Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities

Gong Cheng, *Member, IEEE*, Xingxing Xie, Junwei Han, *Senior Member, IEEE*, Lei Guo, and Gui-Song Xia, *Senior Member, IEEE*

***Abstract*—Remote sensing image scene classification, which aims at labeling remote sensing images with a set of semantic categories based on their contents, has broad applications in a range of fields. Propelled by the powerful feature learning capabilities of deep neural networks, remote sensing image scene classification driven by deep learning has drawn remarkable attention and achieved significant breakthroughs. However, to the best of our knowledge, a comprehensive review of recent achievements regarding deep learning for scene classification of remote sensing images is still lacking. Considering the rapid evolution of this field, this article provides a systematic survey of deep learning methods for remote sensing image scene classification by covering more than 160 papers. To be specific, we discuss the main challenges of remote sensing image scene classification and survey: first, autoencoder-based remote sensing image scene classification methods; second, convolutional neural network-based remote sensing image scene classification methods; and third, generative adversarial network-based remote sensing image scene classification methods. In addition, we introduce the benchmarks used for remote sensing image scene classification and summarize the performance of more than two dozen of representative algorithms on three commonly used benchmark datasets. Finally, we discuss the promising opportunities for further research.**

***Index Terms*—Deep learning, remote sensing image, scene classification.**

I. Introduction

REMOTE sensing images, a valuable data source for earth observation, can help us to measure and observe detailed structures on the Earth's surface. Thanks to the advances of earth observation technology [1], [2], the volume of remote sensing images is drastically growing. This has given particular urgency to the quest for how to make full use of ever-increasing remote sensing images for intelligent earth observation [3], [4]. Hence, it is extremely important to understand huge and complex remote sensing images. As a key and challenging problem for effectively interpreting remote sensing imagery, scene classification of remote sensing images has been an active research area. Remote sensing image scene classification is to correctly label given remote sensing images with predefined semantic categories, as shown in Fig. 1. For the last few decades, extensive research works on remote sensing image scene classification have been undertaken driven by its real-world applications, such as urban planning [5], [6], natural hazards detection [7]–[9], environment monitoring [10]–[12], vegetation mapping [13], [14], and geospatial object detection [15]–[22].

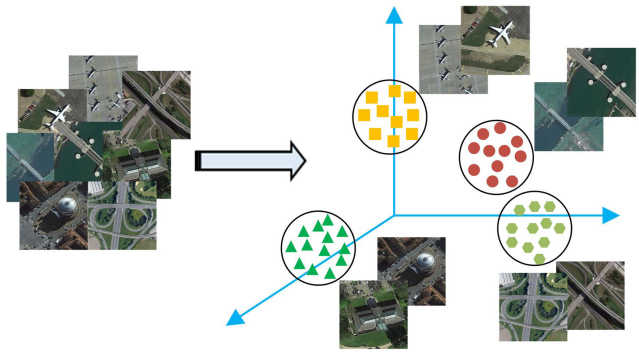


Fig. 1. Illustration of remote sensing image scene classification, which aims at labeling each remote sensing image patch with a semantic class based on its content.

With the improvement of spatial resolution of remote sensing images, remote sensing image classification gradually formed three parallel classification branches at different levels: pixel-level, object-level, and scene-level classification, as shown in Figs. 2 and 3. Here, it is worth mentioning that we use the term of "remote sensing image classification" as a general concept, which includes pixel-level, object-level, and scene-level classification of remote sensing images. To be specific, in the early literatures, researchers mainly focused on classifying remote

Manuscript received April 1, 2020; revised May 30, 2020 and June 18, 2020; accepted June 24, 2020. Date of publication June 29, 2020; date of current version July 8, 2020. This work was supported in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grant JCYJ20180306171131643, in part by the National Science Foundation of China under Grant 61772425 and Grant 61773315, and in part by the Fundamental Research Funds for the Central Universities under Grant 3102019AX09. *(Corresponding author: Junwei Han.)*

Gong Cheng and Xingxing Xie are with the Research and Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China, and also with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: chenggong1119@gmail.com; xiexing@mail.nwpu.edu.cn).

Junwei Han and Lei Guo are with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: junweihan2010@gmail.com; lguo@nwpu.edu.cn).

Gui-Song Xia is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: guisong.xia@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.3005403

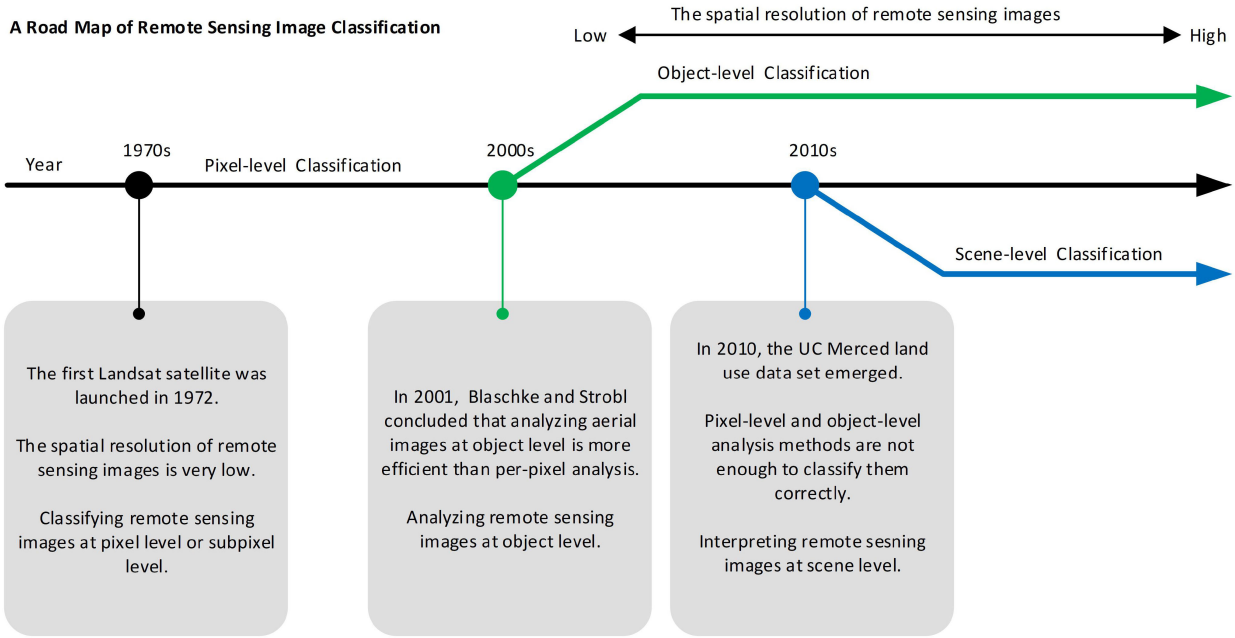


Fig. 2. Road map of remote sensing image classification. With the improvement of spatial resolution of remote sensing images, remote sensing image classification gradually formed three parallel classification branches at different levels: pixel-level, object-level, and scene-level classification. Here, it is worth mentioning that we use "remote sensing image classification" as a general concept.

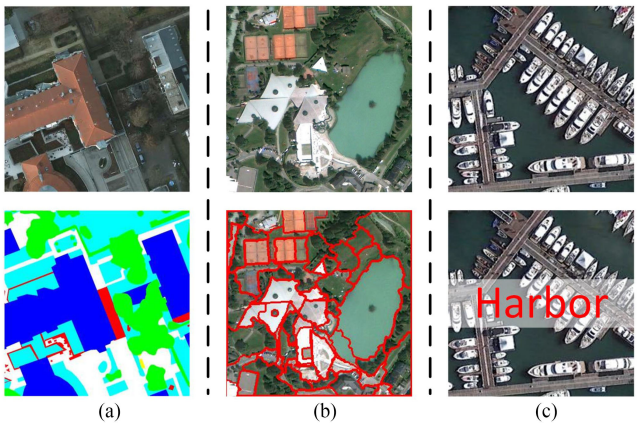


Fig. 3. Three levels of remote sensing image classification. (a) Pixel-level remote sensing image classification focuses on labeling each pixel with a class. (b) Object-level remote sensing image classification aims at recognizing objects in remote sensing images. (c) Scene-level remote sensing image classification seeks to classify each given remote sensing image patch into a semantic class. This survey focuses on scene-level remote sensing image classification.

sensing images at pixel level or subpixel level [23]–[25], through labeling each pixel in the remote sensing images with a semantic class, because the spatial resolution of remote sensing images is very low—the size of a pixel is similar to the sizes of the objects of interest [26]. To date, pixel-level remote sensing image classification [sometimes also called semantic segmentation, as shown in Fig. 3(a)] is still an active research topic in the areas of multispectral and hyperspectral remote sensing image analysis [27]–[31].

Due to the advancement of remote sensing imaging, the spatial resolution of remote sensing images is increasingly finer than common objects of interest, such that single pixels lose their semantic meanings. In such case, it is not feasible to recognize scene images at the pixel level solely and so per-pixel analysis began to be viewed with increasing dissatisfaction. In 2001, Blaschke and Strobl [32] questioned the dominance of per-pixel research paradigm and concluded that analyzing remote sensing images at the object level is more efficient than per-pixel analysis. They suggested that researchers should pay attention to object-level analysis, which aims at recognizing objects in remote sensing images, as shown in Fig. 3(b), where the term "object" refers to meaningful semantic entities or scene units. Subsequently, a series of approaches to analyze remote remote sensing images at object level has dominated remote sensing image analysis for the last two decades [33]–[36]. Amazing achievements of certain specific land use identification tasks have been accomplished by pixel-level and object-level classification algorithms.

However, remote sensing images may contain different and distinct object classes because of the increasing resolutions of remote sensing images. Pixel-level and object-level methods may not be sufficient to always classify them correctly. Under the circumstances, it is of considerable interest to understand the global contents and meanings of remote sensing images. A new paradigm of scene-level analysis of remote sensing images has been recently suggested. Scene-level remote sensing image classification, namely remote sensing image scene classification, seeks to classify each given remote sensing image patch (e.g., 256×256) into a semantic class, as illustrated in Fig. 3(c). Here, the item "scene" represents an image patch cropped from

a large-scale remote sensing image that contains clear semantic information on the earth surface [37], [38].

It is a significant step to be able to represent visual data with discriminative features in almost all tasks of computer vision. The remote sensing domain is no exception. During the previous decade, extensive efforts have been devoted to developing discriminative visual features. A majority of early remote sensing image scene classification methods relied on human-engineering descriptors, e.g., scale-invariant feature transformation (SIFT) [39], texture descriptors (TD) [40]–[42], color histogram (CH) [43], histogram of oriented gradients (HOG) [44], and GIST [45]. Owing to their characteristic of being able to represent an entire image with features, it is feasible to directly apply CH, GIST, and TD to remote sensing image scene classification. However, SIFT and HOG cannot represent an entire image directly because of their local characteristic. To make handcrafted local descriptors represent an entire scene image, these local descriptors are encoded by certain encoding methods (e.g., the improved fisher kernel (IFK) [46], vector of locally aggregated descriptors [47], spatial pyramid matching (SPM) [48], and the popular bag-of-visual-words (BoVW) [49]). Thanks to the simplicity and efficiency of these feature encoding methods, they have been broadly applied to the field of remote sensing image scene classification [50]–[55], whereas the representation capability of handcrafted features is limited.

In this case, unsupervised learning, such as *k*-means clustering, principal component analysis (PCA) [56], and sparse coding [57], which automatically learns features from unlabeled images, become an appealing alternative to human-engineering features. A considerable amount of unsupervised learning-based scene classification methods have emerged [58]–[66], and made substantial progress for scene classification. Nevertheless, these unsupervised learning approaches cannot make full use of data class information.

Fortunately, due to the advances in deep learning theory and the increased availability of remote sensing data and parallel computing resources, deep learning-based algorithms have increasingly prevailed the area of remote sensing image scene classification. In 2006, Hinton and Salakhutdinov [67] created an approach to initialize the weights for training multilayer neural networks, which builds a solid foundation for the development of deep learning later. During the period 2006 to 2012, simple deep learning models have been developed (e.g., deep belief nets [68], autoencoder [67], and stacked autoencoder [69]).

The feature description capabilities of these simple deep learning models have been demonstrated in many fields, involving remote sensing image scene classification. Since the AlexNet, a deep convolutional neural network (CNN) designed by Krizhevskey *et al.* [70] in 2012, obtained the best results in the large-scale visual recognition challenge (LSVRC) [71], a great many advanced deep CNNs have come forth and broken a number of records in many fields. In the wake of these successes, CNN-based methods have emerged in remote sensing image scene classification [72]–[74] and achieved advanced classification accuracy. Nevertheless, CNN-based methods generally demand massive annotated training data, which greatly limits their application scenarios. More recently,

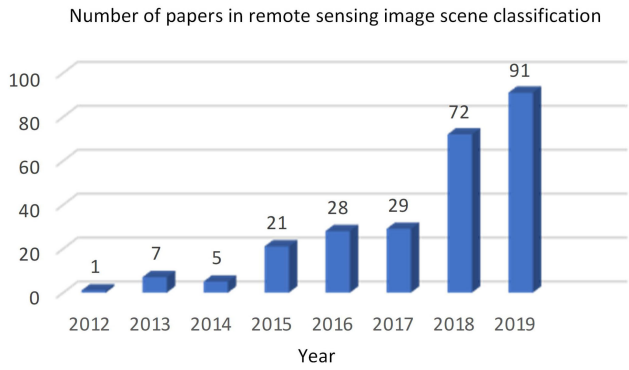


Fig. 4. Number of publications in remote sensing image scene classification from 2012 to 2019. Data from Google scholar advanced search: allintitle: ("remote sensing" or "aerial" or "satellite" or "land use") and "scene classification."

generative adversarial networks (GANs) [82], a promising unsupervised learning method, have achieved significant success in many applications. To remedy the abovementioned limitations, GANs have been employed by some researchers on the field of remote sensing image scene classification [83], [84].

Currently, driven by deep learning, a great number of methods of remote sensing image scene classification have sprung up (see Fig. 4). The number of papers in remote sensing image scene classification dramatically increased after 2014 and 2017, respectively. There are two reasons for the increase. On one hand, around 2014, deep learning techniques began to be applied to remote sensing data analysis. On the other hand, in 2017, large-scale remote sensing image scene classification benchmarks appeared, which have greatly facilitated the development of deep learning-based remote sensing image scene classification.

In the past several years, numerous reviews of remote sensing image classification methods have been published, which are summarized in Table I. For example, Tuia *et al.* [25] surveyed, tested, and compared three active learning-based remote sensing image scene classification methods: committee, large margin, and posterior probability. GóChova *et al.* [2] surveyed multimodal remote sensing image classification and summarized the leading algorithms for this field. In [78], Maulik *et al.* conducted a review of remote sensing image scene classification algorithms based on support vector machine (SVM). Li *et al.* [75] surveyed the pixel-level, subpixel-level, and object-based methods of image classification and emphasized the contribution of spatio-contextual information to remote sensing image scene classification.

As an alternative way to extract robust, abstract, and high-level features from images, deep learning models have made amazing progress on a broad range of tasks in processing image, video, speech, and audio. After this, a number of deep learning-based scene classification algorithms were proposed, such as CNN-based methods and GAN-based methods. A number of reviews of scene classification approaches have been published. Penatti *et al.* [85] assessed the generalization ability of pretrained CNNs in classification of remote sensing images. In [38], Hu *et al.* surveyed how to apply the CNNs that trained on the ImageNet dataset to remote sensing image scene classification. Zhu *et al.*

TABLE I
SUMMARIZATION OF A NUMBER OF SURVEYS OF REMOTE SENSING IMAGE ANALYSIS

No.	Survey Title	Year	Publication	Content
1	A survey of active learning algorithms for supervised remote sensing image classification [25]	2011	IEEE JSTSP	Surveying and testing the main families of active learning methods
2	A review of remote sensing image classification techniques: the role of spatio-contextual information [75]	2014	EuJRS	Review of pixel-wise, subpixel-wise and object-based methods for remote sensing image classification and exploring the contribution of spatio-contextual information to scene classification
3	Multimodal classification of remote sensing images: a review and future directions [2]	2015	Proceedings of the IEEE	Offering a taxonomical view of the field of multimodal remote sensing image classification
4	Deep learning for remote sensing data: A technical tutorial on the state of the art [76]	2016	IEEE GRSM	Reviewing deep learning-based remote sensing data analysis techniques before 2016
5	Deep learning in remote sensing: A comprehensive review and list of resources [77]	2017	IEEE GRSM	Reviewing the progress of deep learning-based remote sensing data analysis before 2017
6	Advanced spectral classifiers for hyperspectral images: A review [27]	2017	IEEE GRSM	Review and comparison of different supervised hyperspectral classification methods
7	Remote sensing image classification: a survey of support-vector-machine-based advanced techniques [78]	2017	IEEE GRSM	Review of remote sensing image classification based on SVM
8	AID: a benchmark data set for performance evaluation of remote sensing image scene classification [79]	2017	IEEE TGRS	Review of aerial image scene classification methods before 2017 and proposing the AID data set
9	Remote sensing image scene classification: benchmark and state of the art [80]	2017	Proceedings of the IEEE	Reviewing the progress of scene classification of remote sensing images before 2017 and proposing the NWPU-RESISC45 data set
10	Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines [28]	2017	IEEE TGRS	Survey of the progress in the classification of spectral–spatial hyperspectral images
11	Deep learning for hyperspectral image classification: An overview [29]	2019	IEEE TGRS	Review of hyperspectral image classification based on deep learning
12	Deep learning in remote sensing applications: A meta-analysis and review [81]	2019	ISPRS JPRS	Providing a review of the applications of deep learning in remote sensing image analysis
13	**Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities**	**2020**	**IEEE JSTARS**	**A systematic review of recent advances in remote sensing image scene classification driven by deep learning**

[77] presented a tutorial about deep learning-based remote sensing data analysis. In order to make full use of pretrained CNNs, Nogueira *et al.* [86] analyzed the performance of CNNs for remote sensing image scene classification with different learning strategies: full training, fine tuning, and using CNNs as feature extractors. In [76], Zhang *et al.* reviewed the recent deep learning-based remote sensing data analysis. Considering the number of scene categories and the accuracy saturation of the existing scene classification datasets, Cheng *et al.* [80] released a large-scale scene classification benchmark, named NWPU-RESISC45, and provided a survey of recent advance in remote sensing image scene classification before 2017. In [79], Xia *et al.* proposed a novel benchmark, called AID, for aerial image classification and reviewed the existing methods of scene classification before 2017. Ma *et al.* [81] provided a review of the applications of deep learning in remote sensing image analysis. In addition, there have been several hyperspectral image classification surveys [27]–[29].

However, a thorough survey of deep learning for scene classification is still lacking. This motivates us to deeply analyze the main challenges faced for remote sensing image scene classification, systematically review those deep learning-based scene classification approaches, most of which are published during the last five years, introduce the mainstream scene classification benchmarks, and discuss several promising future directions of scene classification.

The rest of this article is organized as follows. Section II discusses the current main challenges of remote sensing image scene classification. A brief review of deep learning models and a comprehensive survey of deep learning-based scene classification methods are provided in Section III. The scene classification datasets are introduced in Section IV. In Section V, the comparison and discussion of the performance of deep learning-based scene classification methods on three widely used scene classification benchmarks are given. In Section VI, we discuss the promising future directions of scene classification. Finally, we conclude this article in Section VII.

II. MAIN CHALLENGES OF REMOTE SENSING IMAGE SCENE CLASSIFICATION

The ideal goal of scene classification of remote sensing images is to correctly label the given remote sensing images with their corresponding semantic classes according to their contents, for example, categorizing a remote sensing image from urban into residential, commercial, or industrial area.

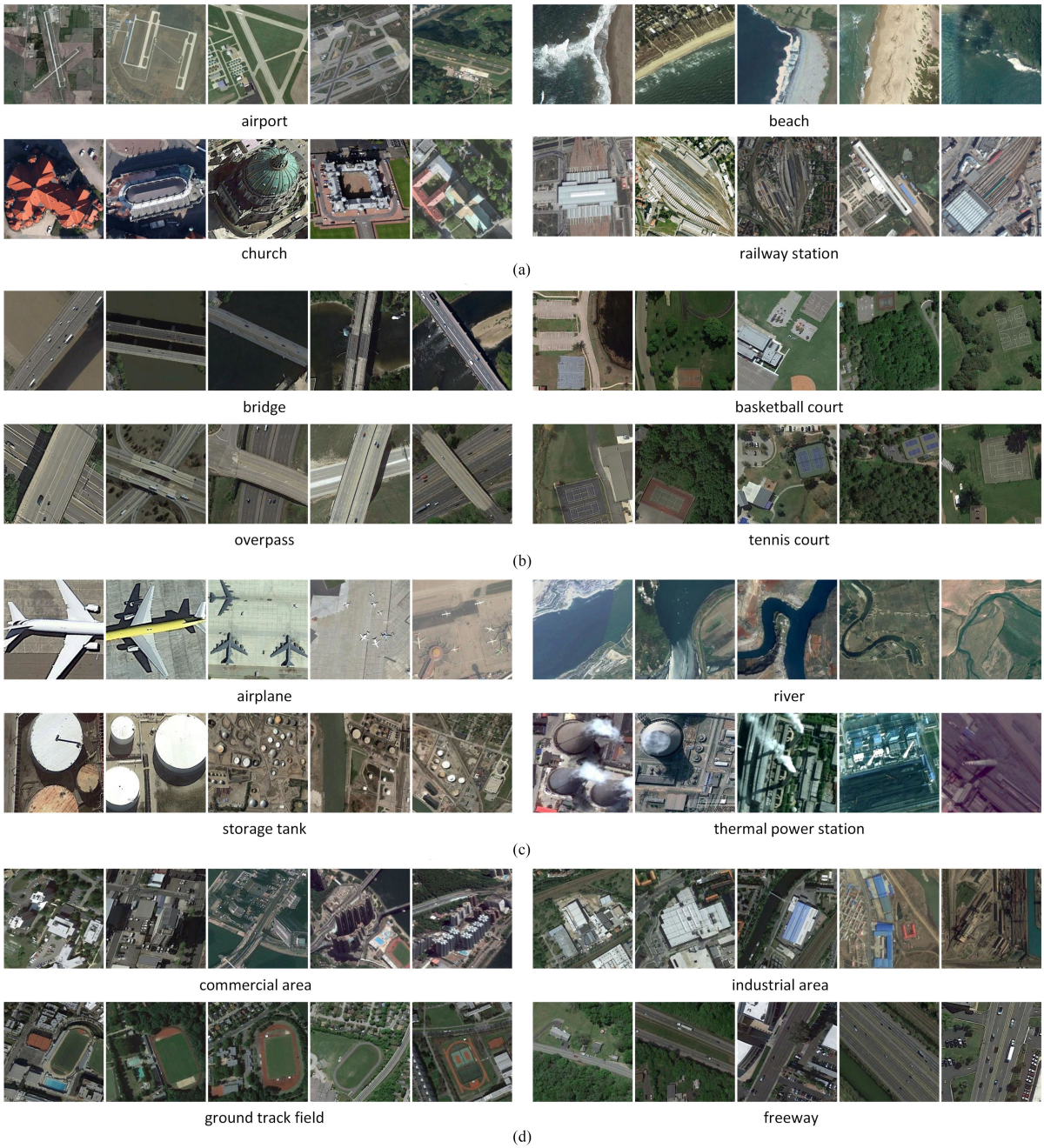


Fig. 5. Challenges of remote sensing image scene classification, which include (a) big within-class diversity, (b) high between-class similarity (also known as low between-class separability), (c) large variance of object/scene scales, and (d) coexistence of multiple ground objects. These images are from the NWPU-RESISC45 dataset [80].

Generally speaking, a remote sensing image contains a variety of ground objects. For instance, roads, trees, and buildings may be included in an industrial scene. Different from object-oriented classification, scene classification is a considerably challenging problem because of the variance and complex spatial distributions of ground objects existing in the scenes. Historically, extensive studies of remote sensing image scene classification have been made. However, there has not yet been an algorithm that can achieve the goal of classifying remote sensing image scenes with satisfactory accuracy. The challenges of remote sensing image scene classification include the following:

1) big intraclass diversity;
2) high interclass similarity (also known as low between-class separability);
3) large variance of object/scene scales;
4) coexistence of multiple ground objects, as shown in Fig. 5.

In terms of within-class diversity, the challenge mainly stems from the large variations in the appearances of ground objects within the same semantic class. Ground objects commonly

vary in style, shape, scale, and distribution, which makes it difficult to correctly classify the scene images. For example, in Fig. 5(a), the churches appear in different building styles, and the airports and railway stations show in different shapes. In addition, when airborne or space platforms capture remote sensing images, there may be large differences in color and radiation intensity appearing within the same semantic class on account of the imaging conditions, which can be influenced by the factors such as weather, cloud, mist, etc. The variations in scene illumination may also cause within-class diversity, for example, the appearances of the scene labeled as "beach" show large differences under different imaging conditions, as shown in Fig. 5(a).

For between-class similarity, the challenge is chiefly caused by the presence of the same objects within different scene classes or the high semantic overlapping between scene categories. For instance, in Fig. 5(b), the scene classes of bridge and overpass both contain the same ground objects, namely bridge, and the basketball courts and tennis courts share high semantic information. Moreover, the ambiguous definition of scene classes degenerates interclass dissimilarity. Some complex scenes are also similar with each other in terms of their visual contents. Therefore, it may be extremely difficult to distinguish these scene classes.

The large variance of object/scene scales is also a nonnegligible challenge for remote sensing image scene classification. In remote sensing imaging, sensors operate at the orbits of various altitudes, from a few hundred kilometers to more than ten thousand kilometers, which leads to imaging altitude variation. With the examples illustrated in Fig. 5(c), the scenes of airplane, storage tank, and thermal power station have huge scale differences under different imaging altitudes. In addition, because of some intrinsic factors, the variations in size for each object/scene category can also exist, for example, the rivers shown in Fig. 5(c) are presented in several different subscenes—stream, brook, and creek.

Moreover, owing to the complex and diverse distribution of ground objects and the wide birds-eye perspective of remote sensing imaging equipments, it is quite common that multiple ground objects appear in a single remote sensing image. As illustrated in Fig. 5(d), the scenes of commercial areas may contain buildings, cars, rivers, roads, parking lots, meadows, swimming pools, and playgrounds; roads, trees, bridges, rivers, and cars can coexist in the scenes of industrial areas; the scenes of ground track fields may accompany with the presence of swimming pools, cars, roads, meadows, and trees; the scenes of freeways contain meadows, trees, buildings, cars, rivers, bridges, forests, parking lots, etc. Faced with the situation, it is difficult for single-label remote sensing image scene classification to provide deep understanding for the contents of remote sensing images.

III. Survey on Deep Learning-Based Remote Sensing Image Scene Classification Methods

In the past decades, many researchers have committed to scene classification of remote sensing images, driven by its wide applications. A number of advanced scene classification systems or approaches have been proposed, especially driven by deep learning. Before deep learning came to the attention of this field, scene classification methods mainly relied on handcrafted features (e.g., color histogram (CH), texture descriptors (TD), GIST) or the representations generated by encoding local features via BoVW, IFK, SPM, etc. Later, considering that handcrafted features only extract low-level information, many researchers turned to look at unsupervised learning methods (e.g., sparse coding, PCA, and *k*-means). By automatically learning discriminative features from unlabeled data, unsupervised learning-based methods have obtained good results in the scene classification of remote sensing images. Yet, unsupervised learning-based algorithms do not adequately exploit data class information, which limits their abilities to discriminate between different scene classes. Now, thanks to the availability of enormous labeled data, the advances in machine learning theory and the increased availability of computational resources, deep learning models (e.g., autoencoder, CNNs, and GANs) have shown powerful abilities to learn fruitful features and have permeated many research fields, including the area of remote sensing image scene classification. Currently, numerous deep learning-based scene classification algorithms have emerged and have yielded the best classification accuracy. In this section, we systematically survey about 50 deep learning-based algorithms for scene classification of remote sensing images. In Fig. 6, we present some milestone works. That is one small step for deep learning theory, but one giant leap for the scene classification of remote sensing images [87]. From autoencoder, to CNNs, and then to GANs, deep learning algorithms constantly update scene classification records. To sum up, most of the deep learning-based scene classification algorithms can be broadly divided into three main categories: autoencoder-based methods, CNN-based methods, and GAN-based methods. In what follows, we discuss the three categories of methods at great length.

A. Autoencoder-Based Remote Sensing Image Scene Classification

1) Brief Introduction of Autoencoder: Autoencoder [67] is an unsupervised feature learning model, which consists of a sort of shallow and symmetrical neural network [see Fig. 7(a)]. An autoencoder consists of three layers: input layer, hidden layer, and output layer. It contains two units—encoder and decoder. The transformation from input layer to hidden layer is the process of encoding. The process of encoding can be formulated as (1), where $\boldsymbol{h} \in \mathbb{R}^n$ is the output of hidden layers, f denotes a nonlinear mapping, $\boldsymbol{W} \in \mathbb{R}^{n \times m}$ stands for the encoding weight matrix, $\boldsymbol{x} \in \mathbb{R}^m$ denotes the input of autoencoder, and $\boldsymbol{b} \in \mathbb{R}^n$ is the bias vector. Decoding is the inverse of encoding, which is the transformation from hidden layer to output layer, and can be formulated as (2), where $\tilde{\boldsymbol{x}} \in \mathbb{R}^m$ represents the reconstructed output, the decoding weight matrix is denoted by $\boldsymbol{W}' \in \mathbb{R}^{m \times n}$, and $\boldsymbol{b}' \in \mathbb{R}^m$ stands for the bias vector

$$\boldsymbol{h} = f(\boldsymbol{W} \cdot \boldsymbol{x} + \boldsymbol{b}) \tag{1}$$

$$\tilde{\boldsymbol{x}} = f\left(\boldsymbol{W}' \cdot \boldsymbol{h} + \boldsymbol{b}'\right). \tag{2}$$

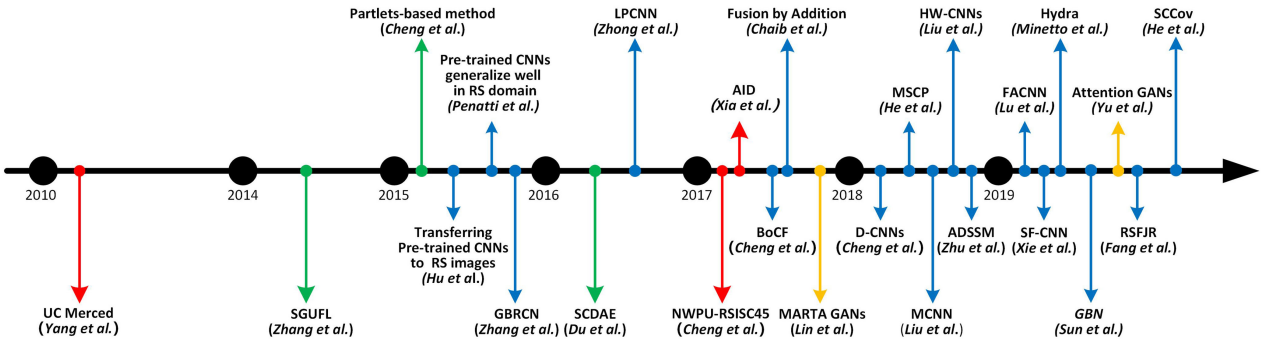


Fig. 6. Milestones of deep learning-based remote sensing image scene classification, including different deep learning-based methods and datasets. The red line represents typical datasets. The green, blue, and orange lines stand for autoencoder-based, CNN-based, and GAN-based remote sensing image scene classification, respectively.

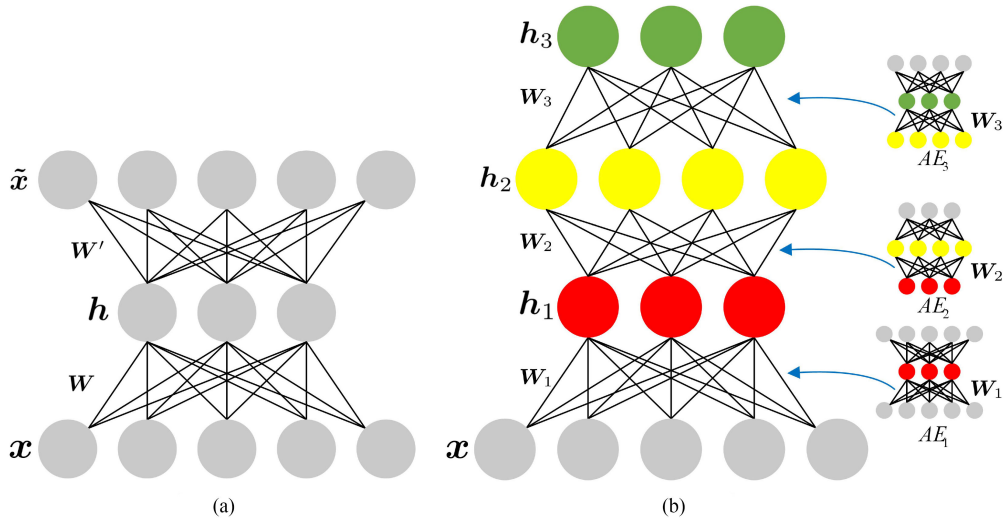


Fig. 7. Architectures of (a) autoencoder and (b) stacked autoencoder. The red, yellow, and green nodes stand for the hidden layers of autoencoders AE_1, AE_2, and AE_3, respectively. When stacking these autoencoders, the output of the hidden layer of the previous autoencoder is the input of the following autoencoder. For example, the output of the hidden layer of AE_1 is the input of AE_2, and the output of the hidden layer of AE_2 is the input of AE_3.

Autoencoder is able to compress high-dimensional features by minimizing the cost function that usually consists of a reconstruction error term and a regularization term. By using gradient descent with back propagation, autoencoder can learn the parameters of networks. In real applications, multilayer stacked autoencoders are used [see Fig. 7(b)] for feature learning. For example, three individual autoencoders AE_1, AE_2, and AE_3 are stacked together to form a stacked autoencoder, as shown in Fig. 7(b). When stacking these autoencoders, the output of the hidden layer of the previous autoencoder is the input of the following autoencoder. For example, the output of the hidden layer of AE_1 is the input of AE_2, and the output of the hidden layer of AE_2 is the input of AE_3. The key to training stacked autoencoders is how to initialize the network. The way of initializing the parameters of networks influences the network convergence especially the early layers, as well as the stability of training. Fortunately, Hinton *et al.* [67] provided a good solution to initialize the weight of the network by using restricted Boltzmann machines.

2) Autoencoder-Based Scene Classification Methods: Autoencoder is able to automatically learn mid-level visual representations from unlabeled data. The mid-level features play an important role in remote sensing image scene classification before deep learning takes off in the remote sensing community. Zhang *et al.* [88] introduced sparse autoencoder to scene classification. Cheng *et al.* [89] used the single-hidden-layer neural network and autoencoder for training more effective sparselets [90] to achieve efficient scene classification and object detection. In [91], Othman *et. al* proposed a remote sensing image scene classification algorithm relied on convolutional features and a sparse autoencoder. Han *et al.* [92] provided the scene classification methods based on hierarchical convolutional sparse autoencoder. Cheng *et al.* [93] demonstrated mid-level visual feature learned from autoencoder-based method is discriminative and able to facilitate scene classification tasks. In light of the limitation of feature representation of a single autoencoder, some researchers stacked multiple autoencoders together. Du *et al.* [94] came up with stacked convolutional denoising

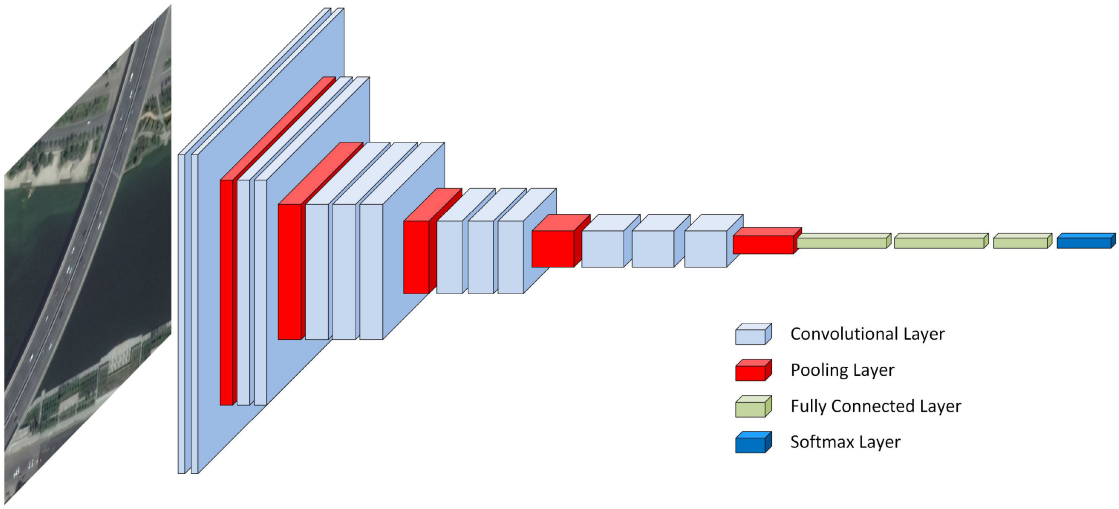


Fig. 8. Aarchitecture of CNNs.

autoencoder networks. After extensive experiments, their proposed framework showed superior classification performance. Yao *et al.* [95] integrated pairwise constraints into a stacked sparse autoencoder to learn more discriminative features for land-use scene classification and semantic annotation tasks.

The autoencoder and the algorithms derived from autoencoder are unsupervised-learning methods and have obtained good results in scene classification of remote sensing images. However, most of the above-mentioned autoencoder-based methods cannot learn the best discrimination features to distinguish different scene classes because they do not fully exploit scene class information.

B. CNN-Based Remote Sensing Image Scene Classification

1) Brief Introduction of CNN: CNNs have shown powerful feature learning ability in the visual domain. Since Krizhevsky and Hinton proposed the Alexnet [70] in 2012, a deep CNN that obtained the best accuracy in the LSVRC, there have appeared an array of advanced CNN models, such as VGGNet [96], GoogleNet [97], ResNet [98], DensNet [99], SENet [100], and SKNet [101]. CNNs are a kind of multilayer network with learning ability that consists of convolutional layers, pooling layers, and fully connected layers (see Fig. 8).

Convolutional layers: Convolutional layers play an important role on feature extraction from images. The convolutional layers input $\boldsymbol{X} \in \mathbb{R}^{n\times w\times h}$ consists of n 2-D feature maps of size $w \times h$. The output $\boldsymbol{H} \in \mathbb{R}^{m\times w'\times h'}$ of convolutional layers is m 2-D feature maps of size $w' \times h'$ via convolutional kernels $\boldsymbol{W}$. $\boldsymbol{W} \in \mathbb{R}^{m\times l\times l\times n}$ is m trainable filters of size $l \times l \times n$ (typically l =1, 3, or 5). The entire process of convolution is described as (3), where $*$ denotes 2-D convolution operation, additionally by using $\boldsymbol{b}$ to denote the m dimensional bias term. In general, a nonlinear activation function f is performed after the convolution operation. As the convolutional structure deepens, the convolutional layers can capture different level features (e.g., edges, lines, corners, structures, and shapes) from the input feature maps

$$\boldsymbol{H} = f(\boldsymbol{W} * \boldsymbol{X} + \boldsymbol{b}). \tag{3}$$

Pooling layers: Pooling layers are to execute a max or average operation over a small area of each input feature map, which can be defined as (4), where *pool* represents the pooling function (e.g., average pooling, max pooling, and stochastic pooling), $\boldsymbol{H}_{l-1}$ and $\boldsymbol{H}_l$ denotes the input and output of the pooling layer respectively. Usually, pooling layers are applied between two successive convolutional layers. Pooling operation can create invariance, such as small shifts and distortions. In the object detection and scene classification tasks, the characteristic of invariance provided by pooling layers is very important

$$\boldsymbol{H}_l = \text{pool}(\boldsymbol{H}_{l-1}). \tag{4}$$

Fully connected layers: Fully connected layers usually appear in the top layer of CNNs, which can summarize the features extracted from the bottom layers. Fully connected layers process its input $\tilde{\boldsymbol{X}}$ with linear transformation by weight $\tilde{\boldsymbol{W}}$ and bias $\tilde{\boldsymbol{b}}$, then map the output of linear transformation by a nonlinear activation function f. The entire process can be formulated as (5). In the task of classification, to output the probability of each class, a softmax classifier is connected to the last fully connected layer generally. The softmax classifier is used to normalize the fully connected layer output $\boldsymbol{y} \in \mathbb{R}^c$ (c is the number of classes) between 0 and 1, which can be described as (7), where e is the exponential function. The output of softmax classifier denotes the probability that the given input image belongs to each class. The dropout method [59] operates on the fully connected layers to avoid overfitting because a fully connected layer usually contains a large number of parameters

$$\boldsymbol{y} = f(\tilde{\boldsymbol{W}} \cdot \tilde{\boldsymbol{X}} + \tilde{\boldsymbol{b}}) \tag{5}$$

$$P\left(y_i\right) = \frac{e^{y_i}}{\sum_{i=1}^{c} e^{y_i}}. \tag{6}$$

2) CNN-Based Scene Classification Methods: In the wake of CNNs successfully being applied to large-scale visual classification tasks, around 2015, the use of CNNs has finally taken off in the remote sensing image analysis field [76], [77]. Compared with traditional advanced methods, e.g., SIFT [39], HOG [44], and BoVW [49], CNNs have the advantage of end-to-end feature learning. Meanwhile, it can extract high-level visual features that handcrafted feature-based methods cannot learn. By using different strategies of exploiting CNNs, a variety of CNN-based scene classification methods [73], [102]–[107] have emerged. Generally, the CNN-based methods of remote sensing image scene classification can be divided into three groups: using pretrained CNNs as feature extractors, fine-tuning pretrained CNNs on target datasets, and training CNNs from scratch.

Using pretrained CNNs as feature extractors: In the beginning, CNNs appeared as feature extractors. Penatti *et al.* [85] introduced CNNs in 2015 into remote sensing image scene classification, and evaluated the generalization capability of off-the-shelf CNNs in classification of remote sensing images. Their experiments show that CNNs can obtain better results than low-level descriptors. Later, Hu *et al.* [38] treated CNNs as feature extractors and investigated how to make full use of pretrained CNNs for scene classification. In [108], Marmanis *et al.* introduced a two-stage CNN scene classification framework. It used pretrained CNNs to derive a set of representations from images. The extracted representations were then fed into shallow CNN classifiers. Chaib *et al.* [109] fused the deep features extracted with VGGNet to enhance scene classification performance. In [110], Li *et al.* fused pretrained CNN features. The fused CNN features show better discrimination than raw CNN features in scene classification. Cheng *et al.* [104] designed the bag of convolutional features (BoCF) for remote sensing image scene classification by using off-the-shelf CNN features to replace traditional local descriptors such as SIFT. Yuan *et al.* [111] rearranged the local features extracted by an already trained VGG19Net for remote sensing image scene classification. In [112], He *et al.* proposed a novel multilayer stacked covariance pooling algorithm (MSCP) for remote sensing image scene classification. MSCP can combine multilayer feature maps extracted from pretrained CNN automatically. Lu *et al.* [113] introduced a feature aggregation CNN (FACNN) for scene classification. FACNN learns scene representations through exploring semantic label information. These methods all used pretrained CNNs as feature extractors and then fused or combined the features extracted by existing CNNs. It is worth noticing that the strategy of using off-the-shelf CNNs as feature extractors is simple and effective on small-scale datasets.

Fine-tuning pretrained CNNs: However, when the amount of training samples is not adequate to train a new CNN from scratch, fine-tuning an already trained CNNs on target datasets is a good choice. Castelluccio *et al.* [114] delved into the use of CNNs for remote sensing image scene classification by experimenting with three learning approaches: using pretrained CNNs as feature extractors, fine tuning, and training from scratch. And they concluded that fine-tuning gave better results than full training when the scale of datasets is small. This made researchers interested in fine-adjusting scene classification networks or optimizing its loss functions. Cheng *et al.* [73] designed a novel objective function for learning discriminative CNNs (D-CNNs). The D-CNNs shows better discriminability in scene classification. In [115], Liu *et al.* coupled CNN with a hierarchical Wasseratein loss function (HW-CNNs) to improve CNNs discriminatory ability. Minetto *et al.* [72] devised a new remote sensing image scene classification framework, named Hydra, which is an ensemble of CNNs and achieve the best results on the NWPU-RESISC45 dataset. Wang *et al.* [74] introduced attention mechanism into CNNs and designed the ARCNet (attention recurrent convolutional network) for scene classification. It is capable of highlighting key areas and discard noncritical information. In [116], to handle the problem of object scale variation in scene classification, Liu *et al.* formulated the multiscale CNN (MCNN). Fang *et al.* [117] designed a robust space-frequency joint representation (RSFJR) for scene classification by adding a frequency domain branch to CNNs. Because of fusing features from the space and frequency domains, the proposed method is able to provide more discriminative feature representations. Xie *et al.* [118] designed a scale-free CNN (SF-CNN) for the task of scene classification. SF-CNN can accept the images of arbitrary size as input without any resizing operation. Sun *et al.* [119] proposed a gated bidirectional network (GBN) for scene classification, which can get rid of the interference information and aggregate the interdependent information among different CNN layers. In the abovementioned methods, CNNs can learn discriminative features and obtain better performance by fine adjusting their structures, optimizing their objective function, or fine-tuning the modified CNNs on the target datasets.

Training CNNs from scratch: Even though fine-tuning pretrained CNNs can achieve remarkable performance, there exist some limitations relying on pretrained CNNs: learned features are not fully suitable for the characteristics of target datasets and it is inconvenient for researchers to modify pretrained CNNs. In [120], Chen *et al.* introduced knowledge distillation into scene classification to boost the performance of light CNNs. Zhang *et al.* [121] illustrated a lightweight and effective CNN that introduces the dilated convolution and channel attention into Mobilenetv2 [122] for scene classification. In addition, it is of considerable interest to design more effective and robust CNNs for scene classification. He *et al.* [123] introduced a novel skip-connected covariance (SCCov) network for remote sensing image scene classification. The SCCov is to add skip connection and covariance pooling to CNNs, which can reduce the amount of parameters and achieve better classification performance. In [102], Zhang *et al.* presented a gradient boosting random convolutional network (GBRCN) for scene classification via assembling different deep neural networks.

These CNN-based methods have obtained astonishing scene classification results. However, they generally require numerous annotated samples to fine-tune already trained CNNs or train a network from scratch.

C. GAN-Based Remote Sensing Image Scene Classification

1) Brief Introduction of GAN: Generative adversarial network (GAN) [82] is another important and promising machine

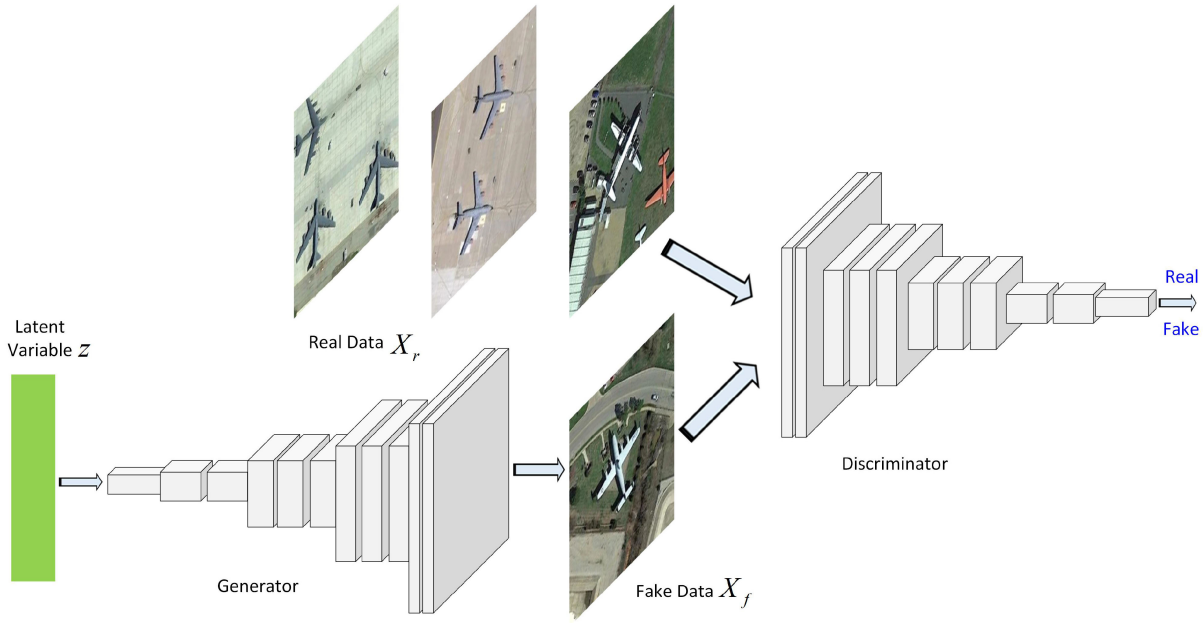


Fig. 9. Architecture of GANs.

learning method. As its name implies, GAN models the distribution of data via adversarial learning based on a minimax two-player game, and generates real-like data. GANs contain a pair of components—the discriminator D and generator G. As shown in Fig. 9, G can be analogs to a group of counterfeiters who take the role of generating fake currency, while D can be thought of as polices who determine whether the currency is made by G or bank. G and D constantly pit against each other in this game until D cannot distinguish between the counterfeit currency and genuine articles. GANs see the competition between G and D as the sole training criterion. G takes an input $\boldsymbol{z}$, which is a latent variable obeying a prior distribution $p_{\boldsymbol{z}}(\boldsymbol{z})$, then maps $\boldsymbol{z}$ with noise into data space by using a differential function $G(\boldsymbol{z};\boldsymbol{\theta}_g)$, where $\boldsymbol{\theta}_g$ denotes the generator G's parameters. D outputs the probability of the input data $\boldsymbol{x}$ that comes from real data rather than generator through a mapping $D(\boldsymbol{x};\boldsymbol{\theta}_d)$ with parameters $\boldsymbol{\theta}_d$, where $\boldsymbol{\theta}_d$ denotes the discriminator D's parameters. The entire process of the two-player minimax game is described as (7), where p_{data} is the distribution of data $\boldsymbol{x}$ and $\mathrm{V}(G, D)$ is an object function. From D's perspective, given an input data generated by G, D will play a role in minimizing its output. While if a sample is real data, D will maximize its output. This is the reason why the term $\log(1 - D(G(\boldsymbol{z})))$ is plugged into (7). Meanwhile, to fool D, G makes an effort to maximize D's output when a generated data is input to D. Thus, the relationship that D wants to maximize $\mathrm{V}(G, D)$ and G struggles to minimize $\mathrm{V}(G, D)$ is formed

$$\min_G \max_D V(G,D) = E_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + E_{\boldsymbol{z}\sim p_t(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \quad (7)$$

2) GAN-Based Scene Classification Methods: As a key method for unsupervised learning, since the introduction by Goodfellow *et al.* [82] in 2014, GANs have been gradually applied to many tasks such as an image to image translation, sample generation, image superresolution, and so on. Facing the tremendous volume of remote sensing images, CNN-based methods need to use massive labeled samples to train models. However, annotating samples is labor-intensive. Some researchers began to employ GANs to scene classification. In 2017, Lin *et al.* [84] proposed a multiple-layer feature-matching generative adversarial networks (MARTA GANs) for the task of scene classification. Duan *et al.* [83] used an adversarial net to assist in mining the inherent and discriminative features from remote sensing images. The dug features are able to enhance the classification accuracy. Bashmal *et al.* [132] provided a GAN-based method, called Siamese-GAN, to handle the aerial vehicle images classification problems under cross-domain conditions. In [133], to generate high-quality remote sensing images for scene classification, Xu *et al.* added the scaled exponential linear unites to GANs. Ma *et al.* [134] designed the Sifting-GAN, which can generate a large variety of authentic annotated samples for scene classification. Teng *et al.* [135] presented a classifier-constrained adversarial network for cross-domain semisupervised scene classification. Han *et al.* [136] introduced a generative framework, named SSGF, to scene classification. Yu *et al.* [137] devised an attention GAN for scene classification. Attention GAN achieves better scene classification performance by enhancing the representation power of the discriminator.

In the area of remote sensing scene image classification, most of GAN-based methods usually use GANs for sample generation or feature learning in an adversarial manner. Compared with CNN-based scene classification methods, only a small number of literatures about GAN-based scene classification method have been reported so far, and the performance of GAN-based scene classification is inferior to CNN-based methods. In addition,

TABLE II
13 PUBLICLY AVAILABLE DATA SETS FOR REMOTE SENSING IMAGE SCENE CLASSIFICATION

Data sets	Image number per class	Number of scene classes	Total image number	Image size	Training ratios		Data sources	Year
UC Merced [49]	100	21	2100	256 × 256	50%	80%	Aerial orthoimagery	2010
WHU-RS19 [124]	50∼61	19	1005	600×600	40%	60%	Google Earth	2012
RSSCN7 [125]	400	7	2800	400×400	20%	50%	Google Earth	2015
Brazilian Coffee Scene [85]	1438	2	2876	64×64	50%		SPOT sensor	2015
SAT-4/-6 [126]	125000/67500	4/6	500000/405000	28×28	80%		National Agriculture Imagery Program	2015
SIRI-WHU [127]	200	12	2400	200×200	50%		Google Earth	2016
RSC11 [128]	about 100	11	1232	512×512	50%		Google Earth	2016
AID [79]	220∼420	30	10000	600×600	20%	50%	Google Earth	2017
NWPU-RESISC45 [80]	700	45	31500	256×256	10%	20%	Google Earth	2017
RSI-CB128/-CB256 [129]	about 800/690	45/35	36000/24000	128×128/256×256	50%	80%	Google Earth & Bing Maps	2017
OPTIMAL-31 [74]	60	31	1860	256×256	80%		Google Earth	2018
EuroSAT [130]	2000∼3000	10	27000	64×64	80%		Sentinel-2	2019
BigEarthNet [131]	328∼217119	44	590326	120×120	60%		Sentinel-2	2019

most of GAN-based scene classification methods cannot be trained end-to-end because they often require labels for training an additional classifier. However, the powerful self-supervised feature learning capacity of GANs provides a promising future direction for the scene classification.

IV. SURVEY ON REMOTE SENSING IMAGE SCENE CLASSIFICATION BENCHMARKS

Datasets play an irreplaceable role on the advance of scene classification. Meanwhile, they are crucial for developing and evaluating various scene classification methods. As the number of high-resolution remote sensing sensors increases, the access to massive high-resolution remote sensing images makes it possible to build large-scale scene classification benchmarks. In the past few years, the researchers from different groups have proposed several publicly available high-resolution benchmark datasets for scene classification of remote sensing images [49], [74], [79], [80], [85], [124]–[131] to facilitate this field forward. Starting with the UC-Merced dataset [49], some representative datasets include WHU-RS19 [124], SAT-4&6 [126], RSSCN7 [125], Brazilian Coffee Scene [85], RSC11 [128], SIRI-WHU [127], RSCI-CB [129], AID [79], NWPU-RESISC45 [80], OPTIMAL-31 [74], EuroSAT [130], and BigEarthNet [131]. The characteristics of these 13 datasets are listed in Table II. Among them, the UC-Merced data [49], AID dataset [79], and NWPU-RESISC45 dataset [80] are three commonly used benchmark datasets, which will be introduced below in detail.

A. UC-Merced Dataset

The UC-Merced dataset[1] [49] was released in 2010 and contains 21 scene classes. Each category consists of 100 land-use images. In total, the dataset comprises 2100 scene images, of which the pixel resolution is 0.3 m. These images were obtained from United States Geological Survey National Map of 21 U.S. regions and fixed at 256×256 pixels. Fig. 10 lists the samples of each category from the dataset. Up to now, the dataset continues to be broadly employed for the scene classification. When conducting algorithm evaluation, two widely used training ratios are 50% and 80%, and the remaining 50% and 20% are used for testing.

B. AID Dataset

The AID [79] dataset[2] is a relatively large-scale dataset for the aerial scene classification. It was published in 2017 by Wuhan University and consists of 30 scene classes. Each scene class consists of 220 to 420 images, which were cropped from Google Earth imagery and fixed at 600×600 pixels. In total, the dataset comprises 10 000 scene images. Fig. 11 lists the samples of each category from the dataset. Different from the UC-Merced dataset, the AID dataset is multisourced because these aerial images were captured with different sensors. Moreover, the dataset is also multiresolution and the pixel resolution of each scene categories varies from about 8 m to about 0.5 m. When conducting the algorithm evaluation, two widely used training ratios are 20% and 50%, and the remaining 80% and 50% are used for testing.

C. NWPU-RESISC45 Dataset

To the best of our knowledge, the NWPU-RESISC45 dataset[3] [80], released by Northwest Polytechnical University, is currently the largest scene classification dataset. It consists of 45 scene categories. Each category consists of 700 images, which

[1]Online. [Available]: http://weegee.vision.ucmerced.edu/datasets/form.html

[2]Online. [Available]: www.lmars.whu.edu.cn/xia/AID-project.html

[3]Online. [Available]: http://www.escience.cn/people/gongcheng/NWPU-RESISC45.html

Fig. 10. Some example images from the UC-Merced dataset.

were obtained from Google Earth and fixed at 256×256 pixels. In total, the dataset comprises 31 500 scene images, which is chosen from more than 100 countries and regions. Apart from some specific classes with a lower spatial resolution (e.g., island, lake, mountain, and iceberg), the pixel resolution of most the scene categories varies from about 30 to 0.2 m. Fig. 12 lists the samples of each category from the dataset. The release of NWPU-RESISC45 dataset has allowed deep learning models to develop their full potential. When conducting algorithm evaluation, two widely used training ratios are 10% and 20%, and the remaining 90% and 80% are used for testing.

V. Performance Comparison and Discussion

A. Evaluation Criteria

There exist three commonly used criteria for evaluating the performance of the task of remote sensing image scene classification: overall accuracy (OA), average accuracy (AA), and confusion matrix. The metric of OA is an evaluation of the performance of the classifiers over the entire test dataset, which is formulated as the total number of accurately classified samples N_c divided by the total number of tested samples N_t, as described in (8). OA is a commonly used criterion for evaluating the performance of the methods for the scene classification of remote sensing images. The criterion of AA is defined as the sum of the accuracies of each category A_i divided by the total number of class c, as described in (9). When the sample number of each category is equal on the test set, OA and AA have the same value. The confusion matrix is a detailed classification result table about the performance of each single classifier. For each element x_{ij} in the table, the proportion of the images that are predicted to be the ith category while actually belonging to the jth class is computed. Therefore, the confusion matrix can directly visualize the performance of each category and through it we can easily get which classifiers are getting it right and what types of errors they are making. In this survey, we only use OA as evaluation criterion because the confusion matrix will take a lot of space

$$\mathrm{OA} = N_c/N_t \tag{8}$$

$$\mathrm{AA} = \frac{1}{c}\sum_{i=1}^{c} A_i. \tag{9}$$

B. Performance Comparison

In recent years, a variety of scene classification algorithms have been published. Here, 27 deep learning-based scene classification methods are selected for the performance comparison on three widely used benchmark datasets. Among the 27 deep

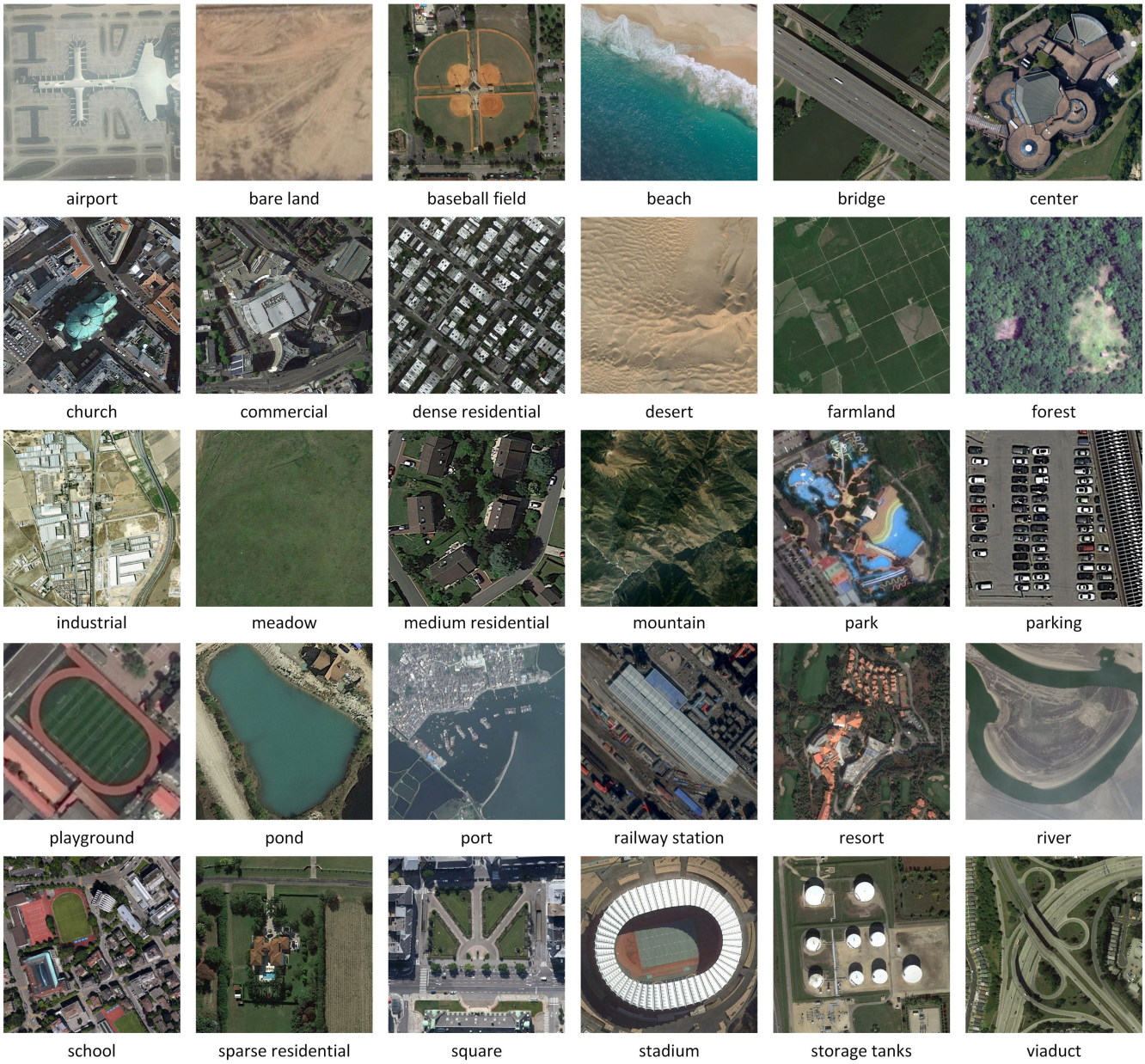


Fig. 11. Some example images from the AID dataset.

learning methods, 3 of them are autoencoder-based methods, 22 of them are CNN-based methods, and 2 of them are GAN-based methods.

Tables III–V report the classification accuracy comparison of deep learning-based scene classification methods on the UC-Merced dataset, the AID dataset, and the NWPU-RESISC45 dataset, respectively, measured in terms of OA.

C. Discussion

As can be seen from Tables III–V, the performance of remote sensing image scene classification has been successively advanced. In the early days, deep learning-based scene classification approaches were mainly based on autoencoder, and researchers usually use the UC-Merced dataset to evaluate autoencoder-based algorithms. As an early unsupervised deep learning method, the structure of autoencoder was relatively simple, so its feature learning capability was also limited. The accuracies of the autoencoder-based approaches had plateaued on the standard benchmarks.

Fortunately, after 2012, CNNs, a powerful supervised learning method, have proved to be capable of learning abstract features from raw images. Despite their powerful potential, it took some time for CNNs to take off in the remote sensing image scene classification domain, until 2015. A short while later, CNN-based algorithms mainly used CNNs as feature extractors, which outperformed autoencoder-based methods. However, only using CNNs as feature extractors did not make full use of the potential

TABLE III
OVERALL ACCURACY (%) COMPARISON OF 21 SCENE CLASSIFICATION METHODS ON THE UC-MERCED DATA SET

Method		Year	Publication	Training ratio	
				50%	80%
Autoencoder-based	SGUFL [88]	2014	IEEE TGRS	-	82.72±1.18
	partlets-based method [37]	2015	IEEE TGRS	88.76±0.79	-
	SCDAE [94]	2016	IEEE TCYB	-	93.7±1.3
CNN-based	GBRCN [102]	2015	IEEE TGRS	-	94.53
	LPCNN [103]	2016	JARS	-	89.90
	Fusion by Addition [109]	2017	IEEE TGRS	-	97.42±1.79
	ARCNet-VGG16 [74]	2018	IEEE TGRS	96.81±0.14	99.12±0.40
	MSCP [112]	2018	IEEE TGRS	-	98.36±0.58
	D-CNNs [73]	2018	IEEE TGRS	-	98.93±0.10
	MCNN [116]	2018	IEEE TGRS	-	96.66±0.9
	ADSSM [138]	2018	IEEE TGRS	-	99.76±0.24
	FACNN [113]	2019	IEEE TGRS	-	98.81±0.24
	SF-CNN [118]	2019	IEEE TGRS	-	99.05±0.27
	SCCov [123]	2019	IEEE TNNLS	-	99.05±0.25
	RSFJR [117]	2019	IEEE TGRS	97.21±0.65	-
	GBN [119]	2019	IEEE TGRS	97.05±0.19	98.57±0.48
	ADFF [139]	2019	Remote Sensing	96.05±0.56	97.53±0.63
	CNN-CapsNet [140]	2019	Remote Sensing	97.59±0.16	99.05±0.24
	Siamese ResNet50 [141]	2019	IEEE GRSL	90.95	94.29
GAN-based	MARTA GANs [84]	2017	IEEE GRSL	85.5±0.69	94.86±0.80
	Attention GANs [137]	2019	IEEE TGRS	89.06±0.50	97.69±0.69

TABLE IV
OVERALL ACCURACY (%) COMPARISON OF 16 SCENE CLASSIFICATION METHODS ON THE AID DATASET

Method		Year	Publication	Training ratio	
				20%	50%
CNN-based	Fusion by Addition [109]	2017	IEEE TGRS	-	91.87±0.36
	ARCNet-VGG16 [74]	2018	IEEE TGRS	88.75±0.40	93.10±0.55
	MSCP [112]	2018	IEEE TGRS	91.52±0.21	94.42±0.17
	D-CNNs [73]	2018	IEEE TGRS	90.82±0.16	96.89±0.10
	MCNN [116]	2018	IEEE TGRS	-	91.80±0.22
	HW-CNNs [115]	2018	IEEE TGRS	-	96.98±0.33
	FACNN [113]	2019	IEEE TGRS	-	95.45±0.11
	SF-CNN [118]	2019	IEEE TGRS	93.60±0.12	96.66±0.11
	SCCov [123]	2019	IEEE TNNLS	93.12±0.25	96.10±0.16
	CNNs-WD [142]	2019	IEEE GRSL	-	97.24±0.32
	RSFJR [117]	2019	IEEE TGRS	-	96.81±1.36
	GBN [119]	2019	IEEE TGRS	92.20±0.23	95.48±0.12
	ADFF [139]	2019	Remote Sensing	93.68±0.29	94.75±0.25
	CNN-CapsNet [140]	2019	Remote Sensing	93.79±0.13	96.32±0.12
GAN-based	MARTA GANs [84]	2017	IEEE GRSL	75.39±0.49	81.57±0.33
	Attention GANs [137]	2019	IEEE TGRS	78.95±0.23	84.52±0.18

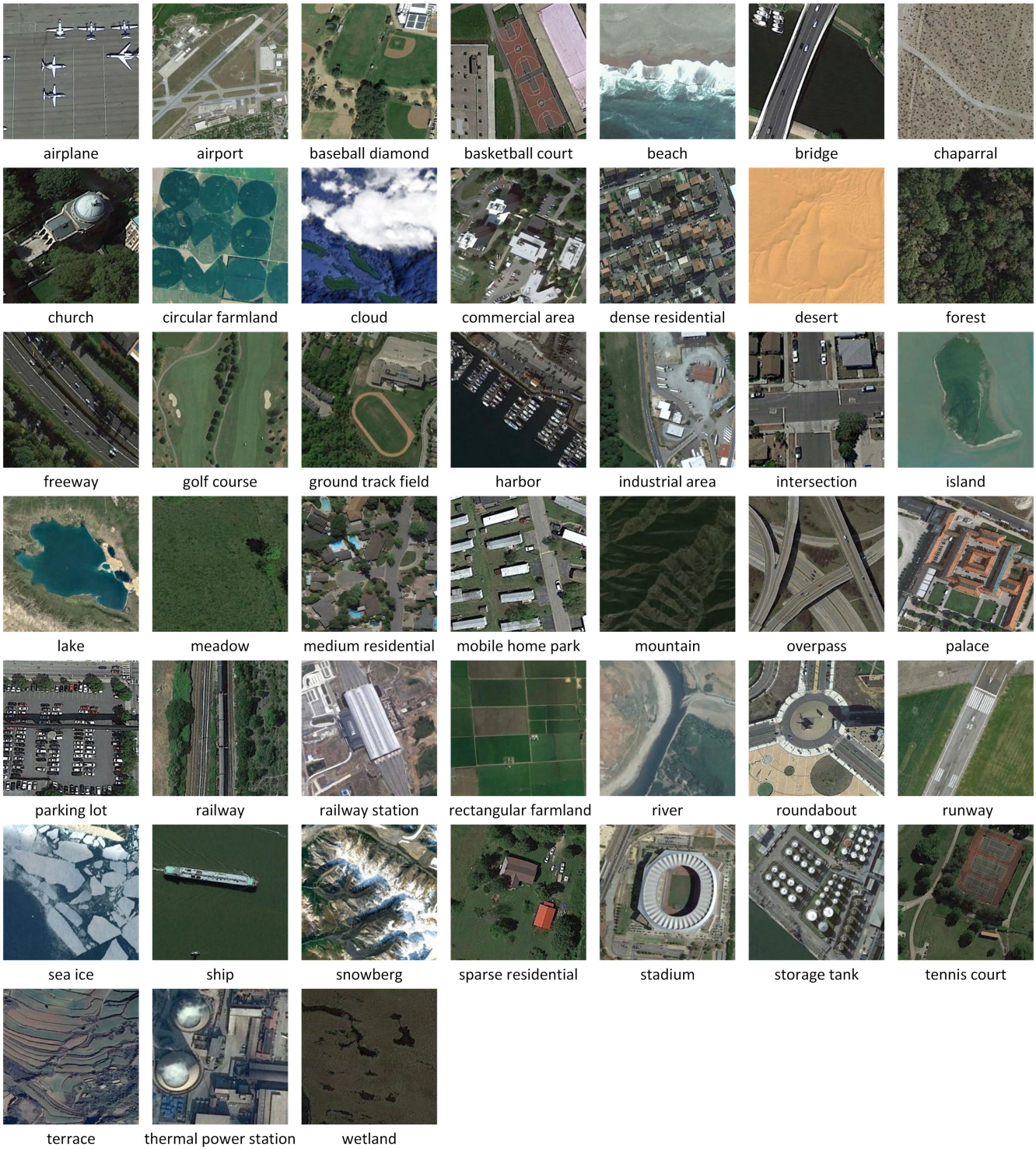


Fig. 12. Some example images from the NWPU-RESISC45 dataset.

of CNNs. Thanks to the release of two large-scale scene classification benchmarks, namely AID and NWPU-RESISC45 in 2017, fine-tuning off-the-shelf CNNs have shown better generalization ability in the task of scene classification than only using CNNs as feature extractors.

Generally, CNN-based methods require large-scale labeled remote sensing images to train CNNs. To deal with this issue, GANs, a novel self-supervised learning method, were introduced into remote sensing image scene classification. Through adversarial training, GANs can model the distribution of real samples and generate new samples. According to the reported accuracy of scene classification in Tables III–V, the development of autoencoder-based methods have reached a bottleneck, CNNs-based methods still dominate and have some upside

TABLE V
OVERALL ACCURACY (%) COMPARISON OF 15 SCENE CLASSIFICATION METHODS ON THE NWPU-RESISC45 DATA SET

Method		Year	Publication	Training ratio	
				10%	20%
CNN-based	BoCF [104]	2017	IEEE GRSL	82.65±0.31	84.32±0.17
	MSCP [112]	2018	IEEE TGRS	88.07±0.18	90.81±0.13
	D-CNNs [73]	2018	IEEE TGRS	89.22±0.50	91.89±0.22
	HW-CNNs [115]	2018	IEEE TGRS	-	94.38±0.17
	IORN [143]	2018	IEEE GRSL	87.83±0.16	91.30±0.17
	ADSSM [138]	2018	IEEE TGRS	91.69±0.22	94.29±0.14
	SF-CNN [118]	2019	IEEE TGRS	89.89±0.16	92.55±0.14
	ADFF [139]	2019	Remote Sensing	90.58±0.19	91.91±0.23
	CNN-CapsNet [140]	2019	Remote Sensing	89.03±0.21	89.03±0.21
	SCCov [123]	2019	IEEE TNNLS	89.30±0.35	92.10±0.25
	DNE [144]	2019	IEEE GRSL	-	96.01
	Hydra [72]	2019	IEEE TGRS	92.44±0.34	94.51±0.21
	Siamese ResNet50 [141]	2019	IEEE GRSL	-	92.28
GAN-based	MARTA GANs [84]	2017	IEEE GRSL	68.63±0.22	75.03±0.28
	Attention GANs [137]	2019	IEEE TGRS	72.21±0.21	77.99±0.19

potential, the performance of GAN-based methods is relatively low on the three benchmarks, and so there remains much room for further improving the performance of GAN-based methods.

Moreover, learning discriminative feature representation is one of the critical driving forces that improve the scene classification performance. Fusing multiple features [109], [117], designing effective cost functions [72], [115], modifying deep learning models [72], [118], and data augmentation [84] are all beneficial for attaining better performance. Meanwhile, with the access to large-scale benchmark data sets, it will become smaller for the gap between the scene classification approaches based on supervised learning and the scene classification approaches relied on unsupervised learning.

The release of publicly available benchmarks, such as the UC-Merced dataset, the AID dataset, and the NWPU-RESISC45 dataset, makes it easier to compare scene classification algorithms. From the perspectives of datasets, the UC-Merced dataset is relatively simple, and the results on the dataset driven by CNNs have reached saturation (above 99% classification accuracy by using the training ratios of 80%). The AID dataset is of moderate difficulty. The classification accuracy on the AID dataset can reach about 97% by using 50% training samples. For NWPU-RESISC45, some advanced methods based on CNNs have reached about 96% classification accuracy when the training ratio is fixed at 20%. Up to the present, the NWPU-RESISC45 dataset is still challenging compared with the UC-Merced dataset and the AID dataset.

The performance of CNN-based methods depends very much on the quantity of training data, so developing larger scale and more challenging remote sensing image scene classification benchmarks can further promote the development of data-driven algorithms.

VI. FUTURE OPPORTUNITIES

Scene classification is an important and challenging problem for remote sensing image interpretation. Driven by its wide application, it has aroused extensive research attention. Thanks to the advancement of deep learning techniques and the establishment of large-scale datasets for scene classification, scene classification has been seeing dramatic improvement. In spite of the amazing successes obtained in the past several years, there still exists a giant gap between the current understanding level of machines and human-level performance. Thus, there is still much work that needs to be done in the field of scene classification. By investigating the current scene classification algorithms and the available datasets, this article discusses several potential future directions for scene classification in remote sensing imagery.

1) Learning discriminative feature representations. Two key factors that influence the performance of scene classification tasks are intraclass diversity and interclass similarity existing in remote sensing images. To tackle the challenges, some representative methods [72], [73], [145] have been introduced over the past few years, such as multitask learning (e.g., unifying classification and similarity/metric learning) and designing/fusing CNNs. Even though these methods are effective to learn discriminative CNN features, the challenges of higher intraclass variation and smaller interclass separability are still not fully solved. These challenges seriously affect the performance of scene classification. In the future, learning more discriminative feature representations to handle the challenges needs to be addressed by various learning ways.
2) Learning multiscale features. In the task of remote sensing image scene classification, the same scene/object class can appear in different scales due to the changes in imaging

distance and the intrinsic properties of scenes/objects in size, so how to learn multiscale features has been a crucial and open problem. Some researches [116], [123], [146]–[149] in multiscale representations have been done over the past few decades, such as multiscale training, multiresolution feature fusion, and changing receptive field. However, these existing methods for learning scale-invariance features are far from the capability of human vision and cannot easily respond to the challenge of large variance of scene/object scale. For example, building deeper CNNs in order to extract high-level features has the side effect that small-sized object information is easily discarded. In the future, designing more robust way to extract multiscale features, especially for small-sized scenes/objects, would be promising for numerous vision tasks.

3) Multilabel remote sensing image scene classification. In the past few decades, extensive efforts have been made for the task of single-label image classification. However, in the real world, it is extremely common that multiple ground objects will appear in a remote sensing image because of the birds-eye imaging method. Therefore, single-label remote sensing image scene classification does not allow for a deep understanding of the intricate content of remote sensing images. In recent years, research has been conducted on multilabel remote sensing image scene classification [150]–[156], but it still faces many challenges that need to be further addressed, such as how to exploit the relationship between different labels, how to learn more generalized discriminative features, and how to build large-scale multilabel remote sensing image scene classification datasets.
4) Developing larger scale scene classification datasets. An ideal scene classification system would be capable of accurately and efficiently recognizing all scene types in all open world scenes. Recent scene classification methods are still trained with relatively limited datasets, so they are capable of classifying scene categories within the training datasets but blind, in principle, to other scene classes outside the datasets. Therefore, a compelling scene classification system should be able to accurately label a novel scene image with a semantic category. The existing datasets [49], [79], [80] contain dozens of scene classes, which are far fewer than those that humans can distinguish. Moreover, a common deep CNN has millions of parameters and it tends to over-fit the tens of thousands of training samples in the training set. Hence, fully training a deep classification model is almost impracticable by using currently available scene classification datasets. A majority of advanced scene classification algorithms mainly rely on fine-tuning already trained CNNs on the target datasets or utilizing pretrained CNNs as feature extractors. Although the transferring solutions behave fairly well on the target datasets with limited types and samples, they are not the most optimal solution compared with fully training a deep CNN model because the model trained from scratch is able to extract more specific features that are adaptable to the target domain when training samples is large enough. Considering this, developing a new large scale dataset with considerably more scene classes for scene classification is very promising.
5) Unsupervised learning for scene classification. Currently, the most advanced scene classification algorithms generally use fully supervised models learned from annotated data with semantic categories and have achieved amazing scene classification results. However, such fully supervised learning is extremely expensive and time-consuming to undertake because data annotation must be done manually by researchers with expert knowledge of the area of remote sensing image understanding. When the number of scene classes is huge, data annotation may become very difficult due to the massive amount of diversities and variations in remote sensing images. Meanwhile, the labeled data are generally full of noise and errors, especially for large-scale datasets, since the diverse knowledge levels of different specialists result in different understandings of the same classes of scene. Fully supervised learning can hardly work well without a large dataset with clean labels. As a promising unsupervised learning method, generative adversarial networks have been used for tackling scene classification with datasets that lack annotations [83], [84], [137]. Consequently, it is valuable to explore unsupervised learning for scene classification.
6) Compact and efficient scene classification models. During the past few years, another key factor in the outstanding progress in scene classification is the evolution of powerful deep CNNs. In order to achieve high accuracy in classification, the layer number of the CNNs has increased from several layers to hundreds of layers. Most advanced CNN models have millions of parameters and require a massive labeled dataset for training and high-performance GPUs, which severely limits the deploying of scene classification algorithms on airborne and satellite-borne embedded systems. In response, some researchers are working to design compact and lightweight scene classification models [120], [121]. In this area, there is much work to be done.
7) Scene classification with limited samples. CNNs have obtained huge success in the field of scene classification. However, most of those models demand large-scale labeled data and numerous iterations to train their parameter sets. This extremely limits their scalability to novel categories because of the high cost of labeling. Also, this fundamentally confines their applicability to rare scene categories (e.g., missile position, military zones), which are difficult to capture. In contrast, humans are adept at distinguishing scenes with little supervision learning, or none at all, such as few-shot [157] or zero-shot learning [158]. For instance, children can quickly and accurately recognize scene types ranging from a single image on TV, in a book, or hearing its description. The current best scene classification approaches are still far from achieving the human's ability to classify scene types with a few labeled samples. Exploring few-shot/zero-shot learning approach for scene classification [159]–[161] still needs to be further developed.

8) Cross-domain scene classification. Current research works have confirmed that CNNs are powerful tools for the task of scene classification and CNN-based methods have attained remarkable performance. However, the big achievements are based on the fact that training and testing data obey the same distribution. What will happen when train and test sets are from different domains? Can CNN models trained on a source domain show good generalization on another target domain? Generally, the performance will drop significantly because there exists a big gap between the source and target domains on data distribution. In fact, these differences between source and target domains are quite common on remote sensing images because of different imaging platforms (e.g., satellites and unmanned aerial vehicles) or different imaging sensors (optical sensors, infrared sensors, and SAR sensors). In the past few years, some researchers have explored cross-domain scene classification to enhance the generalization of CNN models and reduce the distribution gap between the target and source domains [162]–[165]. There is much potential for improving domain adaption-based methods for scene classification, such as mapping the feature representations from target and source domains onto a uniform space while preserving the original data structures, designing additional adaptation layers, and optimizing the loss functions.

VII. Conclusion

Scene classification of remote sensing images has obtained major improvements through several decades of development. The number of papers on remote sensing image scene classification is breathtaking, especially the literature about deep learning-based methods. By taking into account the rapid rate of progress in scene classification, in this article, we first discussed the main challenges that the current area of remote sensing image scene classification faces with. Then, we surveyed three kinds of deep learning-based methods in detail and introduced the mainstream scene classification benchmarks. Next, we summarized the performance of deep learning-based methods on three widely used datasets in tabular forms, and also provided the analysis of the results. Finally, we discussed a set of promising opportunities for further research.

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] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proc. IEEE*, vol. 103, no. 9, pp. 1560–1584, Sep. 2015.

[3] P. Gamba, "Human settlements: A global challenge for EO data processing and interpretation," *Proc. IEEE*, vol. 101, no. 3, pp. 570–581, Mar. 2013.

[4] D. Li, M. Wang, Z. Dong, X. Shen, and L. Shi, "Earth observation brain (EOB): An intelligent earth observation system," *Geo-spatial Inf. Sci.*, vol. 20, no. 2, pp. 134–140, 2017.

[5] N. Longbotham, C. Chaapel, L. Bleiler, C. Padwick, W. J. Emery, and F. Pacifici, "Very high resolution multiangle urban classification analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1155–1170, Apr. 2012.

[6] A. Tayyebi, B. C. Pijanowski, and A. H. Tayyebi, "An urban growth boundary model using neural networks, GIS and radial parameterization: An application to Tehran, Iran," *Landscape Urban Plan.*, vol. 100, no. 1/2, pp. 35–44, Mar. 2011.

[7] T. R. Martha, N. Kerle, C. J. van Westen, V. Jetten, and K. V. Kumar, "Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4928–4943, Dec. 2011.

[8] G. Cheng, L. Guo, T. Zhao, J. Han, H. Li, and J. Fang, "Automatic landslide detection from remote-sensing imagery using a scene classification method based on BOVW and PLSA," *Int. J. Remote Sens.*, vol. 34, no. 1/2, pp. 45–59, 2013.

[9] Z. Y. Lv, W. Shi, X. Zhang, and J. A. Benediktsson, "Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1520–1532, May 2018.

[10] X. Huang, D. Wen, J. Li, and R. Qin, "Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery," *Remote Sens. Environ.*, vol. 196, pp. 56–75, 2017.

[11] T. Zhang and X. Huang, "Monitoring of urban impervious surfaces using time series of high-resolution remote sensing images in rapidly urbanized areas: A case study of Shenzhen," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2692–2708, Aug. 2018.

[12] F. Ghazouani, I. R. Farah, and B. Solaiman, "A multi-level semantic scene interpretation strategy for change interpretation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8775–8795, Nov. 2019.

[13] X. Li and G. Shao, "Object-based urban vegetation mapping with high-resolution aerial photography as a single data source," *Int. J. Remote Sens.*, vol. 34, no. 3, pp. 771–789, 2013.

[14] N. B. Mishra and K. A. Crews, "Mapping vegetation morphology types in a dry savanna ecosystem: Integrating hierarchical object-based image analysis with random forest," *Int. J. Remote Sens.*, vol. 35, no. 3, pp. 1175–1198, 2014.

[15] 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.

[16] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 146, pp. 182–196, 2018.

[17] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019.

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

[19] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.

[20] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.

[21] G. Cheng, P. Zhou, and J. Han, "RIFD-CNN: Rotation-invariant and fisher discriminative convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2884–2893.

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

[23] M. Ji and J. R. Jensen, "Effectiveness of subpixel analysis in detecting and quantifying urban imperviousness from Landsat thematic mapper imagery," *Geocarto Int.*, vol. 14, no. 4, pp. 33–41, 1999.

[24] 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.

[25] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "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.

[26] L. L. Janssen and H. Middelkoop, "Knowledge-based crop classification of a Landsat thematic mapper image," *Int. J. Remote Sens.*, vol. 13, no. 15, pp. 2827–2837, 1992.

[27] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

[28] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[29] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[30] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.

[31] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, Jul. 2019.

[32] T. Blaschke and J. Strobl, "What's wrong with pixels? Some recent developments interfacing remote sensing and gis," *Zeitschrift für Geoinformationssysteme*, vol. 14, no. 6, pp. 12–17, 2001.

[33] T. Blaschke, "Object-based contextual image classification built on image segmentation," in *Proc. IEEE Workshop Adv. Techn. Anal. Remotely Sensed Data*, 2013, pp. 113–119.

[34] G. Yan, J.-F. Mas, B. Maathuis, Z. Xiangmin, and P. 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.

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

[36] T. Blaschke, S. Lang, and G. Hay, *Object-Based Image Analysis: Spatial Concepts for Knowledge-driven Remote Sensing Applications*. Berlin, Germany: Springer Science, 2008.

[37] 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.

[38] 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.

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

[40] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[41] A. K. Jain, N. K. Ratha, and S. Lakshmanan, "Object detection using gabor filters," *Pattern Recognit.*, vol. 30, no. 2, pp. 295–309, 1997.

[42] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[43] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[44] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, vol. 1, pp. 886–893, 2005.

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

[46] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 143–156.

[47] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, 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. 2011.

[48] 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. Vision Pattern Recognit.*, vol. 2, pp. 2169–2178, 2006.

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

[50] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. Int. Conf. Comput. Vision*, 2011, pp. 1465–1472.

[51] W. Shao, W. Yang, G.-S. Xia, and G. Liu, "A hierarchical scheme of multiple feature fusion for high-resolution satellite scene categorization," in *Proc. Int. Conf. Comput. Vision Syst.*, 2013, pp. 324–333.

[52] 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*, 2014, pp. 1–5.

[53] 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 Observ. Remote Sens.*, vol. 7, no. 12, pp. 4620–4631, Dec. 2014.

[54] Y. Zhang, X. Sun, H. Wang, and K. Fu, "High-resolution remote-sensing image classification via an approximate earth mover's distance-based bag-of-features model," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1055–1059, Sep. 2013.

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

[56] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, 1987.

[57] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Res.*, vol. 37, no. 23, pp. 3311–3325, 1997.

[58] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2013.

[59] M. L. Mekhalfi, F. Melgani, Y. Bazi, and N. Alajlan, "Land-use classification with compressive sensing multifeature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2155–2159, Oct. 2015.

[60] V. Risojević and Z. Babić, "Unsupervised quaternion feature learning for remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 4, pp. 1521–1531, Apr. 2016.

[61] 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.

[62] 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.

[63] 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.

[64] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[65] J. Fan, T. Chen, and S. Lu, "Unsupervised feature learning for land-use scene recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2250–2261, Apr. 2017.

[66] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

[67] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Sci.*, vol. 313, no. 5786, pp. 504–507, 2006.

[68] 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.

[69] 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, pp. 3371–3408, 2010.

[70] 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.

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

[72] R. Minetto, M. P. Segundo, and S. Sarkar, "Hydra: An ensemble of convolutional neural networks for geospatial land classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6530–6541, Sep. 2019.

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

[74] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[75] M. Li, S. Zang, B. Zhang, S. Li, and C. Wu, "A review of remote sensing image classification techniques: The role of spatio-contextual information," *Eur. J. Remote Sens.*, vol. 47, no. 1, pp. 389–411, 2014.

[76] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state-of-the-art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[77] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[78] U. Maulik and D. Chakraborty, "Remote sensing image classification: A survey of support-vector-machine-based advanced techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 33–52, Mar. 2017.

[79] G.-S. Xia *et al.*, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

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

[81] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.

[82] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[83] Y. Duan, X. Tao, M. Xu, C. Han, and J. Lu, "GAN-NL: Unsupervised representation learning for remote sensing image classification," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2018, pp. 375–379.

[84] D. Lin, K. Fu, Y. Wang, G. Xu, and X. Sun, "Marta GANs: Unsupervised representation learning for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2092–2096, Nov. 2017.

[85] O. A. 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. Vision Pattern Recognit. Workshops*, 2015, pp. 44–51.

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

[87] B. Zhang *et al.*, "Remotely sensed big data: Evolution in model development for information extraction [point of view]," *Proc. IEEE*, vol. 107, no. 12, pp. 2294–2301, Dec. 2019.

[88] 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. 2014.

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

[90] R. Girshick, H. O. Song, and T. Darrell, "Discriminatively activated sparselets," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 196–204.

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

[92] X. Han, Y. Zhong, B. Zhao, and L. Zhang, "Scene classification based on a hierarchical convolutional sparse auto-encoder for high spatial resolution imagery," *Int. J. Remote Sens.*, vol. 38, no. 2, pp. 514–536, 2017.

[93] G. Cheng, P. Zhou, J. Han, L. Guo, and J. Han, "Auto-encoder-based shared mid-level visual dictionary learning for scene classification using very high resolution remote sensing images," *IET Comput. Vision*, vol. 9, no. 5, pp. 639–647, 2015.

[94] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1017–1027, Apr. 2017.

[95] 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.

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

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

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

[99] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4700–4708.

[100] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7132–7141.

[101] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 510–519.

[102] 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. 2015.

[103] Y. Zhong, F. Fei, and L. Zhang, "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery," *J. Appl. Remote Sens.*, vol. 10, no. 2, 2016, Art. no. 025006.

[104] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.

[105] Y. Yu, Z. Gong, C. Wang, and P. Zhong, "An unsupervised convolutional feature fusion network for deep representation of remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 1, pp. 23–27, Jan. 2018.

[106] Y. Liu, Y. Zhong, F. Fei, Q. Zhu, and Q. Qin, "Scene classification based on a deep random-scale stretched convolutional neural network," *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 444.

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

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

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

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

[111] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Remote sensing image scene classification using rearranged local features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1779–1792, Mar. 2019.

[112] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Dec. 2018.

[113] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.

[114] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*.

[115] Y. Liu, C. Y. Suen, Y. Liu, and L. Ding, "Scene classification using hierarchical Wasserstein CNN," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 5, pp. 2494–2509, May 2019.

[116] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7109–7121, Dec. 2018.

[117] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Robust space–frequency joint representation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7492–7502, Oct. 2019.

[118] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.

[119] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.

[120] G. Chen *et al.*, "Training small networks for scene classification of remote sensing images via knowledge distillation," *Remote Sens.*, vol. 10, no. 5, 2018, Art. no. 719.

[121] B. Zhang, Y. Zhang, and S. Wang, "A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2636–2653, Aug. 2019.

[122] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4510–4520.

[123] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1461–1474, May 2020.

[124] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, "Structural high-resolution satellite image indexing," in *ISPRS TC VII Symposium-100 Years ISPRS*, 2010, pp. 298–303.

[125] 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.

[126] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "DeepSat: A learning framework for satellite imagery," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2015, pp. 1–10.
[127] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.
[128] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote Sens.*, vol. 10, no. 3, 2016, Art. no. 035004.
[129] H. Li, C. Tao, Z. Wu, J. Chen, J. Gong, and M. Deng, "RSCI-CB: A large scale remote sensing image classification benchmark via crowdsource data," 2017, *arXiv:1705.10450*.
[130] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, Jul. 2019.
[131] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5901–5904.
[132] L. Bashmal, Y. Bazi, H. AlHichri, M. M. AlRahhal, N. Ammour, and N. Alajlan, "Siamese-GAN: Learning invariant representations for aerial vehicle image categorization," *Remote Sens.*, vol. 10, no. 2, 2018, Art. no. 351.
[133] S. Xu, X. Mu, D. Chai, and X. Zhang, "Remote sensing image scene classification based on generative adversarial networks," *Remote Sens. Lett.*, vol. 9, no. 7, pp. 617–626, 2018.
[134] D. Ma, P. Tang, and L. Zhao, "SiftingGAN: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 7, pp. 1046–1050, Jul. 2019.
[135] W. Teng, N. Wang, H. Shi, Y. Liu, and J. Wang, "Classifier-constrained deep adversarial domain adaptation for cross-domain semisupervised classification in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 5, pp. 789–793, May 2019.
[136] W. Han, R. Feng, L. Wang, and Y. Cheng, "A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 23–43, 2018.
[137] Y. Yu, X. Li, and F. Liu, "Attention GANs: Unsupervised deep feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 519–531, Jan. 2020.
[138] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 6180–6195, Oct. 2018.
[139] R. Zhu, L. Yan, N. Mo, and Y. Liu, "Attention-based deep feature fusion for the scene classification of high-resolution remote sensing images," *Remote Sens.*, vol. 11, no. 17, 2019, Art. no. 1996.
[140] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-CapsNet," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 494.
[141] X. Liu, Y. Zhou, J. Zhao, R. Yao, B. Liu, and Y. Zheng, "Siamese convolutional neural networks for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1200–1204, Aug. 2019.
[142] Y. Liu, Y. Liu, and L. Ding, "Scene classification by coupling convolutional neural networks with Wasserstein distance," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 722–726, May 2019.
[143] J. Wang, W. Liu, L. Ma, H. Chen, and L. Chen, "Iorn: An effective remote sensing image scene classification framework," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1695–1699, Nov. 2018.
[144] M. A. Dede, E. Aptoula, and Y. Genc, "Deep network ensembles for aerial scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 732–735, May 2019.
[145] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.
[146] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2019.2938758.
[147] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2117–2125.
[148] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2403–2412.
[149] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, "Cross-scale feature fusion for object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, doi: 10.1109/LGRS.2020.2975541.
[150] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 149, pp. 188–199, 2019.
[151] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.
[152] R. Stivaktakis, G. Tsagkatakis, and P. Tsakalides, "Deep learning for multilabel land cover scene categorization using data augmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 7, pp. 1031–1035, Jul. 2019.
[153] Y. Hua, L. Mou, and X. X. Zhu, "Relation network for multilabel aerial image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.
[154] N. Khan, U. Chaudhuri, B. Banerjee, and S. Chaudhuri, "Graph convolutional network for multi-label VHR remote sensing scene recognition," *Neurocomputing*, vol. 357, pp. 36–46, 2019.
[155] B. T. Zegeye and B. Demir, "A novel active learning technique for multilabel remote sensing image scene classification," in *Proc. Image Signal Process. Remote Sens. XXIV*, vol. 10789, 2018, Art. no. 107890B.
[156] G. Cheng, D. Gao, Y. Liu, and J. Han, "Multi-scale and discriminative part detectors based features for multi-label image classification," *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 649–655.
[157] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1199–1208.
[158] M. Ye and Y. Guo, "Zero-shot classification with discriminative semantic representation learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 7140–7148.
[159] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, vol. 2, 2015, pp. 1–8.
[160] M. Zhai, H. Liu, and F. Sun, "Lifelong learning for scene recognition in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1472–1476, Sep. 2019.
[161] A. Li, Z. Lu, L. Wang, T. Xiang, and J.-R. Wen, "Zero-shot scene classification for high spatial resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 4157–4167, Jul. 2017.
[162] N. Ammour, L. Bashmal, Y. Bazi, M. M. Al Rahhal, and M. Zuair, "Asymmetric adaptation of deep features for cross-domain classification in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 4, pp. 597–601, Apr. 2018.
[163] E. Othman, Y. Bazi, F. Melgani, H. Alhichri, N. Alajlan, and M. Zuair, "Domain adaptation network for cross-scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4441–4456, Aug. 2017.
[164] X. Lu, T. Gong, and X. Zheng, "Multisource compensation network for remote sensing cross-domain scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2504–2515, Apr. 2020.
[165] S. Song, H. Yu, Z. Miao, Q. Zhang, Y. Lin, and S. Wang, "Domain adaptation for convolutional neural networks-based remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1324–1328, Aug. 2019.

Gong Cheng (Member, IEEE) received the B.S. degree in biomedical engineering from Xidian University, Xi'an, China, in 2007, and the M.S., and Ph.D. degrees in pattern recognition and intelligent systems from Northwestern Polytechnical University, Xi'an, China, in 2010 and 2013, respectively.

He is currently a Professor with Northwestern Polytechnical University, Xi'an, China. His main research interests include computer vision, pattern recognition, and remote sensing image understanding.

Dr. Cheng is an Associate Editor of IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE and a Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

Xingxing Xie received the B.S. degree in automation from Inner Mongolia University, Huhhot, China, in 2015, and the M.S. degree in pattern recognition and intelligent systems from Northwestern Polytechnical University, Xi'an, China, in 2018. He is currently working toward the doctoral degree with Northwestern Polytechnical University.

His main research interests include computer vision and pattern recognition.

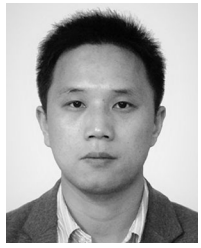
Junwei Han (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in pattern recognition and intelligent systems in 1999, 2001, and 2003, respectively, all from Northwestern Polytechnical University, Xi'an, China.

He is currently a Professor with Northwestern Polytechnical University. He was a Research Fellow with Nanyang Technological University, The Chinese University of Hong Kong, Dublin City University, and the University of Dundee from 2003 to 2010. His research interests include computer vision and brain-imaging analysis.

Dr. Han is an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON HUMAN–MACHINE SYSTEMS, NEUROCOMPUTING, AND MACHINE VISION AND APPLICATIONS.

Lei Guo received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 1982 and 1986, respectively, and the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 1993.

He is currently a Professor with the School of Automation, Northwestern Polytechnical University, Xi'an, China. His research interest focuses on image processing.



Gui-Song Xia (Senior Member, IEEE) received the Ph.D. degree in image processing and computer vision from CNRS LTCI, Télécom ParisTech, Paris, France, in 2011.

From 2011 to 2012, he has been a Postdoctoral Researcher with the Centre de Recherche en Mathématiques de la Decision, CNRS, Paris-Dauphine University, Paris, for one and a half years. He is currently a Full Professor of Computer Vision and Photogrammetry with Wuhan University. He has also been working as Visiting Scholar at DMA, École Normale Supérieure for two months, in 2018. His current research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing imaging.

Dr. Xia serves on the Editorial Boards of the *Pattern Recognition, Signal Processing: Image Communications*, and *EURASIP Journal on Image and Video Processing*.