NaSC-TG2: Natural Scene Classification With Tiangong-2 Remotely Sensed Imagery

Zhuang Zhou, Shengyang Li, Wei Wu, Weilong Guo, Xuan Li,
Guisong Xia, *Senior Member, IEEE*, and Zifei Zhao

***Abstract*—Scene classification is one of the most important applications of remote sensing. Researchers have proposed various datasets and innovative methods for remote sensing scene classification in recent years. However, most of the existing remote sensing scene datasets are collected uniquely from a single data source: Google Earth. In addition, scenes in different datasets are mainly human-made landscapes with high similarity. The lack of richness and diversity of data sources limits the research and applications of remote sensing classification. This article describes a large-scale dataset named "NaSC-TG2," which is a novel benchmark dataset for remote sensing natural scene classification built from Tiangong-2 remotely sensed imagery. The goal of this dataset is to expand and enrich the annotation data for advancing remote sensing classification algorithms, especially for the natural scene classification. The dataset contains 20 000 images, which are equally divided into ten scene classes. The dataset has three primary advantages: 1) it is large scale, especially in terms of the number of each class, and the numbers of scenes are evenly distributed; 2) it has a large number of intraclass differences and high interclass similarity, because all images are carefully selected from different regions and seasons; and 3) it offers natural scenes with novel spatial scale and imaging performance compared with other datasets. All images are acquired from the new generation of wideband imaging spectrometer of Tiangong-2. In addition to RGB images, the corresponding multispectral scene images are also provided. This dataset is useful in supporting the development and evaluation of classification algorithms, as demonstrated in the present study.**

***Index Terms*—Benchmark dataset, deep learning, remote sensing, scene classification, Tiangong-2.**

I. Introduction

Remote sensing technology enables us to measure and understand earth systems using the geometric and physical information in remote sensing images [1]–[3]. Understanding the semantic content of the images from a vast accumulation of remote sensing data is particularly important for practical applications [4], [5]. In this article, we focus on remote sensing scene classification, which is the process of understanding the semantic content based on spatial distribution and structural pattern information of the image and automatically annotating the category to which the image belongs [6].

The classification of remote sensing images is primarily performed at the pixel level [7], [22]. However, there are many drawbacks to pixel-level remote sensing classification. First, with the significant increase in the amount of remote sensing data and the improvement of spatial resolution of images [23]–[25], it is not feasible to interpret remote sensing images pixel by pixel [26], [27]. In addition, single pixel also lacks thematic meanings. Blaschke *et al.* [28] analyzed the disadvantages of pixel-based classification and pointed out that it was more effective to use the object as the smallest unit for remote sensing classification, with "object" referring to a local area of pixels that share the uniformity of spectrum or texture, e.g., super-pixels [29], [30]. For decades, object-oriented methods have been dominant in the classification of high-resolution remote sensing images [31], [35], [36]. It is worth noting that pixel-level as well as object-level classification methods complete the modeling of remote sensing scenes in a bottom-up manner, and train a robust classifier by aggregating spectral, geometrical features, and texture [36].

However, remote sensing scenes often contain different thematic classes [6], especially for high-resolution images. Therefore, it is useful to reveal the context of different thematic classes, such as the semantic information of remote sensing scenes [37], [38]. The main purpose of remote sensing scene classification is to model the spatial distribution and structural pattern of the image to divide the remote sensing images into different semantic categories [39]. Unlike pixel-level and object-oriented classification, scene classification enables a better understanding of remote sensing images [38], [41]. The concept of "scene" usually refers to a certain area in the remote sensing image, reflecting the clear semantic information of the type of surface features [4]–[17], [42], [46].

Despite the recent encouraging progress in remote sensing scene classification [15]–[17], [23]–[26], [37]–[41], [51], [63],

Manuscript received December 14, 2020; revised February 7, 2021; accepted February 18, 2021. Date of publication March 2, 2021; date of current version March 25, 2021. This work was supported in part by the Director's Foundation of Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, under Grant CSU-QZKT-2019-07 and Grant CSU-JJKT-2020-9. *(Corresponding author: Shengyang Li.)*

Zhuang Zhou, Shengyang Li, and Zifei Zhao are with the Technology and Engineering Center for Space Utilization, the Key Laboratory of Space Utilization and the University of Chinese Academy of Sciences, Beijing 100094, China (e-mail: zhouzhuang@csu.ac.cn; shyli@csu.ac.cn; zhaozifei18@csu.ac.cn).

Wei Wu and Xuan Li are with the Technology and Engineering Center for Space Utilization, and the Key Laboratory of Space Utilization, Chinese Academy of Sciences, Beijing 100049, China (e-mail: wuwei@csu.ac.cn; lixuan@csu.ac.cn).

Weilong Guo is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: guoweilong19@csu.ac.cn).

Guisong Xia is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, and the School of Computer Science, Wuhan University, Wuhan 430079, China (e-mail: guisong.xia@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3063096

most existing remote sensing scene datasets still need to be significantly expanded in terms of scale and diversity to a similar level of the ImageNet dataset, which contains tens of millions of labeled images [64].

The limited amount of remote sensing scene data is insufficient for developing and validating the data-driven algorithms represented by deep learning [65]. In addition, most of the existing datasets are collected from Google Earth (e.g., WHU-RS19 [66], AID [67], and NWUP-RESISC45 [68] datasets). Having a relatively single data source and same spatial scale makes the datasets limited to repetitive scenes dominated by human-made landscapes. At present, the dataset constructed for natural scenes has not been available in remote sensing communities, while the surface land covers are mainly natural landscapes. Therefore, natural scene dataset has a more significant application value for remote sensing classification.

In light of all these, in Section II, this article first reviews remote sensing scene classification. It then presents the current representative scene datasets, proposes a natural scene classification benchmark dataset with Tiangong-2 remotely sensed imagery in Section III, and conducts experiments in which different classification methods on the dataset are evaluated, which are presented in Section IV. Finally, in Section V, some concluding remarks are presented based on our primary work.

II. Overview of Remote Sensing Image Scene Classification Methods

The core technology of remote sensing scene classification is how to extract practical features of the image. In terms of feature extraction, there are the following three types of remote sensing scene classification methods—methods based on handcrafted features, feature encoding methods, and deep learning methods [67], [68]. It is worth noting that the three ways are not necessarily independent of one another.

A. Methods Based on Handcrafted Features

In the early stage of the development of remote sensing scene classification technology, researchers designed a series of handcrafted features based on engineering skills according to the characteristics of images and the task of classification [23], [43], [44], [53]. These features represent different characteristics of the scene in terms of color, texture, shape, spatial, and spectral information [69], [77]. The representative handcrafted features used in remote sensing scene classification included color histograms (CH) [73], local binary patterns (LBP) [78], [81], scale invariant feature transform (SIFT) [82], and histogram of oriented gradients (HOG) [83].

1) *CH:* The global CH feature [84] is not only simple to calculate but is also unaffected by image rotation and translation. In terms of image retrieval and scene classification, CH is one of the widely used features, which is mainly due to its insensitivity to orientation changes and image size [60], [76], [78]. However, the color feature cannot represent the local feature of the image or reflect the information of the color spatial distribution.
2) *LBP:* LBP [85], [88] is a kind of operator widely used to describe the local texture features of remote sensing images [53], [60], [74], [75], [79], [80]. It has the advantages of greyscale and rotation invariance. LBP constructs a measure of the relationship between each pixel of the image and its surrounding pixels, and extracts the texture feature of the images.
3) *SIFT:* SIFT feature describes the subregions of an image by identifying gradient information around key points [89]. The processing flow of the SIFT is first to construct a scale space to detect extreme points, then filter the searched extreme points, and finally, provide the description of image features. As a very stable local feature of images, SIFT is invariant to scaling, rotation, and brightness changes [82].
4) *HOG:* HOG is another one of the widely used handcrafted features in image processing [59]. It is obtained by statistical calculation of histogram of gradient direction in the local region of the image. HOG is used as a feature descriptor for computer vision tasks such as object detection [81]. In the field of remote sensing, this handcrafted feature is also widely used in scene classification [2], [3] [15], [90].

As global features, the CHs and LBP features represent the overall statistical characteristics of remote sensing scene images from the perspective of color [44], [78], and texture [82], [83], [85], respectively. The SIFT and HOG features are local structure [89] and shape information [85]. The handcrafted features are usually low-density features that contain a good deal of redundant information, and it is difficult to achieve optimum performance by classifying them directly. To optimize the classification performance, it is usually necessary to include more robust features further developed based on handcrafted features, such as improving the sparsity, rotation, and scale invariance of these handcrafted features [91], [93].

B. Feature Encoding Methods

The objective of feature encoding methods is to develop statistical patterns of higher order by encoding handcrafted features such as CHs, LBP, and SIFT, aiming to extract more significant features of the remote sensing scene and establish a global representation of the image.

1) *Bag-of-visual-words (BoVW):* The BoVW model is one of the most widely used feature encoding methods [53]. In this method, the local feature vector (such as SIFT) of the image is extracted first. Then, the representative vectors in the feature vectors are selected as words to form a visual dictionary. Subsequently, visual word statistics are obtained on the image to judge whether the similarity between the local area of the image and a word exceeds a certain threshold. In this way, the image can be represented as the distribution of words, which completes the image's representation [94]. Given its simplicity and efficiency, the BoVW model and its variants have been widely used in remote sensing scene classification [95], [98].
2) *Probabilistic topic models (PTM):* The PTM introduces an implicit variable based on the BOVW model to represent

the image as the probability distribution of the topic, and to increase the semantic information of the feature. Representative PTMs mainly include probabilistic latent semantic analysis (pLSA) [96] and latent Dirichlet allocation (LDA) [97]. The former uses a graph model to represent the relationship between topics, images, and visual words. It combines probability and statistical theory on the basis of the BoVW model to represent the topic probability distribution of remote sensing image, and then, realizes the scene classification. The latter defines a function for the original topic probability by treating the topic mixing parameters as variables that obey Dirichlet distribution to solve overfitting [98].

C. Methods Based on Deep Learning

In recent years, artificial intelligence (AI) technology represented by deep learning has achieved great success in computer vision. It has also profoundly changed the performance of remote sensing scene classification [61], [98], [99]. The widely used deep learning algorithms mainly include autoencoder [100], convolutional neural network (CNN) [101], and generative adversarial network (GAN) [102]. In general, the deep learning algorithms adopt the multilevel network structure to learn the image features adaptively, and regard the classification of remote sensing scenes as an end-to-end problem [50]. Compared with methods based on handcrafted features and feature encoding methods, the methods based on deep learning can extract more abstract and discriminative semantic features and attain better image classification performance [57], [93].

1) *Autoencoder:* As an unsupervised deep learning algorithm, autoencoder can obtain visual representation of the image from unlabeled remote sensing scene for classification [100]. Cheng *et al.* [77] extracted discriminative features of remotes sensing scene images by using autoencoder and single-hidden-layer neural network to achieve effective classification. Du *et al.* [103] proposed a stacked convolutional denoising autoencoder network to break through the limitation of a single autoencoder in feature representation and optimize the performance of scene classification based on autoencoder. The autoencoder and its variants have achieved better results than handcrafted feature methods in remote sensing scene classification [104], [105]. However, most of the autoencoder methods fail to exploit the information of the remote sensing scene fully and cannot learn the most discriminating features of the image [65].
2) *CNN:* CNN is one of the most widely used deep learning methods. Compared with other algorithms, it has outstanding advantages in the field of image processing [106]. Since the AlexNet designed by Krizhevskey *et al.* [101] achieved historical results in the Large-Scale Visual Recognition Challenge (LSVRC) in 2012, numerous advanced deep CNNs were proposed by the researcher to improve the performance of computer vision tasks continuously [107]. The remote sensing scene classification methods based on CNN achieved the best accuracy and outperformed other methods [108], [110]. The representative CNN include CaffeNet [111], VGGNet [112], GoogLeNet [113], ResNet [114], SENet [115], DensNet [116], and SKNet [117]. In 2015, Penatti *et al.* [48] classified the scene images by using CNN and evaluated the generalization capability of conventional CNNs in remote sensing. Based on the BoVW method, Cheng *et al.* [107] replaced handcrafted features with deep convolutional features as input local descriptors to the model, which improved the accuracy of remote sensing scene classification. Lu *et al.* [118] proposed an aggregated feature CNN to learn the image's representation by exploring the semantic label information of the scene. In view of the size of input images, Xie *et al.* [119] designed a scale-free CNN (SF-CNN) scene classification method, which can adapt to the arbitrary size of remote sensing images without resizing. Chen *et al.* [120] introduced knowledge distillation into scene classification to obtain a lightweight CNN model for remote sensing classification.
3) *GAN:* As an essential and promising deep learning method, GAN can model the data distribution through adversarial learning to generate near-real data [102]. GAN consists of a generator and discriminator, in which the data generated by the trained generator should be as close to the real data as possible, while the discriminator has the ability of accurate discrimination, to extract the essential features of the image [102]. Yu *et al.* [121] designed an attention GAN to enhance the representation ability of the discriminator in the network and improve the performance of remote sensing scene classification. The SiftingGAN proposed by Ma *et al.* [122] can generate a variety of reallike labeled remote sensing images for scene classification.

It is worth noting that the remote sensing scene classification based on deep learning is still dominated by the CNN methods. The classification results of autoencoder methods and GAN methods have not yet achieved the performance comparable to that of CNN methods [65]. Therefore, the deep learning methods used to evaluate the proposed NaSC-TG2 dataset in this article are CNN-based algorithms.

III. Proposed Dataset

Various datasets were built to promote the classification of remote sensing scene images [53], [57], [66]–[68], [98]. However, there are still many apparent limitations to remote sensing scene datasets, such as the data source not being rich, the scenes are mainly artificial landscapes, and the small amount of data [56]. These shortcomings hinder the further development of data-driven algorithms in remote sensing, because almost all the deep learning models need to use large training dataset with diverse images for training to avoid overfitting. Therefore, it is critical to propose a scene dataset that is different from the existing datasets and of a larger scale. This led us to propose a natural scene classification benchmark dataset with Tiangong-2 remotely sensed imagery. In this section, we will briefly review the existing datasets and describe in detail the dataset we propose.

A. Existing Datasets for Remote Sensing Scene Classification

1) UC-Merced Dataset: The dataset contains 21 types of land-use scenes with 2100 images in total [53]. All images in the dataset are manually cropped from the National Map Urban Area Imagery produced by the United States Geological Survey (USGS). The source images were collected from various urban areas around the United States. The 21 land-use categories include agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts [53]. To increase the challenge of the data, the dataset has overlapping land-use categories, such as sparse residential, medium residential, and dense residential, which differ only in their structural density. This is the most influential dataset in the remote sensing communities and has been widely used in the classification and retrieval of remote sensing image scenes [7], [15], [50], [53], [71], [84], [87], [93], [97], [98], [130].

2) WHU-RS19 Dataset: The WHU-RS19 was first released in 2010 [66]. After several expansions, the final version consists of 19 scene classes with a total of 1005 images [23]. The image has 600 × 600 pixels. The scene classes include airport, beach, bridge, commercial area, desert, farmland, football field, forest, industrial area, meadow, mountain, park, parking lot, pond, port, railway station, residential area, river, and viaduct [23]. Since the images in this dataset are extracted from Google Earth in different regions of the world, the corresponding scene images vary greatly in scale, orientation, resolution, and illuminations. These challenges also make the dataset widely used to evaluate a variety of classification methods [8], [18], [19], [22], [59], [61], [88], [92], [124], [125], [131]. However, the disadvantage of this dataset is its small number of images per class.

3) AID Dataset: The AID dataset was also extracted from Google Earth imagery and consists of 30 scene types—airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks, and viaduct [67]. The AID dataset has a larger scale than the UC-Merced and WHU-RS19 datasets. The images of each scene in the dataset are carefully selected from different countries and regions, including China, England, Italy, the United States, France, Germany, and Japan [67]. To further increase the intraclass diversity of scene images, different times, seasons, and imaging conditions of the scene are also taken into account when choosing images [68].

4) SIRI-WHU Dataset: The SIRI-WHU dataset consists of 12 remote sensing scenes with a total of 2400 images [98]. There are 200 images in each class, with a size of 200 × 200 pixels and a spatial resolution of 2 m. All images of 12 land-use types are also extracted from Google Earth, including agriculture, commercial, harbor, idle land, industrial, meadow, overpass, park, pond, residential, river, and water [98]. As the images are mainly from urban areas in China, the number of images in each scene class is relatively small, and the dataset also lacks diversity. Several remote sensing scene classification methods have been validated in this dataset [15], [84], [98].

5) NWPU-RESISC45 Dataset: In 2017, Northwestern Polytechnical University (NWPU) published the NWPU-RESISC45 dataset for remote sensing image scene classification (RESIS) [68]. As the dataset's name indicated, NWPU-RESISC45 contains 45 scene classes with a total of 31 500 images. This dataset is large scale in terms of the total number of images and scene classes. Also, the images contain variations in spatial resolution, object pose, translation, illumination, viewpoint, occlusion, and background, which add to the challenge of classification [68].

6) RSSCN7 Dataset: The RSSCN7 dataset covers seven remote sensing scene classes, including grassland, forest, farmland, parking lot, residential region, industrial region, and river/lake, with a total of 2800 remote sensing images [57]. For each scene, 400 images with 400 × 400 pixels were collected from Google Earth and cropped at four different ratios with 100 images per scale [57]. The main drawback of this dataset comes from the change in the scale of the images.

7) RSC11 Dataset: The RSC11 dataset is extracted from Google Earth in Washington, DC, San Francisco, Los Angeles, Chicago, New York, San Diego, and Houston [92]. The 11 scene classes of the dataset include dense forest, grassland, harbor, high buildings, low buildings, overpass, railway, residential area, roads, sparse forest, and storage tanks [92]. This dataset contains a total of 1232 images, with about 100 images in each class. The size of each image is 512 × 512 pixels, and the spatial resolution is 0.2 m. Since the source images of this dataset are also from Google Earth, it is similar to other ones.

B. NaSC-TG2 Dataset

Tiangong-2 was China's first space laboratory, launched on September 15, 2016, and deorbited on July 19, 2019 [132]. It carried out many space scientific experiments and application tests, including those for earth observation [133], [134]. The wideband imaging spectrometer (WIS) was one of the payloads of Tiangong-2 for earth observation and played an essential role in monitoring large-scale objects at medium ground resolution. As a moderate-resolution optical payload, WIS had a wide field of view and wideband, it has 14 spectral channels in programmable visible and near-infrared (0.40–1.04 μm), two spectral channels in short-wavelength infrared (1.232–1.654 μm), and two spectral channels in thermal infrared (8.125–9.275 μm) [133]. The spatial resolution of the above three bands at nadir point was 100, 200, and 400 m, respectively. With a 300-km swath and 42° field of view, the WIS data are suitable for large-scale land surface monitoring, ocean and coastal water color monitoring, and water temperature observation [135], [136]. The observation range of WIS covers all areas between 42°N and 42°S. WIS acquired a total of 19.6 TB of high-quality observation data, covering a total area of 119.1 million km^2 [137].

On the spatial resolution and spectral band range, we chose the visible and near-infrared spectral channels of the WIS as

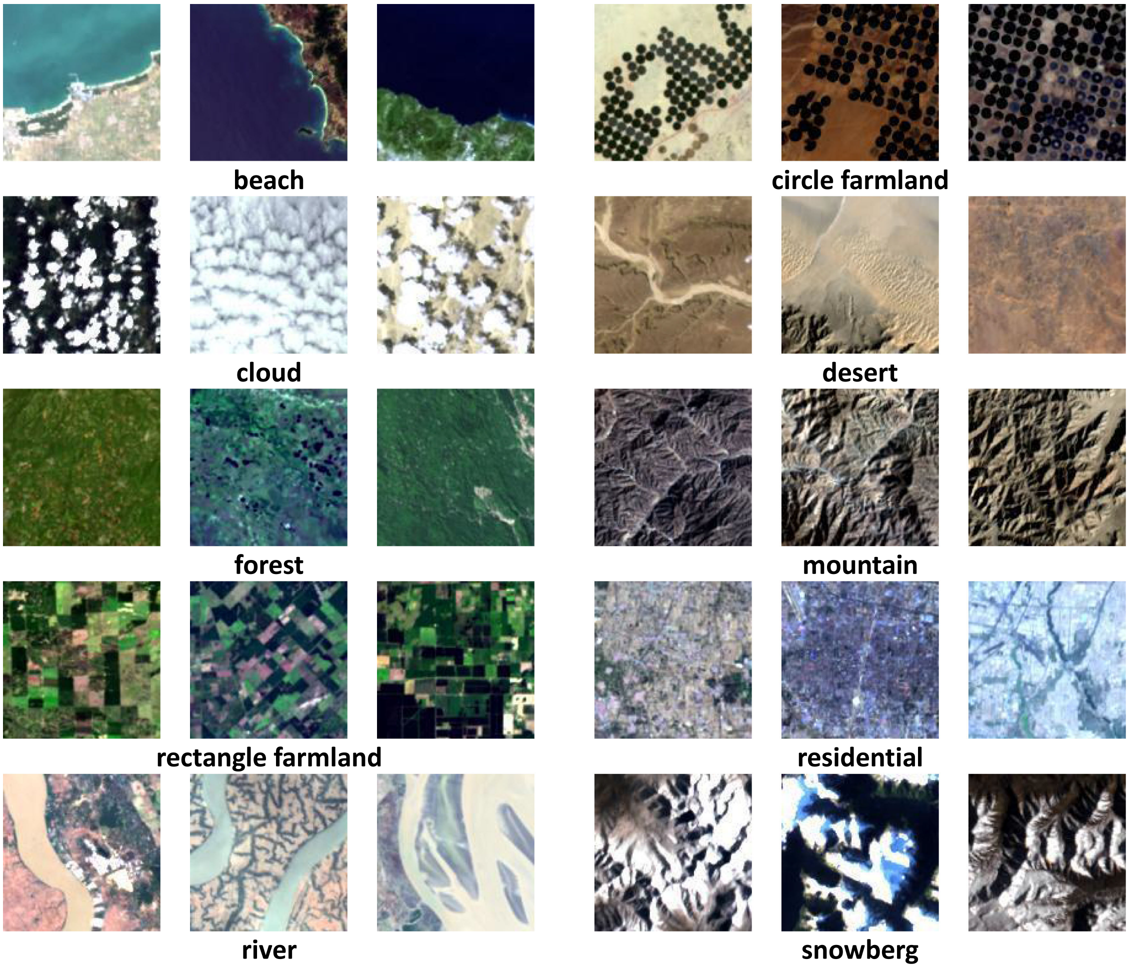


Fig. 1. Examples of each natural scene in the proposed dataset.

TABLE I
EXISTING DATASETS FOR REMOTE SENSING SCENE CLASSIFICATION

Datasets	Image number per class	Number of scene classes	Total image number	Image size
UC Merced	100	21	2100	256×256
WHU-RS19	50~61	19	1005	600×600
AID	220~420	30	10000	600×600
SIRI-WHU	200	12	2400	200×200
NWPU-RESISC45	700	45	61500	256×256
RSSCN7	400	7	2800	400×400
RSC11	about 100	11	1232	512×512
NaSC-TG2	**2000**	**10**	**20000**	**256×256**

the data source. The proposed NaSC-TG2 dataset consists of 20 000 remote sensing images that are divided into ten natural scenes. Each scene includes 2000 images with a size of 128 × 128 pixels, including not only the true-color RGB images but also multispectral images. Because the spatial resolution of the image is 100 m, it can provide remote sensing scenes with a larger spatial scale than other datasets selected from Google Earth, especially suitable for natural scenes. The scenes included in the NaSC-TG2 dataset are beach, circle farmland, cloud, desert, forest, mountain, rectangle farmland, residential, river, and snowberg. All the images are labeled carefully by the remote sensing image interpretation professionals, with samples of each scene shown in Fig. 1. The features of our NaSC-TG2 and other remote sensing scene classification datasets are listed in Table I.

Compared with the existing remote sensing image datasets, the proposed NaSC-TG2 dataset has the following properties.

1) Large Scale: Compared with the tens of millions of labeled images in general image datasets (e.g., ImageNet [64]), the scale of the remote sensing scene datasets needs to be significantly expanded. Otherwise, it will be challenging to realize the full application of data-driven algorithms, such as

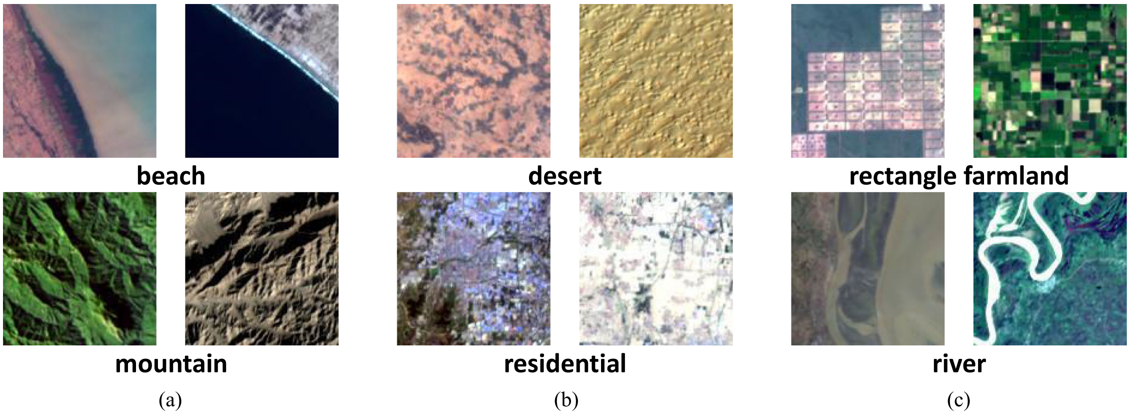


Fig. 2. Large intraclass diversity. (a) Different color images of the same natural scene. (b) Different geometrical structures of the same natural scene. (c) Different object scales of the same natural scene.

deep learning methods, in remote sensing. The scale of almost all the published remote sensing scene datasets is not large enough to adequately train complex deep learning networks from scratch. Therefore, with the proposed NaSC-TG2 dataset, such a large-scale remote sensing scene dataset can supplement the shortage of the number of labeled scene images for the remote sensing communities. Due to the consistent number of each scene in our dataset, a balanced distribution is also more conducive to training networks. Additionally, the validation of a classification method also relies on large-scale labeled data. For smaller scale datasets, the predicted result of whether one image is correct could seriously affect the classification accuracy and cause a large standard deviation, especially when evaluating the classification accuracy of each class. In comparison, the proposed NaSC-TG2 dataset is a large-scale dataset that has an adequate number of images in each scene class, and it can, therefore, provide a better benchmark for evaluating scene classification methods than other datasets.

2) Large Intraclass Differences and High Interclass Similarity: On the one hand, considering the highly complex and changeable conditions on the surface of the earth, the objects in a certain scene may appear in different orientations and sizes, and various scenes may seem to have similar features, e.g., colors and geometrical structures. On the other hand, the imaging conditions of remote sensing sensors are also variable. Therefore, the methods involved in the actual task of remote sensing image classification need to have generalization and robustness, to accurately classify the remote sensing scene images with a large intraclass difference and interclass similarity. The scene images of the dataset proposed here were selected from different regions, seasons, weather conditions, illumination conditions, and scales to maximize the intraclass difference. In addition, when designing scene classes and selecting images, we also considered the similarity between scenes better to match the actual task of remote sensing classification. The comparison of the sample images in Fig. 2 shows that the appearance of the same scene in our dataset has rich variations in color, spatial structure, and object scale. For example, the mountain scene images in different seasons have different colors; the desert scene images in different regions have different geographical structures; the rivers with different scales also show richness in diversity. Besides, there is also the interclass similarity in the NaSC-TG2 dataset, as shown in Fig. 3; some scene images of the dataset sharing similar features, e.g., the circle farmland and desert share similar structural distributions; the forest and the mountain may be very close in color; the cloud and snowberg have similar objects.

3) Natural Scenes With Novel Spatial Scale and Imaging Performance: Almost all existing remote sensing scene dataset images were selected from Google Earth imagery. A single data source leads to great similarities and redundancy among different datasets, and the scene types are mainly artificial landscapes at high spatial resolution. However, the land cover of the earth is dominated by natural objects, and the classification of natural scenes is more practical. The disadvantages of existing datasets restrict the research and development of remote sensing classification methods. The labeled images of the dataset presented in the current study are all extracted from the Tiangong-2 remotely sensed imagery. Compared with other datasets, the NaSC-TG2 has abundant natural scenes with novel spatial scale and imaging performance. The more diverse remote sensing scene images could lead to more comprehensive verification and analysis of the algorithms, especially for the more practical natural scene classification research. In addition to true-color RGB images, the NaSC-TG2 dataset also covers the corresponding 14-band multispectral scene images, providing valuable experimental data for research on high-dimensional scene image classification algorithms.

IV. Benchmarking Representative Methods

In this section, we evaluate some of the representative classification methods by the proposed NaSC-TG2 dataset to examine their performance on natural scenes of remote sensing. In addition, the large scale of the NaSC-TG2 dataset, especially its advantages in the number of images in each class, enables the performance of the classification methods to be more objectively

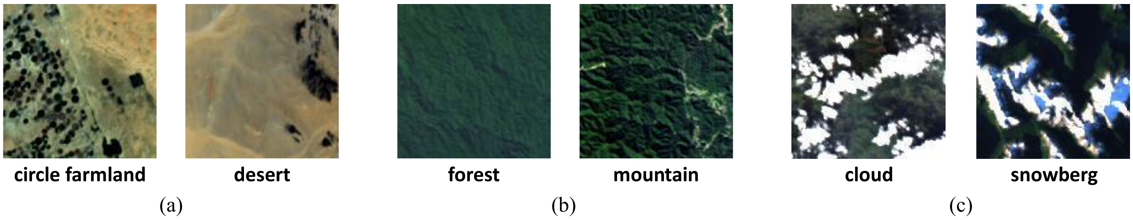


Fig. 3. Small interclass distance. (a) Similar structural distributions between different natural scenes. (b) Similar colors between different natural scenes. (c) Similar objects between different natural scenes.

evaluated. It is worth noting that this experiment focuses only on RGB images in the NaSC-TG2 dataset.

A total of 11 representative methods in the field of remote sensing classification were selected for experimental analysis of the NaSC-TG2 dataset, including three methods based on handcrafted features (CH, LBP, SIFT), three feature encoding methods (BoVW, pLSA, LDA), and five deep learning methods (AlexNet, VGGNet-16, GoogLeNet, ResNet-34, Inception-v3 [138]).

A. Representative Methods

1) Color Histograms: The CH feature is easy to calculate and is widely used in remote sensing image classification. First, the CH feature in the RGB color space of each image was extracted, in which each channel is quantized and forms a total histogram feature. Second, the histogram was further normalized, and the L1 norm of the normalized histogram was 1. Finally, the scene images were classified by a trained classifier based on the extracted CH features.

2) Local Binary Patterns: The LBP feature is calculated by comparing the grey value of the $\mathrm{N} \times \mathrm{N}$ window center pixel and the adjacent $4\times(\mathrm{N}-1)$ pixels. If the value of the surrounding pixel is greater than that of the center pixel, mark the position of the pixel as 1; otherwise, it is 0. The $4\times(\mathrm{N}-1)$ points in the $\mathrm{N} \times \mathrm{N}$ neighborhood can be compared to produce the LBP value of the central pixel in the form of a $2^{4\times(\mathrm{N}-1)}$ -bit binary number, which reflects the texture information of the image with a $2^{4\times(\mathrm{N}-1)}$-dimensional feature vector. Similarly, we classify the feature vector based on a classifier to determine to which scene class the feature vector belongs.

3) Scale Invariant Feature Transform: The SIFT feature extracted from the scene image of the dataset is a feature vector obtained by calculating the gradient histogram of the $\mathrm{N} \times \mathrm{N}$ spatial grids in the image and quantizing it in bins. The classifier takes the SIFT feature of the image as the input and outputs the corresponding label value.

4) Bag-of-Visual-Words: In the experiment based on the BoVW method, we first conducted dense or sparse sampling local areas from the image to extract image patches in size of $\mathrm{N} \times \mathrm{N}$ pixels, followed by calculating the corresponding handcrafted feature as the feature descriptors of these patches, such as CH, LBP, and SIFT. The feature descriptors extracted from all training images are clustered to generate the visual codebook. The clustering algorithm used in this article is the unsupervised k-means cluster. The features of all images in the dataset are then encoded by the mapping relationship between the feature descriptors and the codebook determined by clustering. The trained classifier can predict which scene class the image belongs to according to the coding feature.

5) Probabilistic Latent Semantic Analysis: pLSA introduces a latent variable called topic to improve the BoVW model, which is used to describe the conditional probability distribution of visual words in the dictionary to establish the connection with the dataset's images. In this article, we used a fixed Gaussian distribution as the distribution of visual words in the dictionary. By defining the number of topics and then describing the image with the distribution of topics to reduce the influence of synonym and polysemy, the dimensions of feature can be reduced to be consistent with the number of the topics.

6) Latent Dirichlet Allocation: As a generative topic model, the LDA is improved based on the pLSA model. The main improvement is the addition of a Dirichlet distribution before describing the latent variable topic, which solves the problem of overfitting and enhances the model's robustness. The feature dimensions are consistent with the pLSA model and the number of topics we defined.

7) AlexNet: The AlexNet has a classic network architecture, consisting of five convolutional layers, three pooling layers, two fully connected layers, and a softmax layer [101]. The first two convolutional layers are followed by a normalization layer, and the two normalization layers and the third convolutional layer are followed by the pooling layer. The ReLU (rectified linear units) function is used as the activation function of the network. The output of the second fully connected layer of the AlexNet is a 4096-dimension feature vector, and the classifier predicts the scene class based on the extracted feature vector of the image.

8) VGGNet-16: VGGNet-16 is a CNN architecture containing 13 convolutional layers, five pooling layers, and three fully connected layers [112]. Compared with the AlexNet, the improvement of VGG-16 is to replace the large kernel-sized convolutional filters in the network (11 and 5 in the first and second convolutional layers, respectively) with multiple connected 3×3 kernel-sized filters. The extracted feature from the second fully connected layer of the VGGNet-16 is also a 4096-dimension vector. Based on this feature vector, the classifier predicts the label of the scene image.

9) GoogLeNet: GoogLeNet is a representative CNN model, which was the winner of the ILSVRC-2014 in classification and detection [113]. While AlexNet and VGGNet have fixed convolution kernel sizes, GoogLeNet introduced the concept of an inception module to extract various kinds of features of the image by combining 1 × 1 conv, 3 × 3 conv, 5 × 5 conv, and 3 × 3 max pooling. Due to its 1 × 1 conv at the middle of the network, it can reduce the number of parameters of the network. This allows the network to be deep enough without being easy to overfitting.

10) ResNet-34: ResNet is an innovative architecture called residual network, which solves the gradient optimization problem by improving the structure of the network [114]. For a network with too many layers, the network can easily fall into a vanishing or exploding gradient during the training phase. Through skip connections, the ResNet can be connected to the output directly by skipping a few layers of training. In the way, the network can learn to fit the residual mapping rather than learning the underlying mappings by the instead of layers. According to the number of layers, ResNet has many variants, including ResNet-18, ResNet-34, ResNet-50, ResNet-101, etc. Among them, ResNet-34 is one of the most vibrant networks on its own.

11) Inception-V3: GoogLeNet and Inception-v3 are both the architecture of the convolutional network of the Inception family, and the latter utilizes several techniques, such as factorized convolutions, regularization, dimension reduction, and parallelized computations to loosen the constraints for easier model adaptation [138]. The inception-v3 predicts the class of the scene image based on the 2048-dimensional feature vector extracted from the last pooling layer of the network.

B. Experimental Setup

1) Parameter Settings: For the methods based on handcrafted features in our experiment, the CH and LBP features we extracted are the global descriptors that efficiently represent the entire scene image. Specifically, in the feature extraction of CH, we calculate the statistical histograms in the color space of RGB and quantize each channel into 32 bins, then combine the feature of the three channels to form a 96-dimensional vector. For LBP, we set the window size to 3 × 3, and the grey value of the center pixel in the window is compared with the eight adjacent pixels to obtain 8-bit binary values. The 8-bit binary values can represent 256 patterns, which are the LBP features of an image. Unlike CH and LBP, the SIFT is the local patch descriptor of an image. We extract all the descriptors from the grey image plain using a 16 × 16 size grid with a spacing step of eight pixels. Each dimension of the descriptor is then averaged to obtain a 128-dimensional SIFT feature of the image.

For feature encoding methods, we use CH, LBP, and SIFT as local patch descriptors to extract the spectral, texture, and structural features of the image, respectively. In the process of patch sampling, we use grid sampling, which has been proven to obtain better results in remote sensing scene classification [98]. In our experiment setting, the patch size of all the local descriptors is 16 × 16 pixels with a spacing step of eight pixels to balance the speed and accuracy. The three local feature descriptors and three global feature encoding methods can be combined into nine results. We set the size of the dictionary at 1000, 2000, 3000, 4000, and 5000 to study the way the different sizes affected the classification performance, and then selected the optimal size. For certain parameters, such as the number of topics for pLSA and LDA, we set them both to half the dictionary's size, based on previous experience [67], [68].

The 4096-dimensional vector extracted from the second fully connected layer of the trained AlexNet and VGGNet-16, the 1024-dimensional, 1000-dimensional, and 2048-dimensional vector formed by the last pooling layer of the trained GoogLeNet, ResNet-34, and Inception-v3 are the final global features. All five deep learning models were implemented on a PC with 2 GHz 20-core CPUs and 32 GB of RAM. Two Nvidia Titan RTX GPUs were also used for acceleration.

To make a fair comparison between the different methods, we use the linear support vector machines (SVMs) [47] as the classifier for all 11 kinds of image features. Specifically, the dataset is divided into the training set and test set, according to a certain proportion. The features extracted from the training set are used to train the linear SVM classifier, and the features extracted from the test set are used for evaluating the performance of the trained model.

2) Evaluation Protocols: In this article, we chose the overall accuracy (OA) and confusion matrix as evaluation indicators to quantify different classification methods. OA is defined as the ratio of correctly classified images to the total number of images. It reflects the classification performance of the entire dataset by direct measurement. Since the number of each scene in our dataset is the same, the overall accuracy also represents the average accuracy of all scenes.

The confusion matrix, also known as the error matrix, displays the performance of the classification visually through a specific table layout. The percentage of predicted instances to actual instances is shown in each cell of the matrix. The confusion between different scenes in the predicted results of the method can be seen intuitively from the matrix.

In our evaluation, the training set was randomly selected from the dataset at ratios of 10% and 20%, respectively, for supervised training of the classification model, and the remaining 90% and 80% data were used as the test set for validation. To reduce the random error and obtain reliable results, we repeated the evaluation ten times to calculate the average OA and the corresponding standard deviation.

C. Experimental Results

In this section, we provide the corresponding results and analysis of the different methods on our dataset, specifically including the results of the three categories of feature extraction methods, the confusion matrix, and visual comparative analysis.

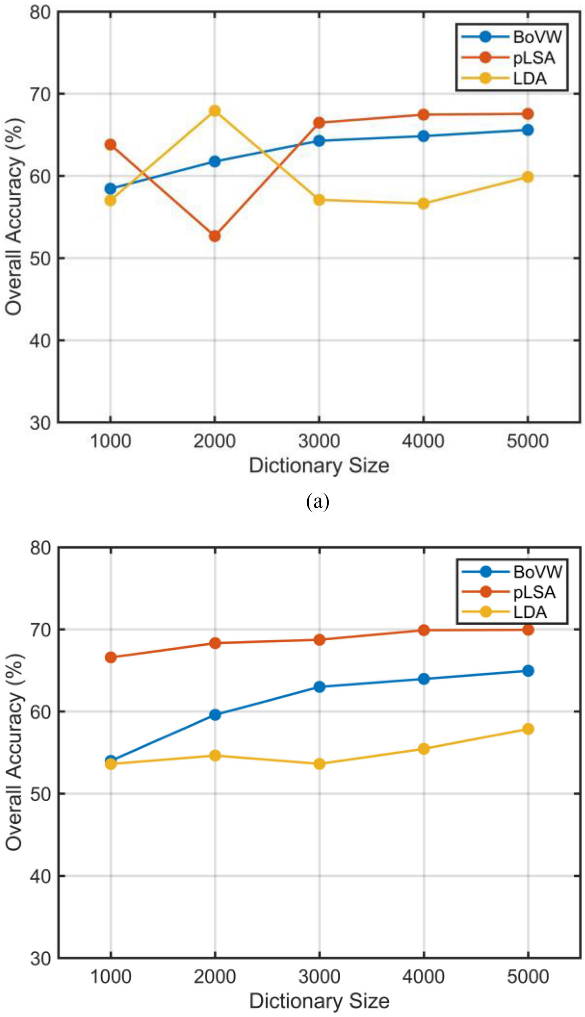
1) Results of the Methods Based on Handcrafted Features: Table II lists the means and standard deviation of OA of the three methods based on handcrafted features, CH, LBP, and SIFT. LBP has the best performance, indicating that for the NaSC-TG2 dataset, the texture descriptor can represent the scene feature better. Considering intraclass differences of the dataset, the colors are not uniform in each scene, and the CH feature performs worse than LBP. The performance

TABLE II
OVERALL ACCURACIES (%) OF THE THREE METHODS BASED ON HANDCRAFTED FEATURES AT 10% AND 20% TRAINING RATIOS

Methods	Training ratios	
	10%	20%
CH	57.12±0.35	58.65±0.27
LBP	65.79±0.53	69.15±0.24
SIFT	37.38±0.50	40.08±0.35

TABLE III
OVERALL ACCURACIES (%) OF NINE FEATURE ENCODING METHODS AT 10% AND 20% TRAINING RATIOS

Methods	Training ratios	
	10%	20%
BoVW (CH)	63.77±0.12	67.41±0.74
pLSA (CH)	53.06±0.21	54.56±0.32
LDA (CH)	64.58±0.40	66.34±0.44
BoVW (LBP)	74.33±0.64	74.77±0.53
pLSA (LBP)	68.13±0.13	69.86±0.76
LDA (LBP)	76.07±0.26	77.62±0.17
BoVW (SIFT)	64.41±0.63	65.01±0.60
pLSA (SIFT)	55.90±0.35	58.19±0.27
LDA (SIFT)	68.94±0.27	70.59±0.32



(a)

(b)

Fig. 4. Overall accuracies of BoVW, pLSA, and LDA methods with the dictionary sizes of 1000, 2000, 3000, 4000, and 5000, respectively, at (a) 10% and (b) 20% training ratios.

of the SIFT descriptor is worse than the other two, and the OA value is lower by more than 20%, indicating that SIFT is not suitable for directly classifying our dataset. Since all scenes are natural landscapes in our dataset, the texture features of different scenes are more robust and distinguishable than color features. In addition, when handcrafted features are directly used for large-scale scene classification, the global features such as texture and color are better than the local features such as SIFT.

2) Results of the Feature Encoding Methods: For feature encoding methods, the dictionary's size is one of the critical parameters that affect the performance of the results. Therefore, the optimal size of the methods at different ratios must be determined in the first place. We use the LBP as the local patch descriptor to compare the performance of the BoVW, pLSA, and LDA at dictionary sizes from 1000 to 5000, in steps of 1000.

Fig. 4 shows the corresponding OA at the different dictionary sizes. For the training ratio of 10%, the dictionary sizes worked best at 5000, 2000, and 5000. For the training ratio of 20%, the dictionary sizes worked best at 5000, 5000, and 5000, for the BoVW, pLSA, and LDA methods, respectively. We carried out the subsequent evaluation based on the optimal dictionary size for the different methods.

The local feature descriptions correspond to the three feature encoding methods, which combined into nine classification results. Table III shows the means and standard deviation of OA for each result. Compared with the classification based on the handcrafted feature directly, the contribution of the different feature descriptors to the classification has changed after encoding. The SIFT feature improved from being lower in performance than CH and LBP to 20% higher than CH in OA. This indicates that the SIFT feature is not suitable for classification directly, but it can generate more robust feature representation through encoding. The classification performance based on LBP is still the best. The differences in texture and spatial structure of scenes are more evident than others for natural targets in our dataset. The texture feature is further enhanced after encoding, and the OA is about 10% higher than other features.

When the feature encoding methods are compared, it can be seen that the LDA method has the best performance, while BoVW being slightly worse. The performance of pLSA is the worst under different training ratios, and the OA is nearly 10% lower than the other two. Compared with BoVW and pLSA, LDA could discover more descriptive topics, leading to more distinguishable descriptors. In addition, more training data could improve the overall accuracy of all methods to a certain extent.

3) Results of the Deep Learning Methods: Table IV displays the means and standard deviation of OA for the five deep learning methods. From the classification results, AlexNet, GoogLeNet, VGG-16, ResNet-34, and Inception-v3 all achieve excellent performance, far better than the methods based on handcrafted features and feature encoding methods. This indicates that the deep learning methods can learn more

TABLE IV
OVERALL ACCURACIES (%) OF FIVE DEEP LEARNING METHODS AT 10% AND 20% TRAINING RATIOS

Methods	Training ratios	
	10%	20%
AlexNet	87.22±0.31	89.39±0.14
VGG-16	87.39±0.22	89.59±0.17
GoogLeNet	85.41±0.15	87.76±0.17
ResNet-34	86.25±0.11	88.37±0.13
Inception-v3	85.55±0.13	86.75±0.15

discriminable features from remote sensing scene images. Of the five, AlexNet and VGG-16 show similar performance on the NaSC-TG2 dataset, while GoogLeNet, ResNet-34, and Inception-v3 perform slightly worse under both training ratios.

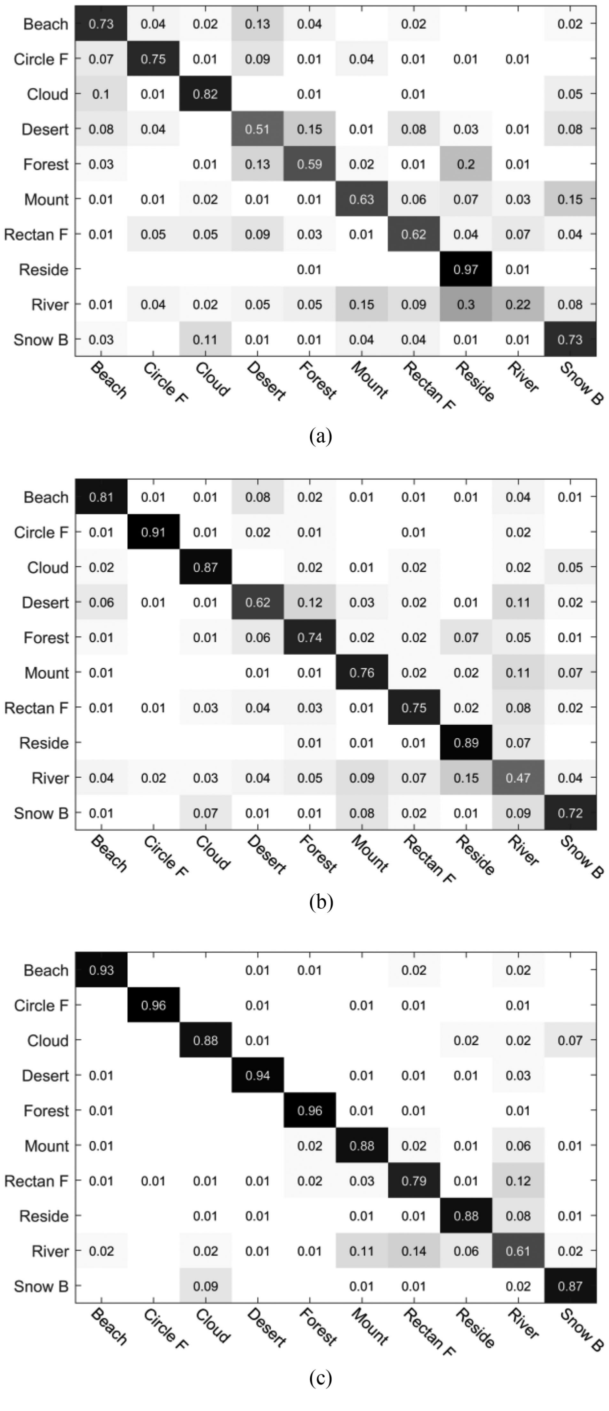
The five deep learning networks are all designed for general digital images, the GoogLeNet (22 layers), ResNet-34 (34 layers), and Inception-v3 (47 layers) are more in-depth than AlexNet (10 layers) and VGG-16 (16 layers). In theory, the deeper networks are more likely to exploit the inherent and discriminative features from images. However, the images from the NaSC-TG2 dataset are natural scenes with large spatial scale objects of the land surface which are quite different from general digital images; the deeper network may not perform the best. Furthermore, because all five CNN methods are trained from scratch, the deeper networks may depend on more extensive training for better performance. The existing experimental conditions could not fully explore the ability of deeper networks. The classification of the remote sensing scenes of our dataset may be more suitable for networks of moderate depth, such as AlexNet, which has eight layers, and VGGNet-16, which has 16 layers.

Besides, due to a large number of images per scene in our dataset, the standard deviation of the OA for all the above methods is relatively small, which makes the evaluation of the various methods more accurate.

D. Confusion Matrix

In addition to the OA, we also calculated the corresponding confusion matrix to visualize the performance of each class from various methods. For the NaSC-TG2 dataset, in Figs. 5 and 6, we show the confusion matrix corresponding to the best results based on handcrafted features, the feature encoding methods, and the deep learning methods under different training ratios.

For the methods based on handcrafted features, only the classification accuracy of the cloud and residential scenes are above 0.8. In terms of LBP feature, cloud and snowberg, forest, and desert are easily confused. For the feature encoding method, the classification accuracy of nearly half of the scenes is more than 0.8, which represents a significant improvement over the method based on handcrafted features, especially for the river, circle farmland, and forest. However, the accuracy of residential is slightly reduced, possibly because the regular distribution of residential areas from the wideband imagery of Tiangong-2 makes the simple texture more representable than the encoding feature. For the deep learning method, all of the scenes can be easily distinguished from one another, and the classification accuracies of most classes are close to 0.9. Only the river scene has a slightly lower classification accuracy, because some rivers are distributed in farmland and residential areas, which may be confused with the surroundings.



(a)

(b)

(c)

Fig. 5. Confusion matrix obtained by methods based on handcrafted features (LBP), feature encoding LDA (LBP), and deep learning (VGG-16) on the proposed dataset under the training ratios of 10%.

Compared with the shallow texture feature, the accuracy of desert, river, and forest scenes increased by more than 35%, indicating that the higher level features extracted from the

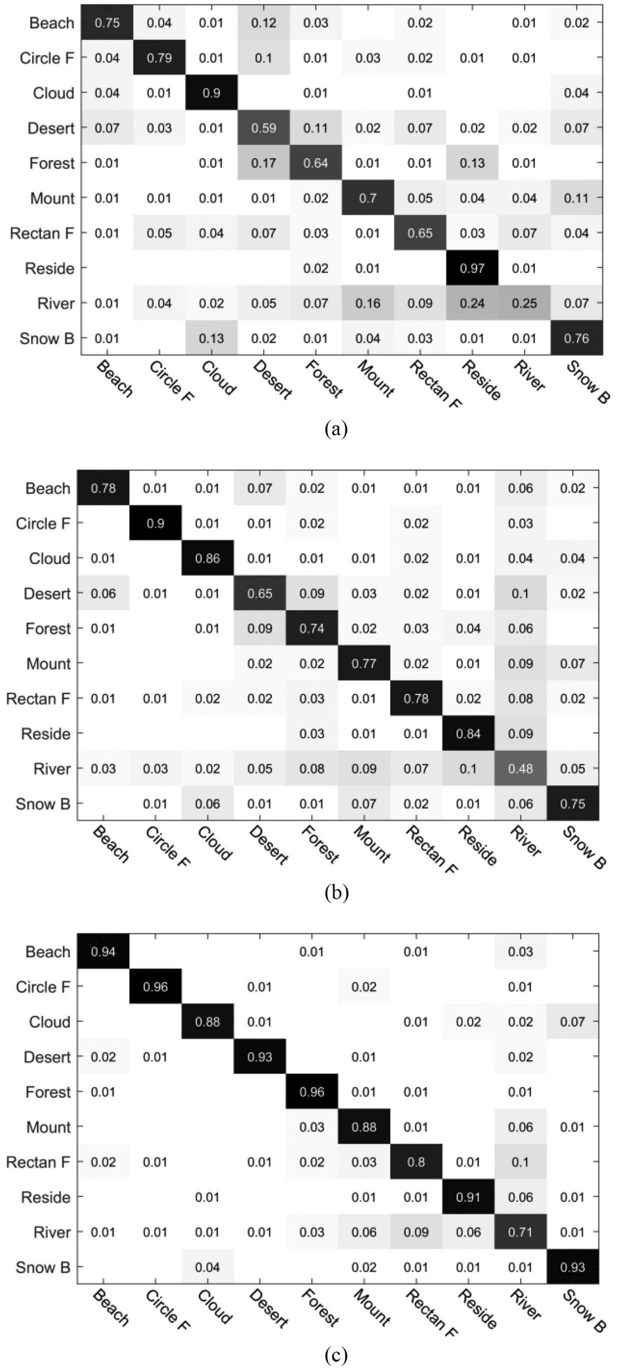


Fig. 6. Confusion matrix obtained by methods based on handcrafted features (LBP), feature encoding LDA (LBP), and deep learning (VGG-16) on the proposed dataset under the training ratios of 20%.

deep learning network can significantly enhance the representative ability of the above scenes and improve the classification accuracy.

The evaluations show that the deep learning methods exhibited the best performance, followed by the feature encoding methods, and the methods based on handcrafted features. The deep learning methods represent the state-of-the-art for remote sensing scene classification and have significant advantages over other methods.

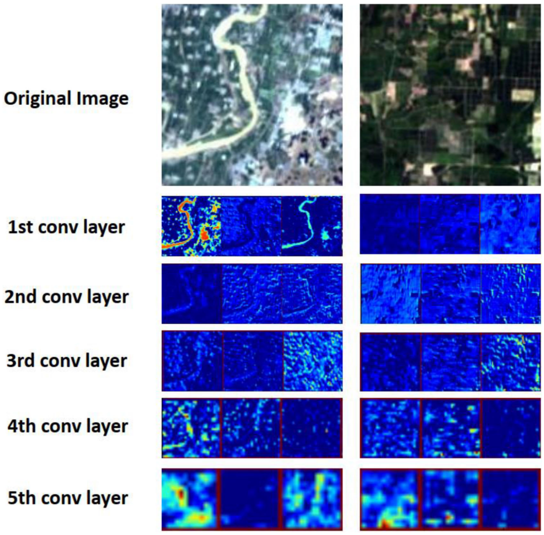


Fig. 7. Visualization of convolutional feature maps of the river (left) and rectangular farmland (right) scene of different layers in the VGG-16 network.

To further analyze the causes of misclassification of the deep learning method, we select representative scene images from the NaSC-TG2 dataset and extract their corresponding convolutional features in different layers in the VGG-16 network for visual comparison in Fig. 7.

Due to the spatial scale changes of actual natural surface objects, the river scene images in our dataset are quite different. From the perspective of the scene, the semantic theme of the image is the river, but the area of the river itself may be much smaller than the others, such as the rivers flowing through residential and farmland areas.

It can be seen from Fig. 7 that in the shallow convolutional features of the VGG-16 network, the representations of the river in the image are enhanced; the river in the feature maps has obvious boundaries and contours. However, as the convolutional features progress to the deeper levels, the representations of the river are gradually weakened; the feature maps may represent more of the overall scene and no longer focus on the object of the river itself.

For the fifth layer of convolutional features of the VGG-16 network, there is no descriptive visual difference between the features corresponding to the river and the rectangular farmland scenes. This may be because the river area in the scene is small, and the semantic knowledge learned by the network is no longer the river but other surrounding contents, which leads to misclassification. To achieve a better classification of the natural scene on our dataset, the classification algorithm needs to pay more attention to the main semantic content of the scene without being disturbed by the area ratio.

V. Conclusion

In this article, we proposed a dataset called "NaSC-TG2," a novel benchmark dataset for remote sensing natural scene classification from Tiangong-2 remotely sensed imagery.

The NaSC-TG2 dataset contains 20000 images, which are divided into ten scene classes, with 2000 images for each. It is large scale, especially in terms of the number of each class, and can be used for data-driven algorithm study. The dataset has a large intraclass difference and high interclass similarity, which matches the actual remote sensing classification task. The scene images from NaSC-TG2 dataset are all taken from the Tiangong-2 wideband imagery. It offers natural scenes with novel spatial scale and imaging performance, enriching the diversity of scenes compared with other datasets used by remote sensing communities, which is suitable for evaluating different remote sensing scene classification methods, especially for natural scenes.

This dataset will be used as experimental data to contribute to the research of classification algorithms. The evaluations of representative classification methods based on the proposed dataset provide baseline results for future algorithm development.

Acknowledgment

The authors would like to thank all the researchers who kindly shared the codes. The dataset they proposed and codes for reproducing all the results in this article are freely available from http://www.msadc.cn/jszc_xzq/

References

[1] Q. Hu *et al.*, "Exploring the use of Google Earth imagery and object-based methods in land use/cover mapping," *Remote Sens.*, vol. 5, no. 11, pp. 6026–6042, 2013.

[2] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[3] N. Joshi *et al.*, "A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring," *Remote Sens.*, vol. 8, no. 1, 2016, Art. no. 70.

[4] R. Khatami, G. Mountrakis, and S. V. Stehman, "A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research," *Remote Sens. Environ.*, vol. 177, pp. 89–100, 2016.

[5] P. Liang, W. Shi, and X. Zhang, "Remote sensing image classification based on stacked denoising autoencoder," *Remote Sens.*, vol. 10, no. 2, 2017, Art. no. 16.

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

[7] B. Luo, S. Jiang, and L. Zhang, "Indexing of remote sensing images with different resolutions by multiple features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 1899–1912, Aug. 2013.

[8] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818–832, Feb. 2013.

[9] E. Raczko and B. Zagajewski, "Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images," *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 144–154, 2017.

[10] 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.

[11] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[12] H. Bagan and Y. Yamagata, "Improved subspace classification method for multispectral remote sensing image classification," *Photogramm. Eng. Remote Sens.*, vol. 76, no. 11, pp. 1239–1251, 2010.

[13] V. Risojevic and Z. Babic, "Fusion of global and local descriptors for remote sensing image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 836–840, Jul. 2013.

[14] A. Avramović and V. Risojević, "Block-based semantic classification of high-resolution multispectral aerial images," *Signal, Image Video Process.*, vol. 10, no. 1, pp. 75–84, 2014.

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

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

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

[18] 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.

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

[20] L. Zou, X. Zhu, C. Wu, Y. Liu, and L. Qu, "Spectral–spatial exploration for hyperspectral image classification via the fusion of fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 659–674, Feb. 2020.

[21] J. Liu, Z. Wu, Z. Wei, L. Xiao, and L. Sun, "Spatial-spectral kernel sparse representation for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 6, pp. 2462–2471, Dec. 2013.

[22] L. Dong *et al.*, "Very high resolution remote sensing imagery classification using a fusion of random forest and deep learning technique—Subtropical area for example," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 113–128, Dec. 2020.

[23] 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, 2011.

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

[25] 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.

[26] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.

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

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

[29] G. P. Petropoulos, C. Kalaitzidis, and K. P. Vadrevu, "Support vector machines and object-based classification for obtaining land-use/cover cartography from hyperion hyperspectral imagery," *Comput. Geosci.*, vol. 41, pp. 99–107, 2012.

[30] G. Yan, J.-F. Mas, B. H. P. Maathuis, Z. Xiangmin, and P. M. V. 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.

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

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

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

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

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

[36] J. Zhao, Y. Zhong, H. Shu, and L. Zhang, "High-resolution image classification integrating spectral-spatial-location cues by conditional random fields," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4033–4045, Sep. 2016.

[37] R. P. D. Lima and K. Marfurt, "Convolutional neural network for remote-sensing scene classification: Transfer learning analysis," *Remote Sens.*, vol. 12, no. 1, 2019, Art. no. 86.

[38] B. Zhao, Y. Zhong, and L. Zhang, "Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery," *Remote Sens. Lett.*, vol. 4, no. 12, pp. 1204–1213, 2013.

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

[40] W. Shao, W. Yang, and G.-S. Xia, "Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification," *Int. J. Remote Sens.*, vol. 34, no. 23, pp. 8588–8602, 2013.

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

[42] 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, Apr. 2011.

[43] Y. Yang and S. Newsam, "Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery," in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 1852–1855.

[44] J. A. D. Santos and R. D. S. Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2010, pp. 203–208.

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

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

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

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

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

[50] 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.

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

[52] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8905–8918, Dec. 2020.

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

[54] 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. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, 2019.

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

[56] 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.

[57] 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.

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

[59] L. Li and N. Shu, "Object-oriented classification of high-resolution remote sensing image using structural feature," in *Proc. 3rd Int. Congr. Image Signal Process.*, 2010, pp. 2212–2215.

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

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

[62] V. Risojević, S. Momić, and Z. Babić, "Gabor descriptors for aerial image classification," in *Proc. Int. Conf. Adaptive Natural Comput. Algorithms*, 2011, pp. 51–60.

[63] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. IV-2/W7, pp. 153–160, 2019.

[64] 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. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[65] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, Jun. 2020.

[66] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maitre, "Structural high-resolution satellite image indexing," in *ISPRS TC VII Symposium-100 Years ISPRS*, vol. 38, pp. 298–303, 2010.

[67] 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.

[68] 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.

[69] G. Cheng *et al.*, "Object detection in remote sensing imagery using a discriminatively trained mixture model," *ISPRS J. Photogramm. Remote Sens.*, vol. 85, pp. 32–43, 2013.

[70] S. Bhagavathy and B. Manjunath, "Modeling and detection of geospatial objects using texture motifs," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 12, pp. 3706–3715, Dec. 2006.

[71] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 3023–3034, May 2014.

[72] H. Li, H. Gu, Y. Han, and J. Yang, "Object-oriented classification of high-resolution remote sensing imagery based on an improved colour structure code and a support vector machine," *Int. J. Remote Sens.*, vol. 31, no. 6, pp. 1453–1470, 2010.

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

[74] S. Newsam, L. Wang, S. Bhagavathy, and B. S. Manjunath, "Using texture to analyze and manage large collections of remote sensed image and video data," *Appl. Opt.*, vol. 43, no. 2, pp. 210–217, 2004.

[75] X. Huang, L. Zhang, and L. Wang, "Evaluation of morphological texture features for mangrove forest mapping and species discrimination using multispectral IKONOS imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 3, pp. 393–397, Jul. 2009.

[76] 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. Vis. Pattern Recognit.*, 2015, pp. 1173–1181.

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

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

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

[80] 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.

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

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

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

[84] B. Zhao, Y. Zhong, L. Zhang, and B. Huang, "The Fisher kernel coding framework for high spatial resolution scene classification," *Remote Sens.*, vol. 8, no. 2, 2016, Art. no. 157.

[85] L. Huang, C. Chen, W. Li, and Q. Du, "Remote sensing image scene classification using multi-scale completed local binary patterns and fisher vectors," *Remote Sens.*, vol. 8, no. 6, 2016, Art. no. 483.

[86] G. Cheng, J. Han, P. Zhou, and L. Guo, "Scalable multi-class geospatial object detection in high-spatial-resolution remote sensing images," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2014, pp. 2479–2482.

[87] K. Qi, H. Wu, C. Shen, and J. Gong, "Land-use scene classification in high-resolution remote sensing images using improved correlatons," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2403–2407, Dec. 2015.

[88] J. Ren, X. Jiang, and J. Yuan, "Learning LBP structure by maximizing the conditional mutual information," *Pattern Recognit.*, vol. 48, no. 10, pp. 3180–3190, 2015.

[89] C. Yao and G. Cheng, "Approximative Bayes optimality linear discriminant analysis for Chinese handwriting character recognition," *Neurocomputing*, vol. 207, pp. 346–353, 2016.

[90] 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.

[91] J. Zhang, T. Li, X. Lu, and Z. Cheng, "Semantic classification of high-resolution remote-sensing images based on mid-level features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2343–2353, Jun. 2016.

[92] 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.

[93] 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.

[94] S. Xu, T. Fang, D. Li, and S. Wang, "Object classification of aerial images with bag-of-visual words," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 2, pp. 366–370, Apr. 2010.

[95] 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, pp. 45–59, 2012.

[96] R. Bahmanyar, S. Cui, and M. Datcu, "A comparative study of bag-of-words and bag-of-topics models of EO image patches," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 6, pp. 1357–1361, Jun. 2015.

[97] H. Wu, B. Liu, W. Su, W. Zhang, and J. Sun, "Hierarchical coding vectors for scene level land-use classification," *Remote Sens.*, vol. 8, no. 5, 2016, Art. no. 436.

[98] 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.

[99] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 844–856, Jul. 2013.

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

[101] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

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

[103] 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.

[104] 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.

[105] 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.

[106] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] Y. Jia *et al.*, "Caffe," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[112] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

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

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

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

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

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

[118] 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.

[119] 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.

[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] 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. 2019.

[122] 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, Jan. 2019.

[123] 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.

[124] S. Cui, "Comparison of approximation methods to Kullback–Leibler divergence between Gaussian mixture models for satellite image retrieval," *Remote Sens. Lett.*, vol. 7, no. 7, pp. 651–660, 2016.

[125] G.-S. Xia, Z. Wang, C. Xiong, and L. Zhang, "Accurate annotation of remote sensing images via active spectral clustering with little expert knowledge," *Remote Sens.*, vol. 7, no. 11, pp. 15014–15045, 2015.

[126] 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.

[127] Y. Zhang, X. Zheng, G. Liu, X. Sun, H. Wang, and K. Fu, “Semi-supervised manifold learning based multigraph fusion for high-resolution remote sensing image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 464–468, Feb. 2014.

[128] R. Negrel, D. Picard, and P. Gosselin, “Evaluation of second-order visual features for land-use classification,” in *Proc. 12th Int. Workshop Content-Based Multimedia Indexing*, 2014, pp. 1–5, doi: 10.1109/CBMI.2014.6849835.

[129] L. Gueguen, “Classifying compound structures in satellite images: A compressed representation for fast queries,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1803–1818, Apr. 2015.

[130] B. Fernando, E. Fromont, and T. Tuytelaars, “Mining Mid-level features for image classification,” *Int. J. Comput. Vis.*, vol. 108, no. 3, pp. 186–203, 2014.

[131] J. Zou, W. Li, C. Chen, and Q. Du, “Scene classification using local and global features with collaborative representation fusion,” *Inf. Sci.*, vol. 348, pp. 209–226, 2016.

[132] M. Gao, “Earth observation payloads and data applications of Tiangong-2 space laboratory,” in *Proc. Tiangong-2 Remote Sens. Appl. Conf.*, 2018, pp. 1–13.

[133] J. Wei, L. Ding, X. He, Y. Tang, X. Feng, and M. Lin, “Design, performance and in-orbit evaluation results of Tiangong-2 wide-band imaging spectrometer,” in *Proc. Tiangong-2 Remote Sens. Appl. Conf.*, 2018, pp. 14–27.

[134] S. Li *et al.*, “Mapping high mountain lakes using space-borne near-nadir SAR observations,” *Remote Sens.*, vol. 10, no. 9, 2018, Art. no. 1418.

[135] Y. Tang, J. Wei, X. Huang, X. Feng, and Q. Song, “Research on on-board calibration system of Tiangong-2 wide-band imaging spectrometer,” in *Proc. Tiangong-2 Remote Sens. Appl. Conf.*, 2018, pp. 28–39.

[136] L. Li, X. Liu, X. Huang, and Y. Tang, “Automatic multi-spectral image registration for Tiangong-2 wide-band imaging spectrometer,” in *Proc. Tiangong-2 Remote Sens. Appl. Conf.*, 2018, pp. 71–81.

[137] K. Liu, B. Qin, and S. Li, “The application of the Tiangong-2 wide-band imaging spectrometer data in the ecological environment evaluation—A case study of Kunming,” in *Proc. Int. Workshop Environ. Geosci.*, 2018, pp. 469–476.

[138] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

Zhuang Zhou received the B.Eng. degree in electrical engineering and automation from the China University of Mining and Technology, Xuzhou, China, in 2013, and the M.S. degree in cartography and geography information system from Beijing Normal University, Beijing, China, in 2016.

He is currently an Engineer with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China. His research interest includes remote sensing image classification.

Shengyang Li received the M.Eng. degree in computer science and technology from the Shandong University of Science and Technology, Qingdao, China, in 2003, and the Ph.D. degree in remote sensing image processing and analysis from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2006.

He is currently a Professor with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences. His research interests include computer vision, target detection and tracking in video satellite, and remote sensing image analysis and understanding.

Wei Wu received the Ph.D. degree in cartography and geography information system from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2019.

She is currently an Assistant Researcher with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China. Her research interests include remote sensing image intelligent recognition, big data mining, and analytics.

Weilong Guo received the B.Eng. degree in software engineering from Jilin University, Jilin, China, in 2018. He is currently working toward the M.Eng. degree with the School of Artificial Intelligence, Chinese Academy of Sciences, Beijing, China.

His research interests include intelligent analysis and understanding of image and video.

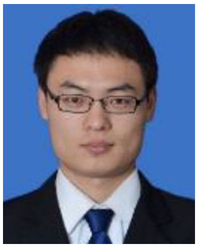
Xuan Li received the B.Eng. degree in electronic and information engineering and the D.Eng. degree in information and communication engineering from the Changchun University of Science and Technology, Changchun, China, in 2011 and 2017, respectively.

He is currently an Engineer with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences, Beijing, China. His research interests include remote sensing image target recognition.

Guisong Xia (Senior Member, IEEE) received the Ph.D. degree in image processing and computer vision from CNRSLTCI, Télécom ParisTech, Paris, France, in 2011.

From 2011 to 2012, he was a Postdoctoral Researcher with the Centre de Recherche en Mathmatiques de la Decision, CNRS, Paris Dauphine University, Paris, France. Since 2018, he has also been a Visiting Scholar with the Department of Mathematics and Applications (DMA), École Normale Suprieure (ENS-Paris), Paris, France. He is currently a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) and the School of Computer Science, Wuhan University, Wuhan, China. His research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing imaging.

Dr. Xia is a member of the Editorial Board for the journals *Pattern Recognition*, *Signal Processing: Image Communications*, and the *EURASIP Journal on Image and Video Processing*.



Zifei Zhao received B.Eng. and M.Eng. degrees in photogrammetry and remote sensing from the Shandong University of Science and Technology, Qingdao, China, 2015 and 2018, respectively. He is currently working toward the Ph.D. degree with the Technology and Engineering Center for Space Utilization, Chinese Academy of Sciences (CAS), Beijing, China.

His research interests include remote sensing image processing and satellite video analysis, such as object detection.