# Object-Guided Remote Sensing Image Scene Classification Based on Joint Use of Deep-Learning Classifier and Detector

Xiaoliang Yang, Weidong Yan, Weiping Ni, Xifeng Pu, Han Zhang, and Maoyu Zhang

***Abstract*—Due to the extremely complex composition of remote sensing scenes, REmote Sensing Image Scene Classification (RESISC) is still a challenging task. To further improve classification accuracy, this article introduces a deep-learning detector into RESISC and proposes to classify remote sensing images according to the detected class-specific signature objects. Inspired by the classification procedure of human vision system, we design a classification framework that utilizes class-specific signature objects of scene classes to guide scene classification. When performing image classification, the proposed framework first adopts a deep-learning classifier to create an initial judgment of the scene class for an image and then determines the scene class based on the class-specific signature objects detected from the image. The proposed method can compete with the state-of-the-art methods on three RESISC benchmark datasets, including NWPU-RESISC45, AID, and OPTIMAL-31.**

***Index Terms*—Class-specific signature object, deep-learning detector, remote sensing image, scene classification.**

## I. Introduction

WITH the rapid development of remote sensing technology, the automation level of remote sensing image interpretation is constantly improving. REmote Sensing Image Scene Classification (RESISC), which aims to label remote sensing images covering multiple land-cover types or ground objects with semantic classes, is a fundamental step in automatic interpretation of remote sensing images. During past few years, there has been increased interest in RESISC owing to its wide applications, such as land resource management, urban planning, and environmental protection [1]–[3], etc.

The existing scene classification methods are mainly carried out in the feature space. Through extracting the designed global and local features, the image is represented as a feature vector in the classifier. In the early 1970s, methods with human-engineering features are mainly adopted in RESISC. To continually improve classification performance, significant efforts have been made to design human-engineering features. The proposed human-engineering features have developed from the initial low-level features [4]–[6] to the mid-level features [7]–[11]. It is worth noticing that scene classification methods with human-engineering features perform well on some scenes with uniform structures and spatial arrangements, but it is difficult for them to depict the high diversity and the nonhomogeneous spatial distributions in remote sensing images [12]. In addition, these methods require a considerable amount of feature engineering skills, which may lead to the lack of the flexibility and adaptivity to different scenes.

Currently, deep learning has brought a revolution in computer vision. The deep convolutional neural network (CNN), which is acknowledged as the most successful and widely used deep-learning approach, has dramatically improved the state-of-the-art in natural image classification, object detection, and visual object recognition [13]. Since deep learning requires very little feature engineering by hand and can take advantage of increases in the amount of available data, it is now the dominant approach for almost all classification and detection tasks. Especially, many recent works have demonstrated that fine-tuning is an effective way to accomplish many other new classification tasks with limited training data. Based on this, researchers have investigated transferring CNNs for RESISC. Compared to conventional methods with human-engineering features, deep-learning methods achieve far better performance (e.g., almost 100% classification accuracy on the most popular UC-Merced dataset [14] with deep ConvNets features [15]).

Due to the high intraclass diversity and low interclass variation of remote sensing scenes, RESISC is still a challenging task. As reported in [16]–[19], since large-scale datasets have rich image variations, deep-learning methods that directly fine-tune the existing CNNs on RESISC dataset suffered an accuracy degradation. This is mainly because they only utilize the feature from the last layer of CNN to classify images and ignore the features from different hierarchical layers of CNN. To enhance the performance, multilayer convolutional features of CNN are adopted [20], [21]. These methods treat the multilayer convolutional features as equally important, which may also bring some interference information, such as feature redundancy and mutual exclusion among the convolutional features.

The above-mentioned CNN methods [16]–[21] tend to generate a global representation of image with the same contribution of each part, in spite of the negative effects of redundant regions. It is obvious that not all regions of the image are useful for the scene classification. To this end, we investigate utilizing class-specific signature objects to guide

Manuscript received December 26, 2019; revised March 24, 2020 and April 22, 2020; accepted May 9, 2020. Date of publication May 28, 2020; date of current version June 11, 2020. *(Corresponding author: Xiaoliang Yang.)*

The authors are with the Remote Sensing Data Analysis Laboratory, Northwest Institute of Nuclear Technology, Xi'an 710024, China (e-mail: yangxiaoliang@nint.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2020.2996760

scene classification. A novel RESISC method based on the joint use of a deep-learning classifier and a detector is proposed. The idea behind our method is to focus on the critical parts of images and abandon the useless ones.

To realize this idea, we first create an initial judgment of the scene class for the image by using an individual CNN classifier to predict the probability order of possible scene classes. Then, the scene class is determined based on the signature objects detected by a deep-learning detector. To solve the issue that how to alleviate the performance degradation arising from false positives when using the detected signature objects to help classification, a classification strategy that jointly utilizes the results of the initial classification and the detection is also designed.

The main advantage of our method is that it can achieve high classification accuracy with a conciseness implementation. The proposed method is actually a combination of two existing CNNs. Implementing the combination of two existing networks is much easier than designing complicated CNN architectures. In addition, the proposed method has high extendibility. Both the classifier and the detector in our method are not fixed, and therefore they can be replaced by other better CNN architectures whenever necessary. Extensive experiments demonstrate that the proposed method obtains state-of-the-art results on three RESISC benchmark datasets (NWPU-RESISC45, AID, and OPTIMAL-31).

The remainder of this article is organized as follows. Section II reviews the related studies of scene classification methods. A brief review of deep-learning detection methods is also given. In Section III, the proposed scene classification method is presented. Section IV presents the experimental results and analysis. Finally, we draw conclusion for this article in Section V.

## II. Related Work

According to the type of used features, the existing RESISC methods can be divided into two categories, methods with human-engineering features and methods with machine learning.

### A. *Scene Classification Methods With Human-Engineering Features*

For this kind of methods, local or global low-level features such as color [22], texture [23], and structure information or their combination are mainly used [6], [24]. Among these features, the scale-invariant feature transform [25] (SIFT) is the most widely used feature. It is a local feature that describes the local variations of complex structures in remote sensing images. Based on the comparison of SIFT and Gabor texture features for classifying the IKONOS satellite images, Yang and Newsam [4] found that SIFT performs better. Using combinations of complementary features is another effective way to improve the scene classification results. For instance, Luo *et al.* [6] jointly used six different kinds of feature descriptors (i.e., simple radiometric features, Gaussian wavelet features, Gray level co-occurrence matrix, Gabor filters, shape features, and SIFT) for indexing remote sensing images and the experimental results indicate that multiple features can give a better description of images. Yu *et al.* [26] proposed a color-texture-structure descriptor which contains the spectral, textural, and structural information. Although these methods are effective on some scenes with uniform spatial distributions, it is difficult for them to properly describe complex remote sensing images.

To fill up the semantic gap, the Bag-of-words (BoW) model, which is originally developed for text analysis, is applied into RESISC. The BoW model constructs a holistic scene representation with three steps. First, the local features of the image are extracted. Second, the visual codebook is generated through encoding each extracted feature to its nearest visual word with a clustering method (e.g., *k*-means clustering). Third, the image is represented by a BoW vector which can be regarded as a histogram counting the occurrences of each visual word. The BoW provides a discriminative representation of image, and it is a leading strategy in the last decade [8]–[10], [14]. To achieve more discriminative feature representation of the BoW model, extensive studies have been carried out. Chen and Tian [9] proposed a translation and rotation-invariant pyramid-of-spatial-relations model to describe both relative and absolute spatial relationships of local features. Researchers in [27] used a soft-assignment to smoothly distribute the features to the codewords. To overcome the limitation that the BoW model lacks spatial information, Lazebnik *et al.* proposed the spatial pyramid matching (SPM) scheme [28]. The SPM uses multiscale spatial average pooling to form the image feature vector and brings a substantial gain in classification accuracy.

To achieve better classification performance, probabilistic topic models are also employed in RESISC [29], [30]. They model each image as a mixture of topics and determine the scene label to an image based on its topic distribution. In [31], LDA is used to model each geo-category in a supervised framework. Yi *et al.* [32] presented a semantic clustering algorithm for high-resolution remote sensing images based on the PLSA. In [33], an automatic framework that combines a probabilistic topic model with a multiscale image representation is proposed for semantic clustering of geo-objects.

The methods with human-engineering features have made significant improvements on the scene classification accuracy. However, due to the limit representative ability, the human-engineering features are difficult to depict the high-diversity and the nonhomogeneous spatial distributions in remote sensing images.

### B. *Scene Classification Methods With Machine Learning*

Instead of using human-engineering features, scene classification methods with machine learning use features that are automatically learned from the image dataset. Autoencoder [34] is a representative method that has been successfully applied to RESISC. It is an unsupervised feature learning method that can automatically learn features from unlabeled samples. A typical autoencoder achieves feature representation of the image through minimizing the reconstruction error between the input data at the encoding layer and its reconstruction at the decoding layer. Zhang *et al.* [35] improved the classification performance through training a sparse autoencoder on the image patches sampled by their saliency degree. Othman *et al*. [36] also used a sparse autoencoder for feature representation. These autoencoders do not make full use of semantic information contained in category labels, and therefore, the performance improvement of the autoencoder is limited.

In recent years, deep learning has attracted great attention of the remote sensing community. Deep convolutional networks, which are designed to process data that come in form of multiple arrays, have dramatically improved the state-of-the-art in natural image classification [37]–[39]. Deep convolutional networks can learn high-level features using a general-purpose learning procedure. For classification tasks, high-level features amplify aspects of the input that are important for discrimination and suppress irrelevant variations. Due to the success, researchers have transferred CNNs to RESISC.

A direct way to utilize CNNs is to fine-tune the existing CNNs for RESICSC [16]–[19]. In [16] and [17], researchers fine-tuned the existing AlexNet, GoogLeNet, and VGGNet on large-scale datasets NWPU-RESISC45 and AID, respectively. Liang *et al*. [19] proposed a transfer learning scheme to fine-tune the existing CNNs. However, these methods ignored the features from different hierarchical layers.

Taking the pretrained CNNs as fixed feature extractor is also effective. Penatti *et al*. in [15] achieved the state-of-the-art scene classification result by using CNN features from fully connected (FC) layers. In contrast to [15], Hu *et al*. [20] evaluated the CNN features from FC layers but also from convolutional layers. Researchers in [21] demonstrated that incorporating different levels' convolutional features can improve the classification accuracy. In [40], a novel feature representation method, named Bag of Convolutional Features (BoCF), was proposed for RESISC. To make the visual words have more semantic properties, BoCF generates visual words with deep convolutional features. Researchers in [41] incorporated the CNN features of top layers into features of bottom layers to improve the representation for small objects.

Recently, a mass of works focus on designing new CNN architectures [42]–[45]. In [42], researchers created an ensemble of CNNs (named Hydra) for RESISC. Their method improved the classification accuracy by fine-tuning hydra's heads multiple times. Through exploring the attention mechanism, Wang *et al*. [43] designed a recurrent attention structure and proposed an attention recurrent convolutional network (ARCNet) for scene classification. Researchers in [44] proposed a gated bidirectional network (GBNet) to hierarchically aggregate multilayer convolutional features and enhance the complementary information. To boost the performance of RESISC, Cheng *et al*. [45] learned discriminative CNNs (D-CNNs) through adding a metric learning regularization term on the objective function. These methods usually need a lot of network designing skills.

### *C. Deep-Learning Detection Methods*

The development of deep learning promotes the breakthrough on the task of object detection. Currently, most object detectors are based on deep learning. Apart from its standalone utility, deep-learning detector also provides a useful building block for larger systems that employ an object detection component. In this article, we introduce a deep-learning detector into scene classification to detect class-specific signature objects in remote sensing images. The details of our method will be presented in Section III. Here we briefly review the existing deep-learning detection methods.

Generally there are two types of deep-learning detection frameworks. The first type is the two-stage framework, which generates the region proposals first and then classifies each proposal to different object classes. R-CNN [46], SPP-net [47], Fast R-CNN [48], Faster R-CNN [49], R-FCN [50], FPN [51], and MASK R-CNN [52] are the representative two-stage detectors. Especially R-CNN, which obtained a mean average precision (mAP) of 53.3% with more than 30% improvement over the previous best result (DPM [53]) on PASCAL VOC 2012 [54], has epoch-making significance in the field of object detection. The flowchart of R-CNN mainly has three steps: region proposal generation based on selective search [55], CNN-based deep feature extraction, classification and localization. The reason for the success of R-CNN lies in that it turns the object detection into classification of region proposals. In R-CNN, SPP-net, and Fast R-CNN, region proposal computation is a bottleneck in improving efficiency. To solve this problem, Ren *et al*. [49] proposed Faster R-CNN, which utilizes a region proposal network. With the proposal of Faster R-CNN, the two-stage detection frameworks can really be trained in an end-to-end way. Compared to Faster RCNN, FPN [51] modifies the backbone network by adding top-down and lateral connections to build a feature pyramid that facilitates end to end learning across different scales.

To overcome the speed limit of the two-stage framework, the one-stage framework, which maps straightly from image pixels to bounding box coordinates and class probabilities based on global regression, has been proposed. YOLO [56] is one of the representative one-stage detectors. It uses the whole topmost feature map to predict both confidences for multiple categories and bounding boxes. However, the detection accuracy of YOLO is not high because of its difficulty in dealing with small objects in groups. To achieve better detection accuracy and speed, Liu *et al*. [57] proposed a Single Shot MultiBox Detector (SSD) for multiple categories, whose key feature is the use of multiscale convolutional bounding box outputs attached to multiple feature maps at the top of the network. Compared to YOLO, SSD is significantly more accurate on PASCAL VOC and Microsoft COCO [58]. By adopting better feature extractor backbone (e.g., ResNet101), adding deconvolution layers with skip connections to introduce additional large-scale context [59], and designing better network structure (e.g. Stem Block and Dense Block) [60], the problem that SSD is not skilled at dealing with small objects can be relieved.

## III. Proposed Method

According to the results reported in [43]–[45], although CNN-based methods have achieved high accuracies on the UC-Merced and WHU-RS19 datasets, their overall accuracies on large-scale remote sensing scene classification datasets, such as

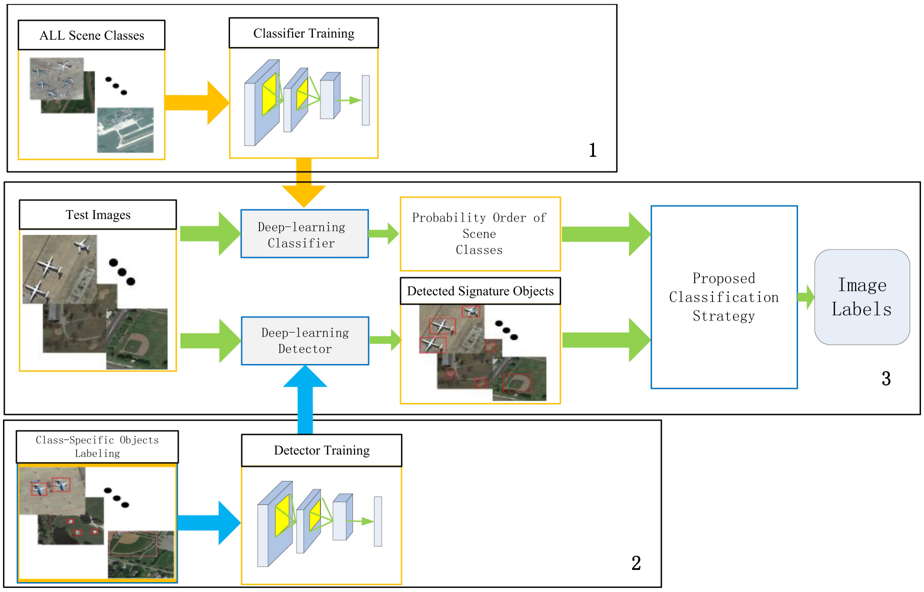


Fig. 1. Overall framework of the proposed method.

NWPU-RESISC45, and AID, are still not saturated. To further improve scene classification accuracy, we investigate utilizing class-specific signature objects to guide scene classification and develop an object-guided method based on using the off-the-shelf deep CNN models. Fig. 1 illustrates the overall framework of the proposed method.

As can be seen from Fig. 1, we first create an initial judgment of the scene class for the image, which is conducted by using an individual CNN classifier to predict the probability order of possible scene classes. Then, the scene class is determined based on the class-specific signature objects detected by a deep-learning detector. The proposed framework is mainly composed of three stages: deep-learning classifier training, deep-learning detector training, and scene classification based on a designed classification strategy.

## *A. Deep-Learning Classifier Training*

In the first stage, a deep-learning classifier is trained to create an initial judgment of the scene class for the image. As shown in Fig. 1, the deep-learning classifier plays a foundation role in our framework. On one hand, it helps to select scene classes that need participation in the second stage: deep-learning detector training. According to the classification accuracy of the deep-learning classifier on the validation set, we set a threshold of classification accuracy and select the scene classes whose accuracies are smaller than the threshold. In general, if a scene class obtains high classification accuracy, it will not easily confuse with other scene classes. Therefore, scene classes with high classification accuracy are not selected to train the detector. Since fewer scene classes participate in the second stage, the burden of labeling images for training the detector can be alleviated.

On the other hand, the deep-learning classifier will directly determine the overall accuracy of the proposed method: 1) for the scene classes that do not participate in training the detector, their classification accuracies are only determined by the classifier; 2) for the classes participated in training the detector, assigning scene labels also depends on the probabilities predicted by the classifier.

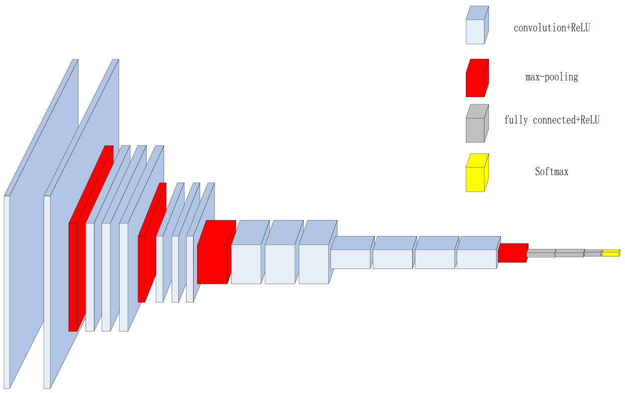


Fig. 2. Architecture of VGGNet-16.

In the proposed framework, a variety of CNNs (such as AlexNet, VGGNet, and DenseNet, etc.) can be adopted as the classifier. To illustrate this issue, VGGNet-16 is first tested. The reason for choosing VGGNet-16 is that fine-tuned VGGNet-16 has achieved good performance on NWPU-RESISC45 and AID.

As shown in Fig. 2, VGGNet-16 is a deep CNN model that consists of 13 convolutional layers and 3 FC layers. The first few stages of VGGNet-16 are composed of two types of layers: convolutional layers and pooling layers. The following stages are composed of FC layers, and the last FC layer is a softmax layer that computes the scores for each defined class.

To further enhance the classification performance, we also test DenseNet, as it outperformed other individual CNN architectures (such as VGGNet and ResNet) on NWPU-RESISC45 [42]. As shown in Fig. 3, DenseNet builds a dense block to

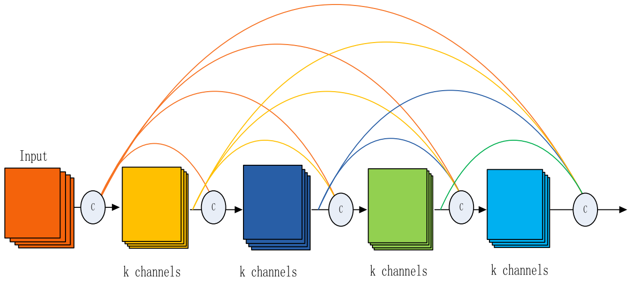


Fig. 3. Architecture of a dense block.

introduce direct connections between any two layers with the same feature-map size. Through using the dense block, the feature propagation is strengthened.

When training the classifier, the online data augmentation is applied. It randomly transforms the samples every epoch during training. Following work [42], we also use random flips in vertical and horizontal directions, random zooming and random shifts over different image crops.

## B. Deep-Learning Detector Training

In the second stage, we train a deep-learning detector to detect class-specific signature objects in remote sensing images. Although the deep-learning classifiers have strong classification capabilities, they would still get undesirable results when most of regions of the two images are similar whereas class-specific signature objects occupy only a small region.

For example, Fig. 4 shows three groups of remote sensing images from AID dataset. Due to high intraclass diversity and low interclass variation of the scene images, a deep-learning classifier may confuse the two images in each group. However, most humans would easily classify each image into correct scene class according to the class-specific signature objects in the image. This example suggests that class-specific signature objects are helpful for scene classification. Motivated by this, we train a deep-learning detector to detect class-specific signature objects.

In the proposed framework, which objects can be defined as the signature objects are determined based on human experience. Compared to designing human-engineering features, determination of signature objects for a scene class is easier and does not require considerable domain expertise. This is because determination of signature objects takes advantage of human ingenuity and makes full use of semantic information contained in scene labels. For example, if a person is required to label *airport* images, it would be a natural thought to find airplanes and runways in the images. However, it is difficult for a machine to imitate this determination process.

Choosing an appropriate deep-learning detector is important to achieve good detection performance. When choosing the detector, the following two features of remote sensing images should be taken into consideration. First, the background may occupy a large area in a remote sensing image, whereas the class-specific signature objects occupy only a small portion of the entire image. Second, class-specific signature objects will exhibit different characteristics at different scales.

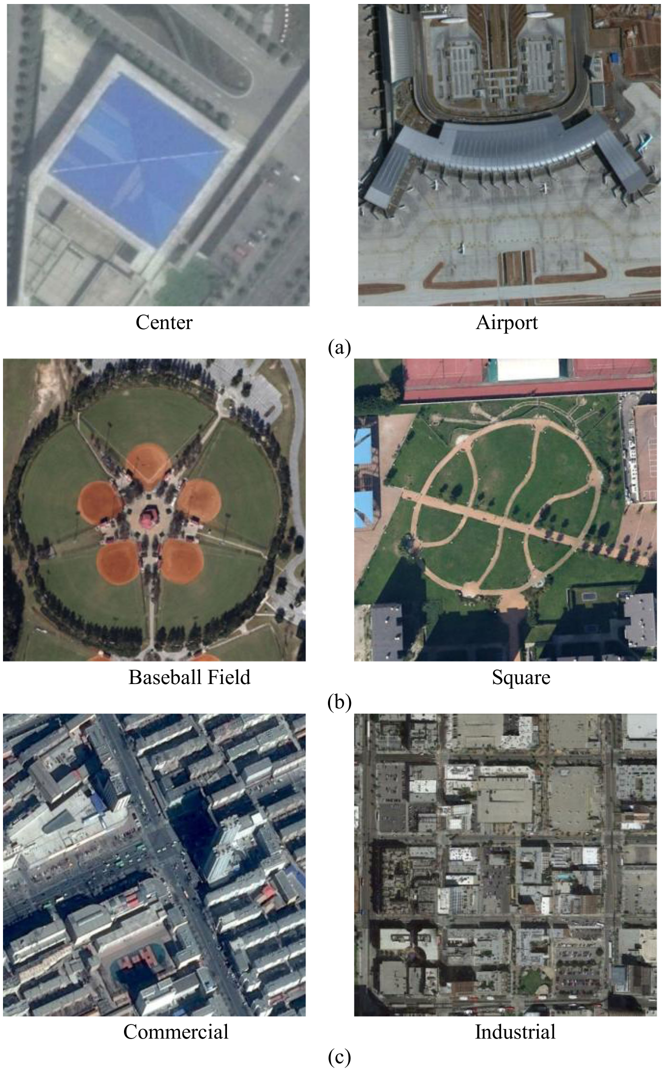


Fig. 4. Spatial and structural complexity of remote sensing images. (a) Similar objects between different scenes. (b) Similar textures between different scenes. (c) Similar structural distributions between different scenes.

Compared to the one-stage detectors, the two-stage detectors usually perform better [61]. In this article, we adopt FPN (ResNet101) [51] as the deep-learning detector because it is robust to scale changes and has higher detection quality (mAP) than Faster R-CNN and R-FCN. The architecture of this detector is shown in Fig. 5. FPN uses a pyramidal representation and combines features of shallow layers with deeper layers to extract rich semantics from all levels. It can obtain state-of-the-art representation without sacrificing speed and memory.

As is known, the scene classification dataset only provides image-level supervision information, whereas training the detector requires the instance-level supervision information. Therefore, we manually label the signature objects for training the detector. In the proposed scheme, the image-level supervision information is employed to help us label the signature objects, and the label of each class-specific object is named as SceneClass_ObjectName.

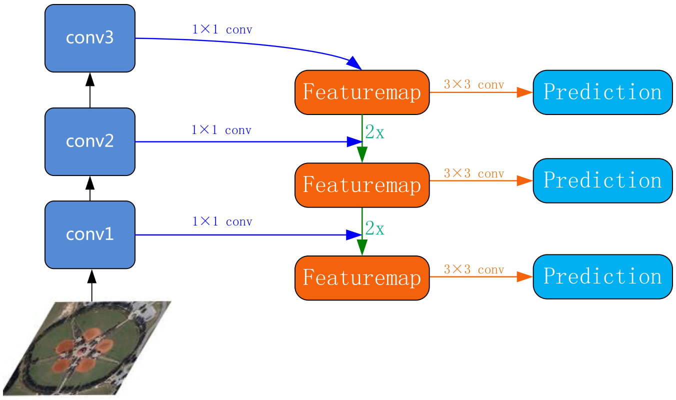


Fig. 5. Architecture of FPN.

Through using the end-to-end detector, we transform scene classification task from feature engineering to sample engineering, so that researchers can get rid of the complex feature design and focus on determining which objects can be defined as class-specific signature objects for a scene class.

### C. Scene Classification Based on the Designed Classification Strategy

In the third stage of the proposed framework, we propose a new classification strategy to determine the scene label of a test image. The basic idea for the proposed classification strategy is that the detected class-specific signature objects should be consistent with the classification result. For example, if a test image is classified to *basketball court*, then it is very possible to detect basketball courts from the image.

A simple classification strategy is to determine the scene label as the category with the biggest number of class-specific signature objects. A drawback of this strategy is that it does not take the false positives produced by the detector into consideration. Due to the limited training samples, the detector may produce a few false positive, which would significantly degrade the scene classification accuracy of this strategy. To alleviate the problem arising from the false positives, we propose to jointly use the initial classification result of the deep-learning classifier and the detection result to determine the scene class of an image.

In the proposed strategy, we consider three situations of the detection result. First, if the detected objects all belong to one scene class, then the test image is assigned to the label of that scene class. The second situation is that the detected objects belong to multiple scene classes. In this situation, we use the deep-learning classifier to predict the probability for every scene class. According to the probability order, we search for class-specific signature objects belonging to each scene class in turn. Once the class-specific signature objects of a scene class are found, the search is stopped and the test image is classified into that scene class. Through utilizing the probability order, the wrong classifications caused by the false-positives can be alleviated. Third, if there are no class-specific signature objects detected, then the test image is assigned to the label of the deep-learning classifier with the highest response.

The scheme of the proposed classification strategy can be summarized as follows.

1) Assume that the probability order of the scene classes predicted by the deep-learning classifier is $x_1 > x_2 > x_3 \ldots > x_M$, where $M$ is the number of scene classes selected to train the deep-learning detector.
2) Count the number of class-specific signature objects belonging to each scene class. Assume these numbers are $n_1, n_2, n_3, \ldots, n_M$, respectively.
3) If $n_1 + n_2 + n_3 + \ldots + n_M = 0$, which means that there is no class-specific signature objects detected in the test image, then the image is assigned to the label of the deep-learning classifier with the highest response.
4) If detected objects all belong to one scene class, which means only one of the $\{n_1, n_2, n_3, \ldots, n_M\}$ is greater than 0, assume it is $n_k$, $k \in \{1, 2, 3, \ldots, M\}$, then the test image is assigned to the label of the $k$th scene class.
5) If detected objects belong to multiple scene classes, from $n_1$ to $n_M$, assume $n_k$ is the first number greater than 0, then the test image is assigned to the label of the $k$th scene class.

From this scheme, we can see that the key feature of this strategy is that it imitates the classification procedure of human vision system to some extent. It first uses a classifier to give a coarse classification for the image, and then the scene class of the image is finally determined based on the detected class-specific signature objects. Moreover, this classification strategy can be used without training, which reduces the computational burden of the proposed classification method.

## IV. Experiment and Analysis

In this section, we evaluate the proposed method on three challenging datasets: NWPU-RESISC45, AID, and OPTIMAL-31.

### A. Datasets Descriptions

1) *NWPU-RESISC45:* The NWPU-RESISC45 dataset is a publicly available benchmark, which contains 31 500 images, covering 45 scene classes. Each class includes 700 images with a size of each 256 × 256 pixels in the red-green-blue (RGB) color space. The spatial resolution varies from about 30 to 0.2 m per pixel. Fig. 6 shows two samples of each class from this dataset. For each scene class in the NWPU-RESISC45 dataset, image variations in translation, spatial resolution, viewpoint, object pose, illumination, background, occlusion, etc., are big. This dataset has high within-class diversity and between-class similarity, which makes it rather challenging for scene classification.
2) *AID:* The AID dataset is another publicly available large-scale dataset for RESISC, which contains 10 000 images collected from Google Earth imagery, covering 30 scene classes. The numbers of sample images vary from 220 to 420 with a fixed size of 600 × 600 pixels. The spatial resolution varies from about 8 to 0.5 m per pixel. As same as NWPU-RESISC45 dataset, the images in AID are multisource, as Google Earth images are from different

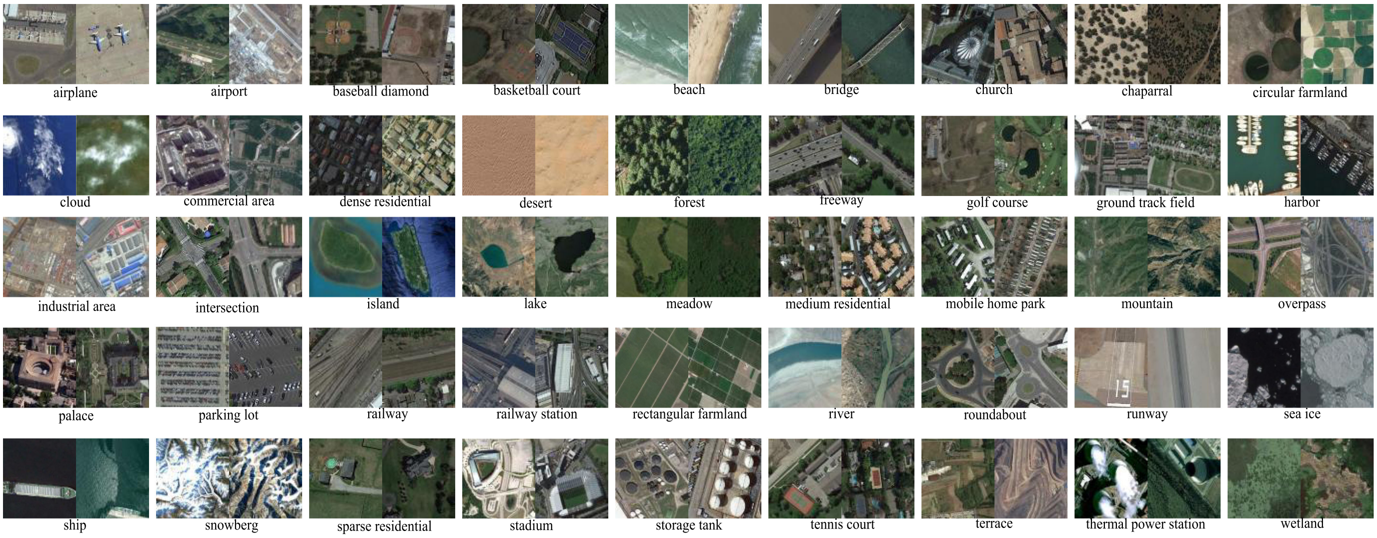


Fig. 6. Samples of NWPU-RESISC45: two examples of each semantic scene class are shown. There are 31500 images within 45 classes.

remote imaging sensors. This brings more challenges for scene classification than the single-source images like UC-Merced dataset.

3) *OPTIMAL-31:* The OPTIMAL-31 dataset is a recently constructed dataset for RESISC, which contains 1860 images also collected from Google Earth, covering 31 scene classes. Each class includes 60 sample images with a fixed size of 256 × 256 pixels in the RGB color space. Compared to UC-Merced and WHU-RS19, OPTIMAL-31 contains more classes so that it has a higher degree of difficulty.

### B. Parameter Settings

According to the discussion in Section III-A, we adopt VGGNet-16 and DenseNet as the classifier, and named the corresponding method as VGGNet16+FPN, and DenseNet+FPN, respectively. When training the classifier of the proposed framework, we fine-tune VGGNet-16 and DenseNet on each dataset. The VGGNet-16 pretrained on ImageNet dataset is obtained from[1]. The DenseNet-161 pretrained on ImageNet dataset is obtained from[2].

When training the detector of the proposed framework, we fine-tune FPN (ResNet101) on the images of selected scene classes. The FPN (ResNet101) pretrained on Microsoft COCO dataset is obtained from[3]. The detailed parameters used for fine-tuning are summarized in Table I. All experiments were implemented on a PC with 44 2.2 GHz 2core CPUs and 128 GB memory. In addition, a NVIDIA Quadro P6000 GPU was also used for acceleration.

TABLE I
PARAMETERS UTILIZED FOR CNN MODEL FINETUNING

| CNN | Batch Size | Learning Rate | Weight Decay |
|---|---|---|---|
| VGGNet-16 | 50 | 0.001<br>(0.01 last layer) | 0.0005 |
| DenseNet-161 | 15 | 0.0001 | 0.00001 |
| FPN (ResNet101) | 15 | 0.0003 | 0.000015 |

### C. Evaluation Metrics

In the experiments, the metrics of overall accuracy and confusion matrix are used to evaluate all classification methods. The overall accuracy is defined as the number of correctly classified images divided by the total number of test images. It is a direct measure to reveal the classification performance on the whole dataset.

The confusion matrix is an informative table used to analyze all the errors and confusions between different scene classes. It is generated by counting each type of correct and incorrect classification of the test images and accumulating the results in the table. Each item $x_{ij}$ denotes the rate of test samples from the $i$th class that are classified as the $j$th class.

To obtain reliable results for the metrics of overall accuracy and confusion matrix, we repeat the experiment ten times for each training–testing ratio and report the mean and standard deviation of the results.

### D. Ablation Studies

To validate the contributions of each classification branch, we conduct ablation studies. In the ablation experiments, 20% of the images in each RS scene category of NWPU-RESISC45 are randomly selected for training. Since our method mainly contains three branches (deep-learning classifier, deep-learning detector, and the designed classification strategy), we divide ablation experiments into three groups.

The first group is to validate the contribution of the classifier. In this group, VGGNet-16+FPN and DenseNet+FPN are compared. In the second group, DenseNet and DenseNet+FPN are

[1] [Online]. Available: https://github.com/BVLC/caffe/wiki/Model-Zoo
[2] [Online]. Available: https://download.pytorch.org/models
[3] [Online]. Available: https://github.com/jwyang/fpn.pytorch

TABLE II
ABLATION EXPERIMENTS ON NWPU-RESISC45

| Methods | 20% for training |
|---|---|
| VGGNet-16+FPN | 92.02% |
| **DenseNet+FPN** | 95.26% |
| DenseNet | 93.03% |
| **DenseNet+FPN** | 95.26% |
| DenseNet+FPN+LDA | 89.36% |
| **DenseNet+FPN** | 95.26% |

TABLE III
PERFORMANCE COMPARISON OF STATE-OF-THE-ART METHODS ON NWPU-RESISC45

| Methods | Training Ratio | |
|---|---|---|
| | 10% | 20% |
| VGGNet-16 CNN Feature [16] | 76.47±0.18% | 79.79±0.15% |
| BOCF [40] | 82.65±0.31% | 84.32±0.17% |
| Fine-tuned VGGNet-16 [16] | 87.15±0.45% | 90.36±0.18% |
| D-CNN with VGGNet-16) [45] | 89.22±0.50% | 91.89±0.22% |
| Hydra(DenseNet+ResNet) [42] | 92.44±0.34% | 94.51±0.21% |
| VGGNet-16+FPN (ours) | 89.18±0.18% | 92.17±0.15% |
| DenseNet+FPN (ours) | 93.17±0.19% | 95.11±0.14% |

compared to validate the contribution of the detector. In the third group, DenseNet+FPN+LDA and DenseNet+FPN are compared to validate the contribution of the designed classification strategy. DenseNet+FPN+LDA regards the detected objects as the visual words and uses LDA for scene classification.

The experimental results are shown in Table II. From Table II, the following can be observed.

1) The deep-learning classifier directly determines the overall accuracy of the proposed method. Through replacing the VGGNet-16 with DenseNet, the classification accuracy improves about 3.2%.This is mainly because the discriminative ability of DenseNet is stronger than that of VGGNet-16.
2) The proposed method (DenseNet+FPN) makes an increase of 2.23% over the individual DenseNet, which confirms that the detection of signature objects actually makes a great contribution to the promotion of the classification accuracy.
3) In the third group, the LDA-based method (DenseNet+FPN+LDA), which regards the detected objects as the visual words, does not achieve results as good as expected, and performs much worse than the proposed method (DenseNet+FPN). A possible explanation is that the false-positives degrade the discriminative ability of the generated topic model. In contrast, the proposed classification strategy is more appropriate than LDA. To reduce the effect of false-positives, it jointly uses the results of initial classification and the detection to determine the scene class.

### *E. Comparison With the State-of-the-Art Methods*

In this section, performance comparison between our method and some state-of-the-art methods are discussed.

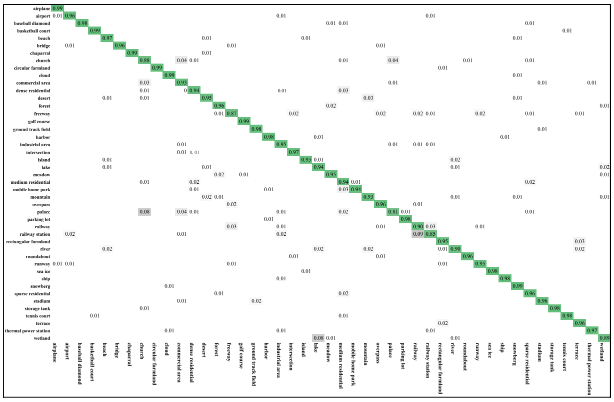


Fig. 7. Confusion matrix on NWPU-RESISC45 under the training ratio of 20%.

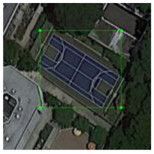
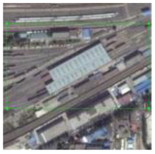
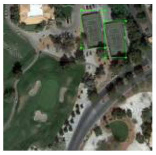
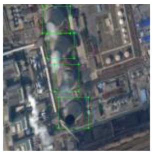
*1) NWPU-RESISC45:* Table III presents the experimental results of the proposed method and the other five comparison methods on NWPU-RESISC45, under the training ratios of 10% and 20%, respectively. It can be observed that the proposed method achieves the best performance for both training ratios. In detail, the proposed method achieves a 93.17 ± 0.19% and 95.11 ± 0.14% accuracy under the 10% and 20% training ratio, respectively. Especially compared to Hydra [42], which recently achieved the state-of-the-art performance for the NWPU-RESISC45, the proposed DenseNet+FPN can make an increase of 0.6%. It is worth noting that the proposed DenseNet+FPN outperforms fine-tuned VGGNet-16, which has been reported as the best scene classification method in [16], by more than 4.5% in accuracy. Fig. 7 shows the confusion matrix of the proposed DenseNet+FPN. It can be seen that the biggest confusion happens between *palace* and *church*, because of their similar global structure and spatial layout.

The main reason for the good performance is that the proposed method uses class-specific signature objects to help scene classification. If signature objects of a scene class are detected in the image, the probability that the image belongs to that class will be greatly increased.

According to the proposed framework, partial scene classes are selected to train the deep-learning detector. Table IV shows the selected classes and the corresponding class-specific signature objects. It can be observed that the majority of the selected classes are artificial scene types. This is mainly because the artificial scene types usually have complex spatial patterns and thus are hard to be distinguished only using the deep-learning classifier. Another reason is that the majority of artificial scene types have class-specific signature objects in their images.

It should be noted that not all artificial scene classes are selected. The reasons are as follows. First, some artificial scene types (e.g., parking lot) are with relatively uniform spatial arrangements and structural distributions, which can lead to high classification accuracies of these classes. Therefore, there is no need to select these scene classes. Second, a little scene types are difficult to define their class-specific signature objects due to their high with-in class diversity. For example, school is not

TABLE IV
SELECTED CLASSES OF NWPU-RESISC45 AND THE CORRESPONDING CLASS-SPECIFIC SIGNATURE OBJECTS

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| airplane | | airport | | baseball diamond | | basketball court |  |
| bridge | | church | | commercial area | | dense residential | |
| golf course | | ground track field | | industrial area | | intersection | |
| lake | | medium residential | | railway station |  | roundabout | |
| runway | | ship | | sparse residential | | stadium | |
| storage tank | | tennis court |  | thermal power station |  | | |

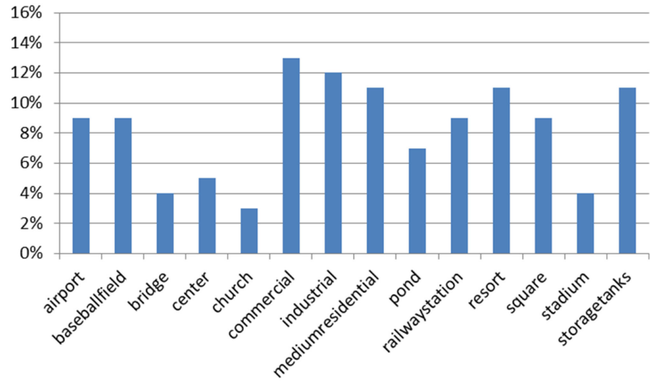


Fig. 8. Accuracy improvement for the selected classes of NWPU-RESISC45.

TABLE V
PERFORMANCE COMPARISON OF STATE-OF-THE-ART METHODS ON AID

| Methods | Training Ratio | |
|---|---|---|
| | 10% | 20% |
| VGGNet-16 CNN Feature [17] | 82.07±0.39% | 85.9±0.29% |
| BOCF [40] | 82.33±0.41% | 86.86±0.32% |
| Fine-tuned VGGNet-16 [17] | 84.83±0.32% | 87.79±0.18% |
| D-CNN with VGGNet-16 [45] | 90.82±0.16% | 96.89±0.10% |
| GBNet+global feature [44] | 92.2±0.23% | 95.48±0.12% |
| Multilevel Fusion CNN [21] | - | 95.36±0.22% |
| VGGNet-16+FPN (ours) | 91.41±0.16% | 96.77±0.12% |
| DenseNet+FPN (ours) | 93.03±0.16% | 97.13±0.11% |

selected because it is difficult and time-consuming for a human with little expertise to find out its class-specific signature objects.

To illustrate the contribution of the deep-learning detector, we compare the proposed VGGNet-16+FPN with Fine-tuned VGGNet-16 on NWPU-RESISC45 under the training ratio of 20%, and list improvements of accuracies for the scene classes that participated in detection (see Fig. 8). As can be seen from Fig. 8, compared to fine-tuned VGGNet-16, the proposed VGGNet-16+FPN obtains an average 3% accuracy improvement and significantly improves performance in these categories: tennis court (90%–>96%), stadium (91%–>96%), airport (86%–>90%), and lake (89%–>93%).

*2) AID:* Table V presents the experimental results of the proposed method and some state-of-the-art comparison methods on AID, under the training ratios of 20% and 50%, respectively. It can be observed that the proposed DenseNet+FPN outperforms other scene classification methods in overall accuracy whether its training ratio is 20% or 50%. Moreover, the proposed VGGNet-16+FPN can also compete with the state-of-the-art

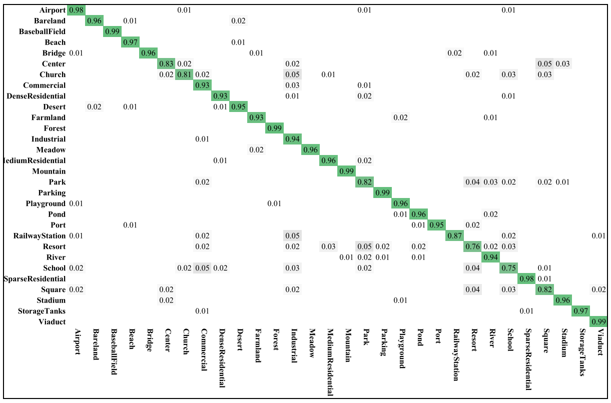


Fig. 9. Confusion matrix on AID under the training ratio of 20%.

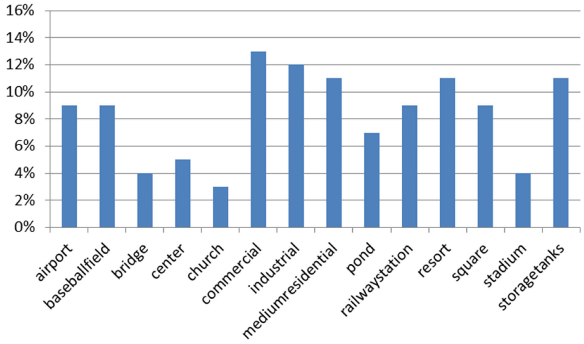


Fig. 10. Accuracy improvement for the selected classes of AID.

TABLE VI
PERFORMANCE COMPARISON OF STATE-OF-THE-ART METHODS ON OPTIMAL-31

| Methods | Training Ratio | |
|---|---|---|
| | 10% | 20% |
| Fine-tuned VGGNet-16 | 87.45±0.45% | |
| ARCNet-ResNet34 [43] | 91.28±0.45% | |
| ARCNet-VGGNet16 [43] | 92.70±0.35% | |
| GBNet+global feature [44] | 93.28±0.27% | |
| VGGNet-16+FPN (ours) | 93.77±0.27% | |
| DenseNet+FPN (ours) | 95.23±0.25% | |

methods. As AID is less challenging than NWPU-RESISC45, the results are higher and the improvements are lower.

We also make a confusion matrix to further analyze the effect of DenseNet+FPN, as shown in Fig. 9. For this dataset, the most notable confusion of the proposed DenseNet+FPN happens between *school* and *commercial*, which may be caused by the densely distributed tall buildings in these scenes.

The improvements of accuracies for the scene classes that participated in detection on this dataset are shown in Fig. 10. It can be seen that, compared to sine-tuned VGGNet-16, the proposed VGGNet-16+FPN obtains an average 8% accuracy improvement and significantly improves performance in these categories: commercial (82%–>95%), industrial (82%–>94%), resort (64%–>75%), and storage tanks (83%–>94%).

*3) OPTIMAL-31:* To further confirm the effectiveness of the proposed method, we evaluate it on OPTIMAL-31, under the training ratio of 80%. As we can be seen in Table VI, in agreement with the results of previous two experiments, the proposed method has the highest accuracy. The proposed DenseNet+FPN obtains a 95.23 ± 0.25% classification accuracy, which is 1.46% higher than that of GBNet+global feature. In addition, the proposed VGGNet-16+FPN takes the second place. This proves that the detection of signature objects can indeed improve the classification accuracy in the domain of remote sensing scene classification. The confusion matrix of DenseNet+FPN is shown in Fig. 11. It can be seen that, misclassification appears especially in these two categories: *freeway* and *overpass*, whose land-cover units are very similar.

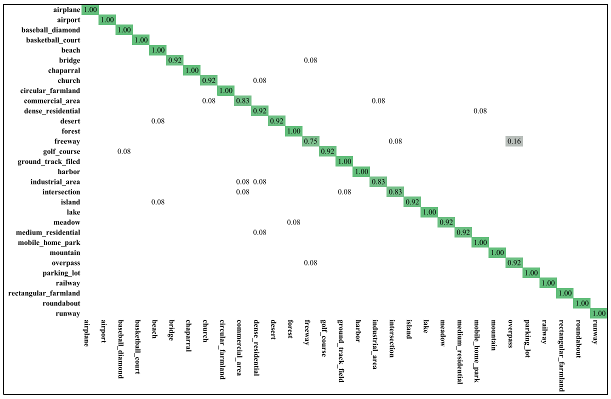


Fig. 11. Confusion matrix on OPTIMAL-31 under the training ratio of 80%.

TABLE VII
TRAINING AND TESTING TIME OF DIFFERENT METHODS ON THREE DATASETS

| | Fine-tuned VGGNet-16 | | VGGNet-16+FPN (ours) | | DenseNet+FPN (ours) | |
|---|---|---|---|---|---|---|
| | Train.(m) | Test.(ms) | Train.(m) | Test.(ms) | Train.(m) | Test.(ms) |
| NWPU-RESISC45 (20%) | 529.2 | 27 | 1684.2 | 71 | 1816.5 | 78 |
| AID (20%) | 294.7 | 66 | 784.67 | 175 | 836.7 | 191 |
| OPTIMAL-31 (80%) | 122.5 | 26 | 395.25 | 70 | 427.8 | 76 |

## *F. Training and Testing Time*

In this section, we analyze the computational efficiency of the proposed method. The training time and testing time is reported in Table VII. All networks are trained with 100 epochs on each dataset. The testing time indicates the average time taken by processing one image. To clearly illustrate the time cost induced by the target detection, we make a comparison between fine-tuned VGGNet-16 and the proposed method (VGGNet-16+FPN). As shown in Table VII, the target detection leads to a considerable increase in time cost, both the training time and the testing time of our method are more than those of fine-tuned VGGNet-16. This deficiency may limit its use in real-time classification work. In our future work, we plan to fuse the deep-learning classifier and detector together to develop a new classification network, which will make the method more efficient.

## V. Conclusion

In this article, an object-guided RESISC method based on the joint use of a deep-learning classifier and a detector is proposed. The framework of this method imitates the classification procedure of human vision system, which first gives a coarse classification of the image using a deep-learning classifier, and then determines the scene class of the image based on class-specific signature objects detected in the image. The proposed method achieves results comparable to the state-of-the-art on three publicly available RESISC datasets, including NWPU-RESISC45, AID, and OPTIMAL-31. The main advantage of the proposed method is its conciseness in implementation. Implementing this method is much easier than duplicating complicated CNN architectures.

In the proposed framework, which objects can be defined as signature objects for a specific class are determined based on the human experience. To further reduce the manual workload, we look forward to combining CNNs with RNNs to automatically decide which objects need to be detected in future studies.

## References

[1] 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.

[2] S. 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, May 2011.

[3] 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.

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

[5] J. dos Santos, O. Penatti, and R. da Silva Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification," *in Proc. 5th Int. Conf. Comput. Vis. Theory Appl.*, May 2010, pp. 203–208.

[6] 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.

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

[8] 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, Mar. 2014.

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

[10] 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, Feb. 2015.

[11] 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.

[12] 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.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

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

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

[16] 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, Apr. 2017.

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

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

[19] Y. Liang, S. Monteiro, and E. Saber, "Transfer learning for high resolution aerial image classification," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, Oct. 2016, pp. 1–8.

[20] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, no. 11, pp. 14680–14707, Nov. 2015.

[21] Y. Yu and F. Liu, "Aerial scene classification via multilevel fusion based on deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 287–291, Feb. 2018.

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

[23] B. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[24] 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.

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

[26] H. Yu, W. Yang, G.-S. Xia, and G. Liu, "A color-texture-structure descriptor for high-resolution satellite image classification," *Remote Sens.*, vol. 8, no. 3, pp. 259–271, Mar. 2016.

[27] J. Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[28] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2169–2178.

[29] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 993–1022, Jan. 2003.

[30] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Aug. 1999, pp. 50–57.

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

[32] W. Yi, H. Tang, and Y. Chen, "An object-oriented semantic clustering algorithm for high resolution remote sensing images using the aspect model," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 3, pp. 522–526, May. 2011.

[33] H. Tag *et al.*, "A multi-scale latent Dirichilet allocation model for object-oriented clustering of VHR panchromatic satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1680–1692, Mar. 2013.

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

[35] 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.

[36] 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, Apr. 2016.

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

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

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

[40] 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, Aug. 2017.

[41] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "RON: Reverse connection with objectness prior networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5244–5252.

[42] R. Minetto, M. 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.

[43] 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.

[44] 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.

[45] 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.

[46] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014. pp. 580–587.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[48] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Conf. Neural Inform. Process. Syst.*, Dec. 2015, pp. 1–8.

[50] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Conf. Neural Inf. Process. Syst.*, Dec. 2016, pp. 1–9.

[51] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 936–944.

[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.

[53] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[54] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 8, no. 2, pp. 303–338, Jun. 2010.

[55] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

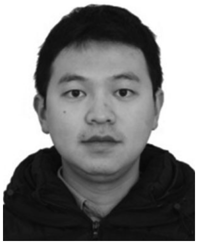
[57] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.

[58] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.

[59] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A. Berg, "DSSD: Deconvolutional single shot detector," Jan. 2017. [Online]. Available: https://arxiv.org/abs/1701.06659

[60] Z. Shen, Z. Liu, J. Li, Y. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1937–1945.

[61] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neur. Netw. Learn.*, vol. 14, no. 8, pp. 1–21, Mar. 2017.

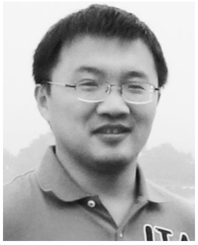


**Xiaoliang Yang** received the B.S. degree from Wuhan University, Wuhan, China, in 2010, and the M.S. and Ph.D. degrees from National University of Defense Technology, Changsha, China, in 2013 and 2017, respectively.

He is currently working as a Research Associate with Northwest Institute of Nuclear Technology, Xi'an, China. His research interests include remote sensing image analysis, InSAR data processing, and deep learning.

**Weidong Yan** was born in 1967. He received the B.S. and M.S. degrees in electronic engineering from the School of Electrical Engineering, National University of Defense Technology, Changsha, China.

He is currently a Researcher Fellow with the Northwest Institute of Nuclear Technology, Xi'an, China. His research interests include remote sensing image analysis and pattern recognition.

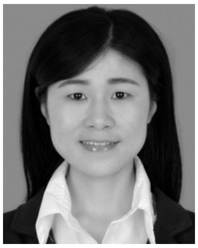


**Weiping Ni** was born in China in 1980. He received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2004, the M.S. degree from National University of Defense Technology Changsha, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent system from Xidian University, Xi'an, China, in 2016.

Since 2014, he has been a Research Associate with the Northwest Institute of Nuclear Technology, Xi'an. His research interests include remote sensing image processing, automatic target recognition, and computer vision.

**Xifeng Pu** received the B.S. degree in mechanical engineering from Xi'an Jiaotong University, Xi'an, China, in 2001, and the Ph.D. degree in safety science and engineering from Beijing Institute of Technology, Beijing, China, in 2013.

He is currently an Associate Professor with the Northwest Institute of Nuclear Technology, Xi'an China. His main research interests include pattern recognition, and remote sensing image analysis.



**Han Zhang** received the B.S. degree in electronics science and technology from Shanghai Jiao Tong University, Shanghai, China, in 2010, and the M.S. degree from National University of Defense Technology, Changsha, China, in 2012.

She is currently working as a Research Associate with Northwest Institute of Nuclear Technology, Xi'an, China. Her research interests include multitemporal and multimodal remote sensing image analysis, pattern recognition, and deep learning.

**Maoyu Zhang** received the B.S. degree in solid physics from Jilin University, Changchun, China, in 1999, and the M.S. degree in electromagnetic and microwave technology from Northwest Institute of Nuclear Technology, Xi'an, China, in 2006.

His research interests include pattern recognition, and deep learning.