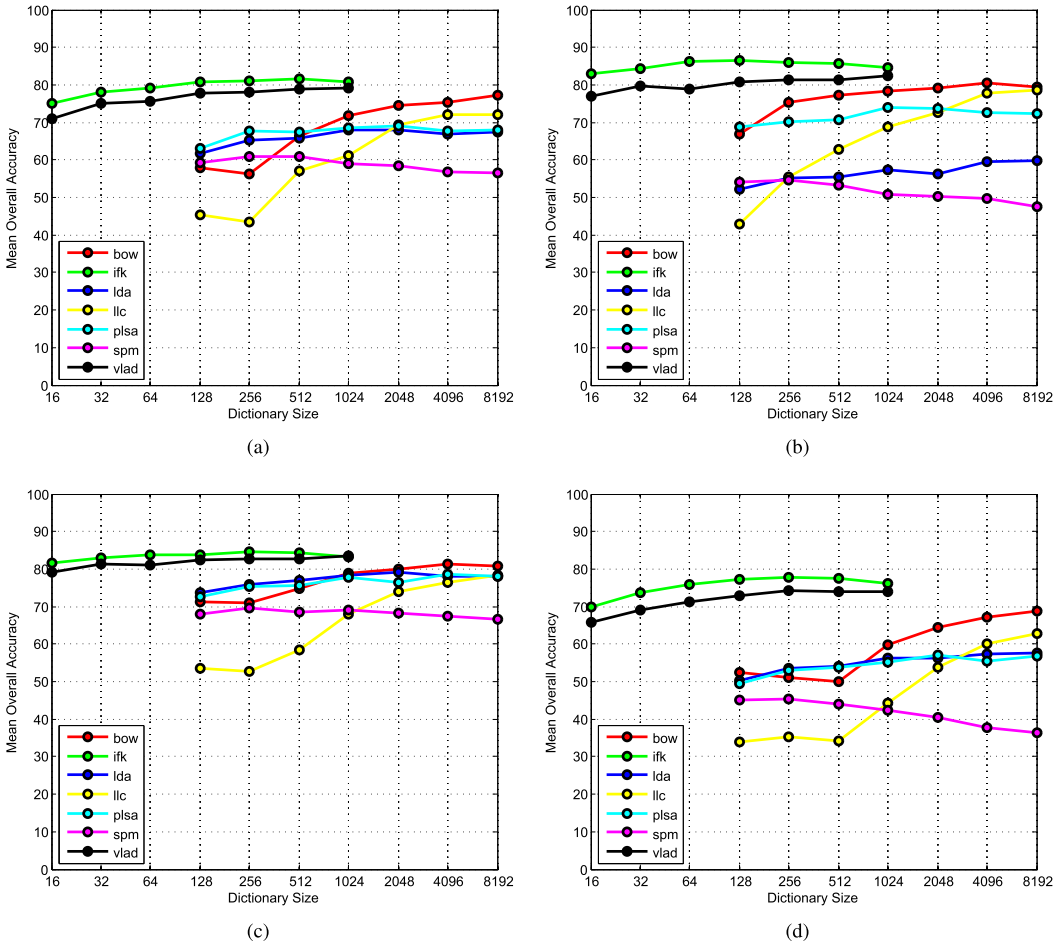


Fig. 5. OA of different midlevel coding methods using different dictionary sizes on (a) UC-Merced, (b) WHU-RS19, (c) RSSCN7, and (d) AID.

nary or undersized dictionary may decrease the performance. For undersized dictionary, different local features will be clustered into the same center, while for oversized dictionary, similar local features will be clustered into different centers, and both will result in less discriminative global features and affect the performance. We can observe little fluctuations of OA when the dictionary size increases for BoVW and LLC. We checked this empirically and found these were mainly due to the local minima of the clustering algorithms, e.g., K-means used in our experiments.

Moreover, note that larger dictionary size will result in much higher dimensional features and thus more time and space complexities for training classification models. Therefore, in our experiments, we set the parameters in order to make a tradeoff between the classification accuracy and the computational complexity of the methods. Thus, in the following experiments, we fix the dictionary size to be 4096 for BoVW and LLC, for IFK and VLAD, we set it at 128, for LDA and pLSA, it is set to be 1024, and for SPM, it is set to be 256.

After finding the proper dictionary size for each coding method, we set the corresponding values for other local patch descriptors and evaluate the 21 kinds of midlevel features obtained by combining seven kinds of global feature coding methods (e.g., BoVW, SPM, LLC, pLSA, LDA, IFK, and VLAD) with three kinds of local feature descriptors (e.g., SIFT, LBP, and CH). Table III shows the means and standard variances of OA on each data set. Surprisingly, when comparing the results using different local feature descriptors, SIFT can give consistent the most robust performances, while in the low-level features, it performs the worst among the low-level methods. This indicates that SIFT is more suitable to be encoded in the midlevel methods to generate more robust feature representation. By comparing the results using different global feature coding methods, IFK, VLAD, and BoVW account for the highest three OA in general, and pLSA and LLC are in the middle, while the left two methods have relatively worse performances. When comparing all the 21 midlevel methods, the features obtained by IFK with an SIFT descriptor give the best or comparable performances on all the data sets in general, which benefits from the combination of the great robustness and invariance of SIFT when describing local patches and the generative and discriminative nature of IFK.

3) Results With High-Level Methods: Table IV illustrates the means and standard variances of OA using the high-level methods (i.e., the features extracted from the first fully connected layer using the pretrained CNN models) on the four data sets. From the classification results, we can see that CaffeNet and VGG-VD-16 give similar performances on all

TABLE III
OA OF DIFFERENT MIDLEVEL METHODS ON THE UC-MERCED DATA SET, THE WHU-RS19 DATA SET, THE RSSCN7 DATA SET, AND OUR AID DATA SET

Methods	UC-Merced (50%)	UC-Merced (80%)	WHU-RS19 (40%)	WHU-RS19 (60%)	RSSCN7 (20%)	RSSCN7 (50%)	AID (20%)	AID (50%)
BoVW (SIFT)	71.90±0.79	74.12±3.30	75.26±1.39	80.13±2.01	76.33±0.88	81.34±0.55	61.40±0.41	67.65±0.49
IFK (SIFT)	77.09±0.82	82.07±1.50	82.14±0.84	86.95±1.31	80.25±0.67	84.41±0.99	70.60±0.42	77.33±0.37
LDA (SIFT)	63.51±1.08	69.52±1.27	56.26±2.21	56.26±3.62	70.28±1.81	77.83±0.66	49.53±0.44	56.52±0.41
LLC (SIFT)	69.41±1.14	71.17±2.09	73.32±2.13	77.42±1.85	73.29±0.97	76.57±0.77	56.36±0.68	59.92±0.63
pLSA (SIFT)	66.12±1.07	68.38±2.37	71.00±1.10	74.45±1.13	74.64±0.87	77.37±0.58	50.97±0.42	54.76±0.40
SPM (SIFT)	56.26±1.56	61.52±2.54	51.16±1.51	53.61±1.95	65.06±1.18	69.04±0.78	38.14±0.75	45.28±0.66
VLAD (SIFT)	73.23±1.02	78.19±1.66	76.37±2.01	80.82±2.15	77.27±0.58	82.31±1.18	65.19±0.61	72.77±0.48
BoVW (LBP)	73.13±1.50	77.12±1.93	70.00±1.28	74.97±1.66	76.98±0.90	81.69±1.11	56.73±0.54	64.25±0.55
IFK (LBP)	77.08±1.37	82.24±1.98	75.33±2.15	80.71±2.51	78.00±0.73	83.34±0.64	65.79±0.36	75.01±0.63
LDA (LBP)	58.31±1.50	64.17±1.51	47.07±3.23	57.74±2.93	67.32±1.77	74.56±0.80	41.52±0.49	46.47±0.55
LLC (LBP)	66.25±1.20	69.50±1.35	70.72±1.22	73.45±2.47	73.04±0.90	76.57±0.83	52.93±0.53	57.01±0.42
pLSA (LBP)	60.15±1.73	62.24±1.47	63.05±1.24	66.89±2.32	70.73±0.55	73.38±0.78	40.32±0.53	44.05±0.33
SPM (LBP)	58.25±1.88	62.36±1.90	54.35±2.18	58.50±2.33	68.87±0.73	72.93±0.93	38.07±0.66	44.90±0.60
VLAD (LBP)	72.75±1.38	76.71±2.03	68.35±1.77	74.89±1.36	76.42±0.90	82.18±0.96	59.44±0.43	69.42±0.85
BoVW (CH)	70.74±1.00	76.17±2.27	61.86±1.71	66.55±1.98	74.97±1.01	81.78±0.67	48.60±0.41	55.74±0.48
IFK (CH)	79.20±1.15	83.79±2.32	74.58±1.90	79.29±2.46	81.00±0.51	86.77±1.11	64.96±0.39	72.76±0.70
LDA (CH)	54.62±1.52	60.71±2.30	48.14±1.95	56.71±1.78	66.33±1.24	72.18±0.88	43.55±0.35	47.16±0.53
LLC (CH)	67.20±1.39	70.14±1.84	62.42±2.07	67.61±1.69	72.03±0.87	76.27±1.00	50.24±0.75	53.32±0.46
pLSA (CH)	54.92±1.16	58.52±1.77	56.70±0.91	59.76±2.04	70.70±0.93	73.16±1.04	44.28±0.58	47.87±0.58
SPM (CH)	58.20±0.96	60.52±1.05	55.95±1.84	59.89±2.17	68.62±0.82	73.32±0.67	41.21±0.38	46.03±0.38
VLAD (CH)	64.75±1.59	68.86±1.93	53.12±1.63	57.11±1.54	67.98±0.63	74.76±0.70	44.67±0.48	53.48±0.78

TABLE IV
OA OF HIGH-LEVEL METHODS ON THE UC-MERCED DATA SET, THE WHU-RS19 DATA SET, THE RSSCN7 DATA SET, AND OUR AID DATA SET

Methods	UC-Merced (50%)	UC-Merced (80%)	WHU-RS19 (40%)	WHU-RS19 (60%)	RSSCN7 (20%)	RSSCN7 (50%)	AID (20%)	AID (50%)
CaffeNet	93.98±0.67	95.02±0.81	95.11±1.20	96.24±0.56	85.57±0.95	88.25±0.62	86.86±0.47	89.53±0.31
VGG-VD-16	94.14±0.69	95.21±1.20	95.44±0.60	96.05±0.91	83.98±0.87	87.18±0.94	86.59±0.29	89.64±0.36
GoogLeNet	92.70±0.60	94.31±0.89	93.12±0.82	94.71±1.33	82.55±1.11	85.84±0.92	83.44±0.40	86.39±0.55

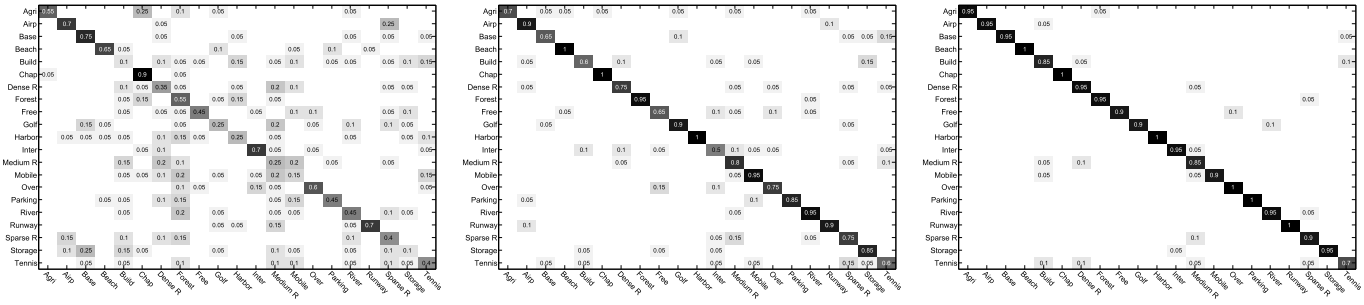
the data sets, while GoogLeNet performs slightly worse. Note that CaffeNet has only 8 layers that is much shallower than the VGG-VD-16 and GoogLeNet, which has 16 and 22 layers, respectively. Superficially, this phenomenon may result in the conclusion that a shallower network works better, which is inconsistent with image classification of natural images. However, note the fact that the networks are all trained by the natural images, we just use them as feature extractors in our experiment. Therefore, the deeper the network, the more likely the learned features oriented to the natural image processing task, which may result in worse performance for classifying aerial scenes.

Compared with the above low-level and midlevel methods, high-level methods show far better performance on both data sets, which indicates that the high-level methods have the ability to learn highly discriminative features. Moreover, note that all the networks we use are pretrained models on the ILSVRC 2012 data set [90], i.e., all the parameters are trained by the natural images, which shows its great generalization ability compared with other methods.

In addition, in all the above methods, the standard deviations of OA on our new data set are much lower than the others, which are mainly caused by the number of testing samples. There are only dozens of images per class for testing the UC-Merced data set and WHU-RS19 data set, and thus, OA will have a greater variation range if the numbers of right predictions in each run are inconsistent with only a few numbers. However, the number of testing images in our new data set is more than ten times larger than the above two, and thus, it will result in much smaller standard variances, which can help to evaluate the performances more precisely.

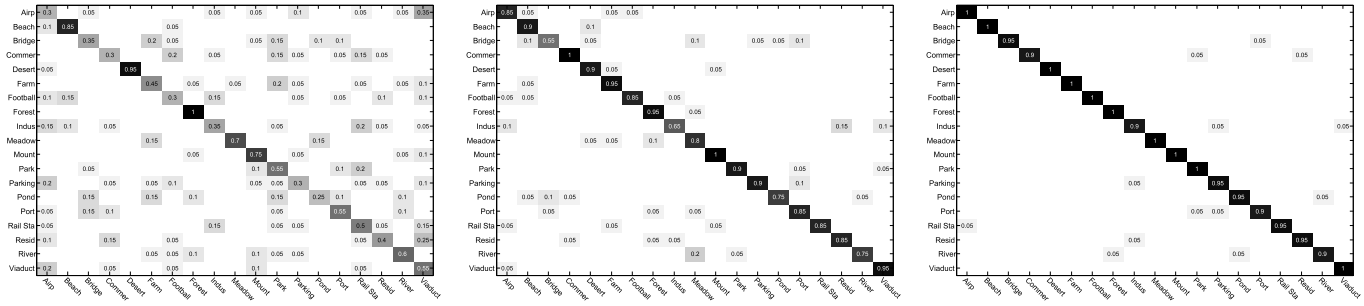
4) Confusion Matrix: Besides giving the OA of various methods, we also compute the corresponding confusion matrix. For each data set, we fix the training set and choose to show the best results of the low-level, midlevel, and high-level methods on the testing set for each data set. Fig. 6 shows the confusion matrix using low level (GIST), midlevel IFK (CH), and high level (VGG-VD-16) on the UC-Merced data set, Fig. 7 shows the results on WHU-RS19 data set, Fig. 8 shows the results on RSSCN7 data set, and Fig. 9 shows our AID data set.

From the confusion matrix on the UC-Merced data set (Fig. 6), we can see that there are only one class (chaparral) obtained the classification accuracy above 0.8, and most classes are easily confused with others using low-level features; when using midlevel features, the classification accuracies of all the scene types increase and more than a half of the classes achieve the classification accuracy above 0.8, while for high-level features, the scene types can be easily distinguished from others that the classification accuracies of most classes



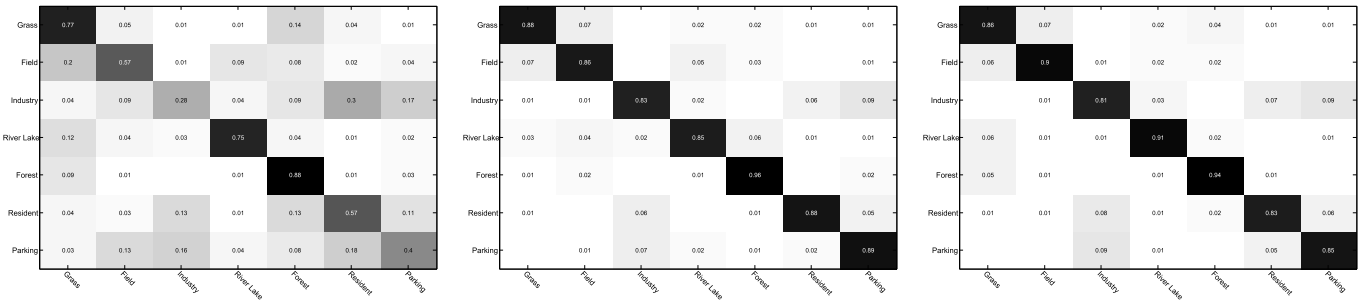
(a) (b) (b)

Fig. 6. Confusion matrix obtained by (a) low level (GIST), (b) midlevel IFK (CH), and (c) high level (VGG-VD-16) on the UC-Merced data set.



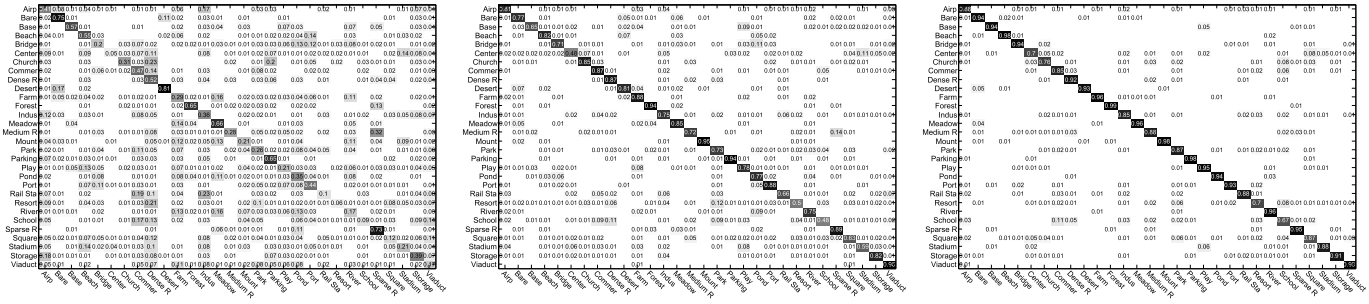
(a) (b) (c)

Fig. 7. Confusion matrix obtained by (a) low level (CH), (b) midlevel IFK (SIFT), and (c) high level (VGG-VD-16) on the WHU-RS19 data set.



(a) (b) (c)

Fig. 8. Confusion matrix obtained by (a) low level (LBP), (b) midlevel IFK (CH), and (c) high level (CaffeNet) on the RSSCN7 data set.



(a) (b) (c)

Fig. 9. Confusion matrix obtained by (a) low level (CH), (b) midlevel IFK (SIFT), and (c) high level (VGG-VD-16) on our AID data set.

are close to or even equal to 1. The most notable confusion is among buildings, dense residential, medium residential, sparse residential, and tennis court, for their similar structures and land cover types.

From the confusion matrix on the WHU-RS19 data set (Fig. 7) and RSSCN7 data set (Fig. 8), we can observe the similar phenomenon that the low-level features cannot distinguish the scene types well, and midlevel features can improve the classification accuracy a lot, while high-level features obtain a quite clean confusion matrix.

When analyzing the confusion matrix on our AID data set (Fig. 9), similarly, the low-level features give the worst performance, while high-level features give the best. Although most scene types can achieve the classification accuracy close

to 1 using high-level features, most of them are natural scene types and thus easy to be distinguished, e.g., beach, forest, and mountain. Note that the most confused scene types in the UC-Merced data set are sparse residential, medium residential and dense residential areas, the corresponding classification accuracies of them in our new data set are around 0.9, which no longer belongs to the moat confused scene types in our new data set. The most difficult scene types in our new data set are almost newly added scene types, i.e., school (0.67), square (0.67), resort (0.7), and center (0.7). The most notable confusion is between school and commercial, for they contain similar structures, e.g., the teaching buildings in the school type, and the shopping malls in the commercial area.

By comparing the confusion matrix among the four data sets, we can conclude that our new data set is more suitable for aerial scene classification than the others for it contains much fine-grained meanwhile challenging scene types.

D. Discussion

From the above experimental results, we can summarize some interesting but meaningful observations as follows.

1) By comparing various scene classification methods, we can observe the layered performances as the name implies: low-level methods have relatively worse performances, while high-level methods perform better on all the data sets, which shows the great potential of high-level methods.
2) By comparing different scene classification data sets, we can find that our new data set is far more challenging than the others for it has relatively higher intraclass variations and smaller interclass dissimilarity. In addition, the large numbers of sample images can help to evaluate various methods more precisely.

The above observations can provide us with very meaningful instructions for investigating more effective high-level methods on more challenging data sets to promote the progress in aerial scene classification.

VI. Conclusion

In this paper, we first give a comprehensive review on aerial scene classification by giving a clear summary of the existing approaches. We find that the results on the current popularly used data sets are already saturated and thus severely limit the progress of aerial scene classification. In order to solve the problem, we construct a new large-scale data set, i.e., AID, which is the largest and most challenging one for the scene classification of aerial images. The purpose of the data set is to provide the research community with a benchmark resource to advance the state-of-the-art algorithms in an aerial scene analysis. In addition, we have evaluated a set of representative aerial scene classification approaches with various experimental protocols on the new data set. These can serve as the baseline results for future works. Moreover, both the data set and the codes are publicly online for freely downloading to promote the development of aerial scene classification.

Acknowledgment

The authors would like to thank all the researchers who kindly shared the codes used in this paper and all the volunteers who helped them constructing the data set.

References

[1] Q. Hu *et al.*, "Exploring the use of Google earth imagery and object-based methods in land use/cover mapping," *Remote Sens.*, vol. 5, no. 11, pp. 6026–6042, 2013.
[2] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.
[3] 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.
[4] 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.
[5] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogram. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.
[6] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.
[7] V. Risojević and Z. Babić, "Aerial image classification using structural texture similarity," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2011, pp. 190–195.
[8] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1465–1472.
[9] 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.
[10] V. Risojević and Z. Babić, "Orientation difference descriptor for aerial image classification," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Apr. 2012, pp. 150–153.
[11] F. Hu, W. Yang, J. Chen, and H. Sun, "Tile-level annotation of satellite images using multi-level max-margin discriminative random field," *Remote Sens.*, vol. 5, no. 5, pp. 2275–2291, 2013.
[12] B. Luo, S. Jiang, and L. Zhang, "Indexing of remote sensing images with different resolutions by multiple features," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 6, no. 4, pp. 1899–1912, Aug. 2013.
[13] W. Shao, W. Yang, G.-S. Xia, and G. Liu, "A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization," in *Computer Vision Systems*. Berlin, Germany: Springer, 2013, pp. 324–333.
[14] W. Shao, W. Yang, and G.-S. Xia, "Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification," *Int. J. Remote Sens.*, vol. 34, no. 23, pp. 8588–8602, 2013.
[15] V. Risojević and Z. Babić, "Fusion of global and local descriptors for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 836–840, Jul. 2013.
[16] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.
[17] B. Zhao, Y. Zhong, and L. Zhang, "Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery," *Remote Sens. Lett.*, vol. 4, no. 12, pp. 1204–1213, 2013.
[18] 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.
[19] A. Avramović and V. Risojević, "Block-based semantic classification of high-resolution multispectral aerial images," *Signal, Image Video Process.*, vol. 10, no. 1, pp. 75–84, Jan. 2016.

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

[21] R. Kusumaningrum, H. Wei, R. Manurung, and A. Murni, "Integrated visual vocabulary in latent Dirichlet allocation–based scene classification for IKONOS image," *J. Appl. Remote Sens.*, vol. 8, no. 1, p. 083690, 2014.

[22] R. Negrel, D. Picard, and P.-H. Gosselin, "Evaluation of second-order visual features for land-use classification," in *Proc. Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2014, pp. 1–5.

[23] L. Zhao, P. Tang, and L. Huo, "A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification," *Int. J. Remote Sens.*, vol. 35, no. 6, pp. 2296–2310, 2014.

[24] L.-J. Zhao, P. Tang, and L.-Z. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 7, no. 12, pp. 4620–4631, Dec. 2014.

[25] Q. Zhu, Y. Zhong, and L. Zhang, "Multi-feature probability topic scene classifier for high spatial resolution remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2014, pp. 2854–2857.

[26] 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.

[27] X. Chen, T. Fang, H. Huo, and D. Li, "Measuring the effectiveness of various features for thematic information extraction from very high resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 4837–4851, Sep. 2015.

[28] Y. Zhong, M. Cui, Q. Zhu, and L. Zhang, "Scene classification based on multifeature probabilistic latent semantic analysis for high spatial resolution remote sensing images," *J. Appl. Remote Sens.*, vol. 9, no. 1, p. 095064, 2015.

[29] H. Sridharan and A. Cheriyadat, "Bag of lines (BoL) for improved aerial scene representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 676–680, Mar. 2015.

[30] J. Hu, G.-S. Xia, F. Hu, and L. Zhang, "A comparative study of sampling analysis in the scene classification of optical high-spatial resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14988–15013, 2015.

[31] J. Hu, T. Jiang, X. Tong, G.-S. Xia, and L. Zhang, "A benchmark for scene classification of high spatial resolution remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 5003–5006.

[32] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.

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

[34] 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.*, Jun. 2015, pp. 44–51.

[35] F. P. S. Luus, B. Salmon, F. van den Bergh, and B. T. J. Maharaj, "Multiview deep learning for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2448–2452, Dec. 2015.

[36] W. Yang, X. Yin, and G. S. Xia, "Learning high-level features for satellite image classification with limited labeled samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4472–4482, Aug. 2015.

[37] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[38] 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.

[39] 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.

[40] 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.

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

[42] K. Nogueira, O. A. Penatti, and J. A. dos Santos. (2016). "Towards better exploiting convolutional neural networks for remote sensing scene classification." [Online]. Available: https://arxiv.org/abs/1602.01517

[43] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2218–2232, Jul. 2009.

[44] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 606–617, Jun. 2011.

[45] T. Blaschke and J. Strobl, "What's wrong with pixels? Some recent developments interfacing remote sensing and GIS," *GeoBIT/GIS*, vol. 6, no. 1, pp. 12–17, 2001.

[46] N. B. Kotliar and J. A. Wiens, "Multiple scales of patchiness and patch structure: A hierarchical framework for the study of heterogeneity," *Oikos*, vol. 59, no. 2, pp. 253–260, 1990.

[47] T. Blaschke, "Object-based contextual image classification built on image segmentation," in *Proc. IEEE Workshop Adv. Techn. Anal. Remotely Sensed Data*, Oct. 2003, pp. 113–119.

[48] Y. Gao, J.-F. Mas, B. Maathuis, X. Zhang, and P. M. Van Dijk, "Comparison of pixel-based and object-oriented image classification approaches—A case study in a coal fire area, Wuda, Inner Mongolia, China," *Int. J. Remote. Sens.*, vol. 27, no. 18, pp. 4039–4055, 2006.

[49] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS J. Photogram. Remote Sens.*, vol. 65, no. 1, pp. 2–16, Jan. 2010.

[50] S. W. Myint, P. Gober, A. Brazel, S. Grossman-Clarke, and Q. Weng, "Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery," *Remote Sens. Environ.*, vol. 115, no. 5, pp. 1145–1161, 2011.

[51] D. C. Duro, S. E. Franklin, and M. G. Dubé, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery," *Remote Sens. Environ.*, vol. 118, pp. 259–272, Mar. 2012.

[52] Y. Zhong, J. Zhao, and L. Zhang, "A hybrid object-oriented conditional random field classification framework for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7023–7037, Nov. 2014.

[53] J. Zhao, Y. Zhong, and L. Zhang, "Detail-preserving smoothing classifier based on conditional random fields for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2440–2452, May 2015.

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

[55] J. A. dos Santos, O. A. B. Penatti, and R. da Silva Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification," in *Proc. VISAPP*, May 2010, pp. 203–208.

[56] M. Liénou, H. Maître, 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.

[57] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *Proc. ISPRS TC 7th Symp. Years ISPRS*, vol. 38. 2010, pp. 298–303.

[58] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, 2010, pp. 270–279.

[59] L. Chen, W. Yang, K. Xu, and T. Xu, "Evaluation of local features for scene classification using vhr satellite images," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, Apr. 2011, pp. 385–388.

[60] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 173–176, Jan. 2011.

[61] V. Risojević, S. Momić, and Z. Babić, "Gabor descriptors for aerial image classification," in *Adaptive and Natural Computing Algorithms*. Berlin, Germany: Springer, 2011, pp. 51–60.

[62] B. Zhao, Y. Zhong, and L. Zhang, "Hybrid generative/discriminative scene classification strategy based on latent Dirichlet allocation for high spatial resolution remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2013, pp. 196–199.

[63] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

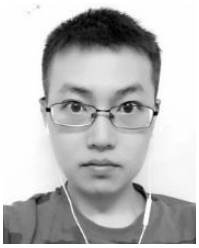
[64] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[65] G.-S. Xia, J. Delon, and Y. Gousseau, "Accurate junction detection and characterization in natural images," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 31–56, Jan. 2014.
[66] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.
[67] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.
[68] G.-S. Xia, J. Delon, and Y. Gousseau, "Shape-based invariant texture indexing," *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 382–403, Jul. 2010.
[69] G. Liu, G.-S. Xia, W. Yang, and L. Zhang, "Texture analysis with shape co-occurrence patterns," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 1627–1632.
[70] T. Ojala, M. Pietikäinen, and T. Mäenpää, "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.
[71] B. Luo, J. F. Aujol, Y. Gousseau, and S. Ladjal, "Indexing of satellite images with different resolutions by wavelet features," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1465–1472, Aug. 2008.
[72] B. Luo, J.-F. Aujol, and Y. Gousseau, "Local scale measure from the topographic map and application to remote sensing images," *Multiscale Model. Simul.*, vol. 8, no. 1, pp. 1–29, 2009.
[73] S. Mallat and L. Sifre, "Combined scattering for rotation invariant texture analysis," in *Proc. ESANN*, Apr. 2012, pp. 68–81.
[74] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult, "Multi-attribute spaces: Calibration for attribute fusion and similarity search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2933–2940.
[75] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1794–1801.
[76] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 143–156.
[77] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. 2006, pp. 2169–2178.
[78] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[79] M. Stricker and M. Orengo, "Similarity of color images," in *Proc. IST/SPIE's Symp. Electron. Imag., Sci. Technol.*, 1995, pp. 381–392.
[80] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
[81] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 517–530.
[82] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 3371–3408, Dec. 2010.
[83] 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.
[84] C. Szegedy *et al.* (2014). "Going deeper with convolutions." [Online]. Available: https://arxiv.org/abs/1409.4842
[85] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[86] B. Zhou, A. L. Garcia, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1. 2015, pp. 487–495.
[87] Y. Yuan, L. Mou, and X. Lu, "Scene recognition by manifold regularized deep learning architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2222–2233, Oct. 2015.
[88] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
[89] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.
[90] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, pp. 1–42, Apr. 2015.
[91] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. (2013). "OverFeat: Integrated recognition, localization and detection using convolutional networks." [Online]. Available: https://arxiv.org/abs/1312.6229
[92] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[93] 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.
[94] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.
[95] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[96] 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.
[97] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2223–2231.
[98] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[99] L. Min, C. Qiang, and S. Yan. (2013). "Network in network." [Online]. Available: https://arxiv.org/abs/1312.4400
[100] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

Gui-Song Xia (M'10–SM'15) received the B.S. degree in electronics engineering and the M.S. degree in signal processing from Wuhan University, Wuhan, China, in 2005 and 2007, respectively, and the Ph.D. degree in image processing and computer vision from the Centre National de la Recherche Scientifique (CNRS) Laboratoire Traitement et Communication de l'Information, TELECOM ParisTech, Paris, France, in 2011.

In 2011, he was a Post-Doctoral Researcher with the Centre de Recherche en Mathmatiques de la Decision, CNRS, Paris-Dauphine University, Paris, for one and a half years. He is currently a Full Professor with the State Key Laboratory of Information Engineering, Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include mathematical image modeling, texture synthesis, image indexing and content-based retrieval, structure from motion, perceptual grouping, and remote sensing imaging.

Jingwen Hu received the B.S. degree in electronic information engineering from Wuhan University, Wuhan, China, in 2013, where she is currently pursuing the Ph.D. degree in signal and information processing.

Her research interests include scene classification for high spatial resolution remote sensing imagery and deep learning.



Fan Hu (S'16) received the B.S. degree in communication engineering from Wuhan University, Wuhan, China, in 2011, where he is currently pursuing the Ph.D. degree with the Signal Processing Laboratory, Electronic Information School.

His research interests include high-resolution image classification, and machine learning, especially deep learning and their applications in remote sensing.

Baoguang Shi received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2012, where he is currently pursuing the Ph.D. degree with the School of Electronic Information and Communications.

His research interests include scene text detection and recognition, script identification, and face alignment.

Xiang Bai (SM'16) received the B.S., M.S., and Ph.D. degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003, 2005, and 2009, respectively, all in electronics and information engineering.

He is currently a Professor with the School of Electronic Information and Communications and the Vice-Director of the National Center of Anti-Counterfeiting Technology, HUST. His research interests include object recognition, shape analysis, and scene text recognition.

Yanfei Zhong (M'11–SM'15) received the B.S. degree in information engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

He has been with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, since 2007, where he is currently a Full Professor. He was a Referee of more than 30 international journals. He has authored over 100 research papers, including more than 45 peer-reviewed articles published in international journals. His research interests include multispectral and hyperspectral remote sensing data processing, high-resolution image processing and scene analysis, and computational intelligence.

Dr. Zhong was a recipient of the National Excellent Doctoral Dissertation Award of China and the New Century Excellent Talents in University of China. He is serving as an Associate Editor of the *International Journal of Remote Sensing* and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Liangpei Zhang (M'06–SM'08) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He was a Principal Scientist of the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He is currently the Head of the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is also a Chang-Jiang Scholar Chair Professor appointed by the Ministry of Education of China. He has authored over 500 research papers and five books. He holds 15 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of the Institution of Engineering and Technology, an Executive Member (Board of Governor) of the China National Committee of International Geosphere–Biosphere Programme, and an Executive Member of the China Society of Image and Graphics. He received the best reviewer awards from the IEEE Geoscience and Remote Sensing Society (GRSS) for his service to the IEEE JOURNAL OF SELECTED TOPICS IN EARTH OBSERVATIONS AND APPLIED REMOTE SENSING (JSTARS) in 2012 and the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2014. He was a recipient of the 2010 Best Paper Boeing Award and the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing. His research teams received the top three prizes of the IEEE GRSS 2014 Data Fusion Contest, and his students have been selected as the winners or finalists of the IEEE International Geoscience and Remote Sensing Symposium student paper contest in recent years. He was the General Chair for the 4th IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing and the Guest Editor of JSTARS. He is the Founding Chair of the IEEE GRSS Wuhan Chapter. He regularly serves as a Co-Chair of the series SPIE conferences on multispectral image processing and pattern recognition, the Conference on Asia Remote Sensing, and many other conferences. He is an Editor of several conference proceedings, issues, and geoinformatics symposiums. He also serves as an Associate Editor of the *International Journal of Ambient Computing and Intelligence*, the *International Journal of Image and Graphics*, the *International Journal of Digital Multimedia Broadcasting*, the *Journal of Geo-spatial Information Science*, and the *Journal of Remote Sensing*, and the Guest Editor of the *Journal of Applied Remote Sensing* and the *Journal of Sensors*. He is currently serving as an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Xiaoqiang Lu (M'14–SM'15) is currently a Full Professor with the State Key Laboratory of Transient Optics and Photonics, Center for OPTical IMagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His research interests include pattern recognition, machine learning, hyperspectral image analysis, cellular automata, and medical imaging.