Hindawi
Journal of Sensors
Volume 2017, Article ID 1796728, 14 pages
https://doi.org/10.1155/2017/1796728

Research Article

Cascade Convolutional Neural Network Based on Transfer-Learning for Aircraft Detection on High-Resolution Remote Sensing Images

Bin Pan,[1] Jianhao Tai,[1] Qi Zheng,[2] and Shanshan Zhao[1]

[1]*School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei, China*
[2]*School of Computer Science, Wuhan University, Wuhan, Hubei, China*

Correspondence should be addressed to Jianhao Tai; taijianhao@whu.edu.cn

Received 7 March 2017; Revised 1 June 2017; Accepted 11 June 2017; Published 27 July 2017

Academic Editor: María Guijarro

Aircraft detection from high-resolution remote sensing images is important for civil and military applications. Recently, detection methods based on deep learning have rapidly advanced. However, they require numerous samples to train the detection model and cannot be directly used to efficiently handle large-area remote sensing images. A weakly supervised learning method (WSLM) can detect a target with few samples. However, it cannot extract an adequate number of features, and the detection accuracy requires improvement. We propose a cascade convolutional neural network (CCNN) framework based on transfer-learning and geometric feature constraints (GFC) for aircraft detection. It achieves high accuracy and efficient detection with relatively few samples. A high-accuracy detection model is first obtained using transfer-learning to fine-tune pretrained models with few samples. Then, a GFC region proposal filtering method improves detection efficiency. The CCNN framework completes the aircraft detection for large-area remote sensing images. The framework first-level network is an image classifier, which filters the entire image, excluding most areas with no aircraft. The second-level network is an object detector, which rapidly detects aircraft from the first-level network output. Compared with WSLM, detection accuracy increased by 3.66%, false detection decreased by 64%, and missed detection decreased by 23.1%.

1. Introduction

Aircraft detection from remote sensing images is a type of small target recognition under a wide range. It has two problems, however, one is the efficiency of large-area image detection; the other is the aircraft feature extraction and expression in complex environments. Nevertheless, the increase of high-resolution remote sensing images has advanced research in this area [1]. The key point of target detection is to find the stable target feature. Traditional aircraft detection methods mainly focus on the feature description, feature selection, feature extraction, and other algorithms [2–5]. Adjustment and optimization of the algorithm can improve the detection accuracy and efficiency [6]. However, these features are common image attributes; thus, it is difficult to fundamentally distinguish between the target and background. Moreover, in a complex environment, the method accuracy is poor, such as that of the histogram of oriented gradient (HOG) [7] and scale-invariant feature transform (SIFT) [3]. Some studies matched the test image by designing a standard sample of the target [2, 8]. However, this method only applies to special scenes; it is not very versatile. In practice, the multifeature fusion method is often used to comprehensively describe the target [9]. Nonetheless, it increases the algorithm complexity and reduces the detection efficiency to some extent [7, 10].

The deep learning method has a strong feature-extraction ability. Through a multilayer neural network and a large number of samples, it extracts the multilevel features of objects. The method has advanced considerably in the field of natural image processing [11, 12]. Furthermore, numerous high-performance algorithms [13, 14], such as the Region Convolutional Neural Network (R-CNN) and Fast and Faster R-CNN [15, 16], have been proposed. However, these methods require many samples to train the network model. When

processing a large-area image, many invalid region proposals must be detected, and this is inefficient. Some target detection methods using remote sensing images based on deep learning have been presented. Weakly supervised and semisupervised feature learning methods have been proposed [17–19]. Zhang et al. proposed an aircraft detection method based on weakly supervised learning coupled with a CNN [20]. Han et al. proposed an iterative detector approach [21], by which the detector is iteratively trained using refined annotations until the model converges. Nevertheless, a nonconvergence situation may occur, which reduces the detection accuracy. Weakly supervised learning extracts features with a small amount of sample data. However, owing to a lack of samples, the features are not sufficient. Therefore, the detection accuracy is limited.

There are three key problems in aircraft detection from remote sensing images by the deep learning method. First, the number of training samples is limited. A means of using a small number of samples to train a high-accuracy model is an important point. Second, aircraft in remote sensing images have obvious features, and some stability features can be selected as constraint conditions. Combining these features with deep learning, the detection method can have a better targeting ability. Third, a remote sensing image covers a large area, its scale is not uniform, and the sensor is miscellaneous. The existing deep learning model and network structure are not directly suitable for remote sensing image aircraft detection [22]. The network structure and detection algorithm must be modified to improve the efficiency and accuracy.

In view of the above problems, a cascade CNN (CCNN) architecture is proposed to improve the accuracy and efficiency. It contains four main parts: (1) A two-level CCNN is designed to rapidly process remote sensing image aircraft detection. The first-level network quickly classifies the scene and eliminates the areas that do not contain aircraft. The second-level network identifies and locates aircraft in the areas not filtered out in the previous step. (2) A transfer-learning method is used to fine-tune the parameters of pretrained classification and object detection models with the samples. (3) A region proposal filtering method of a geometric feature constraint (GFC) based on the geometric features of aircraft is proposed. By using these features to filter the region proposals, numerous nonaircraft region proposals are eliminated. (4) An aircraft sample library and a remote sensing image-scene sample library are established. They are used as the data source of transfer-learning.

2. Related Work

In this section, we review related work that has utilized deep learning for scene classification and target detection. Remote sensing scene classification plays a key role in a wide range of applications and has received remarkable attention from researchers. Significant efforts have been made to develop varied datasets and present a variety of approaches to scene classification from remote sensing images. However, there has yet to be a systematic review of the literature concerning datasets and scene classification methods. In addition, almost all existing datasets have several limitations, including the small scale of scene classes and image numbers, the lack of image variation and diversity, and accuracy saturation. In response to this problem, Cheng et al. [23] proposed a large-scale dataset, called "NWPU-RESISC45," which is a publicly available benchmark for remote sensing image scene classification (RESISC). This dataset contains 31,500 images, covering 45 scene classes with 700 images in each class, which provides strong support for RESISC. Paisitkriangkrai et al. [24] designed a multiresolution convolutional neural network (CNN) model for the semantic labeling of aerial imagery. To avoid overfitting from training a CNN with limited data [25] investigated the use of a CNN model pretrained on general computer vision datasets for classifying remote sensing scenes. However, the CNN model in [25] was not further fine-tuned and was directly used for land-use classification. Yao et al. [26] proposed a unified annotation framework combining discriminative high-level feature learning and weakly supervised feature transfer. Specifically, they first employ an efficient stacked discriminative sparse autoencoder (SDSAE) to learn high-level features on an auxiliary satellite image data set for a land-use classification task.

Another way to improve the efficiency of aircraft detection is to first detect the airport range and then detect the aircraft within the airport. Airport detection from remote sensing images has gained increasing research attention [21, 27, 28] in recent years due to its strategic importance for both civil and military applications; however, it is a challenging problem because of variations in airport appearance and the presence of complex cluttered backgrounds and variations in satellite image resolution. Yao et al. [29] proposed a hierarchical detection model with a coarse and a fine layer. At the coarse layer, a target-oriented saliency model is built by combining contrast and line density cues to rapidly localize airport candidate areas. At the fine layer, a learned condition random field (CRF) model is applied to each candidate area to perform fine detection of the airport target. CNNs have demonstrated their strong feature representation power for computer vision. Despite the progress made in natural scene images, it is problematic to directly use CNN features for object detection in optical remote sensing images because it is difficult to effectively manage the problem of variations in object rotation variation. To address this problem, Cheng et al. [30] proposed a novel and effective approach to learning a rotation-invariant CNN (RICNN) model that advances object detection performance, which is achieved by introducing and learning a new rotation-invariant layer based on existing CNN architectures. Cosaliency detection, the goal of which is to discover the common and salient objects in a given image group, has received tremendous research interest in recent years. However, most existing cosaliency detection methods assume that all images in an image group should contain cosalient objects in only one category and are thus impractical for large-scale image sets obtained from the Internet. Yao et al. [31] improved cosaliency detection and advanced its development. Their results can outperform state-of-the-art cosaliency detection methods performed on manually separated image subgroups.

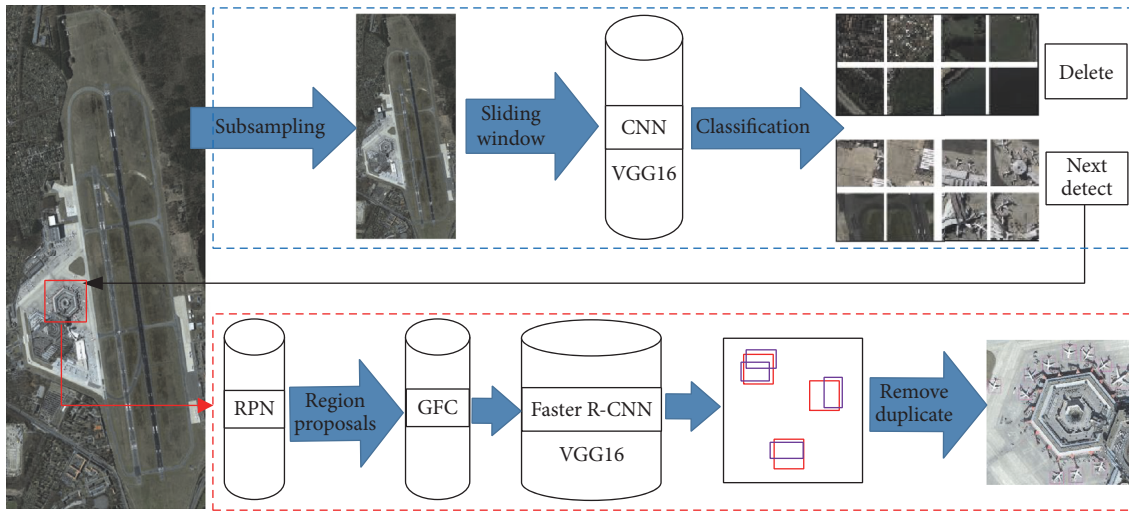


Figure 1: *Architecture of two-level CCNN for large-area remote sensing rapid aircraft detection*. The target detection process consists of two parts. First, the image is downsampled. The first-level CNN structure is used to classify the scene, with a sliding window on the downsampled image, block-by-block. The nontarget window image is excluded and the index position of the target window is reserved and passed to the next level of the network. The second level receives the target window and performs aircraft target detection on the original image. An RPN is used to get region proposals, and the GFC model is used to filter multiple region proposals and exclude the region proposals that do not satisfy the geometric characteristics of aircraft. Then, the Faster R-CNN classification model is used to classify the remaining region proposals to generate the target area. Finally, using overlap constraints, delete the redundant target area to get the final detection results.

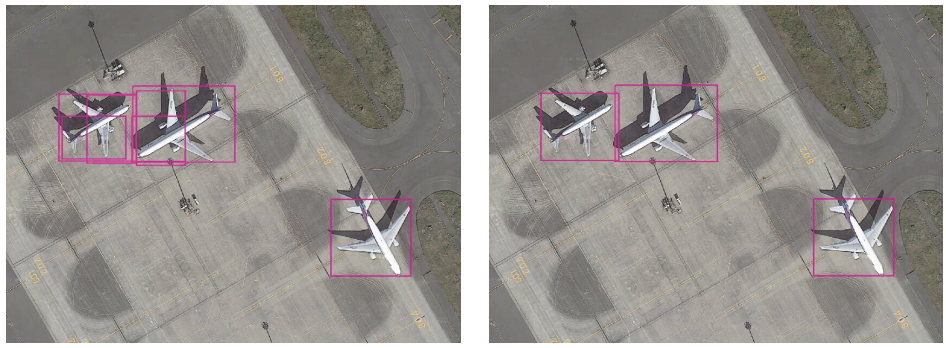
3. Methods

3.1. Cascade Network for Aircraft Detection. In general, the image used for aircraft detection includes the entire airport and surrounding areas; most image areas do not contain aircraft. This situation reduces the detection accuracy and efficiency. We propose a two-level CCNN method that combines the trained models and GFC method to solve this problem. As shown in Figure 1, the first level is a coarse filter. A new image with lower resolution is obtained through twofold downsampling of the original image. The new image is traversed using a sliding window, and the CNN image classifier is used to classify each sliding window. Most of the negative areas are filtered out, and the positive area is fed into the second-level network for aircraft detection.

The second-level network is a target detector. In this level, Faster R-CNN is used to perform aircraft detection for the respective area. The output of the first-level network is the input of the second-level network. The corresponding region on the original image is sent to the second network, and the window size is 800 × 800 pixels. The region proposals are generated by the region proposal network (RPN). The main difference between Faster R-CNN and Fast R-CNN is the generation of the region proposal; that is, the former uses the efficient RPN method, whereas the latter uses the less efficient selective search (SS) method. In the RPN, each image produces approximately 20,000 region proposals of three ratios and three scales. Then, according to the region proposal classification score, the top 300 region proposals are selected as the Faster R-CNN input for target detection. More than 95% of the 20,000 region proposals are background. Scoring and sorting such a large number of region proposals is obviously time-intensive. The GFC is used to delete the invalid region proposals. The final detection for the remaining region proposals is performed by the aircraft detector.

The general sliding window method uses a small size window, such as 12 × 12, 24 × 24, and 48 × 48. Small windows and small steps require considerable time to traverse the entire image, which reduces the detection efficiency. In an experiment, we used a large-size sliding window. In the first level, the window size was 400 × 400, which was suitable for image classification. According to the training samples, the downsampled images had a higher classification accuracy; moreover, the speed of traversing the entire image was improved. In the second level, the window size was 800 × 800. The range of this window was equal to the downsampling range. It required almost the same amount of time to detect a 400 × 400 image and an 800 × 800 image. Thus, it was not necessary to decompose the original image into four 400 × 400 smaller images. The window size and sliding step size were determined by the image resolution. It had to satisfy the condition that at least one window could contain a complete maximum aircraft.

Because of the overlap between adjacent windows, some aircraft may be detected several times. The repetitive detection areas should be optimized. The redundant objects are deleted, and only one optimal result is retained. The regions for the same aircraft must meet the following conditions. (1) Every region should satisfy the condition of GFC; that is, the aspect ratio, width, and height must meet the thresholds of (r_1, r_2), (w_1, w_2), and (h_1, h_2), respectively. (2) The overlap between two regions should be larger than 30%, as defined in (1). (3) The area is the largest, as defined in (2), in which area(r_i) means the area of the given region. The process is



(a) (b)

Figure 2: *Process of deleting duplicate detected regions and overlapping regions between sliding windows*. (a) Result before deleting the duplicate. (b) Result after deleting the duplicate.

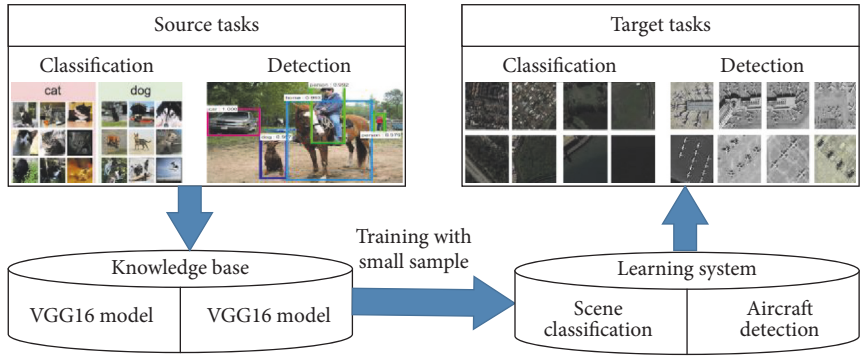


Figure 3: *Structure of transfer-learning for scene classification and target detection*. Source tasks are the original deep learning methods of classification and detection, and the knowledge base is comprised of the trained networks. The learning system involves transfer-learning and fine-tuning, and the target tasks are remote sensing image scene classification and aircraft detection.

shown in Figure 2, in which the two aircraft at the upper left are in overlapping windows. The incomplete regions are deleted, and the most appropriate one is preserved.

$$\text{area}\left(r_i \cap r_j\right) \geq 30\% \tag{1}$$

$$\text{Region} = \max \text{area}\left(r_i\right). \tag{2}$$

3.2. Transfer-Learning and Fine-Tuning the Model Parameters. Transfer-learning refers to a model training method. The parameters of the pretrained network model are fine-tuned by a small number of samples to prompt the original model to adapt to a new task [32, 33]. The deep learning object detection method can achieve very high detection accuracy. One reason is that a large number of samples are used to train the detection model, which contains a sufficient number of target features. However, for the remote sensing image, no large sample library, such as ImageNet, exists that can be used to train a high-accuracy model. Moreover, training a new high-accuracy object detection model consumes considerable time and computing resources. Transfer-learning can solve the problem of lacking samples and significantly shorten the training time [34].

In this paper, the transfer-learning method is, respectively, used to fine-tune the pretrained image classification model and object detection model. As shown in Figure 3, the source models are the respective classification and object detection models. Both include the VGG16 network structure [35] and are trained with ImageNet. They were trained with the samples given in Section 4.1. The parameters of the original model were fine-tuned by the learning system to complete the tasks of scene classification and aircraft detection.

The CNN structure comprises many layers. The bottom layers are used to extract generic features, such as structure, texture, and color. The top layers contain specific features associated with the target detection task. Using transfer-learning, the top layer parameters are adjusted with a small number of aircraft samples to change the target detection task. Through a backpropagation algorithm, the bottom-layer parameters are adjusted to extract aircraft features. In general, we use softmax as the supervisory loss function, which is a classic cross-entropy classifier. It is defined in (3), where $f(x_i)$ is the activation of the previous layer node that pushed through a softmax function. In (4), the target output y_i is a $1 - K$ vector, and K is the number of outputs. In addition, L is the number of layers, and λ is a regularization term. The goal is to minimize loss. Finally, stochastic gradient descent

is used to fine-tune the model and backpropagation is used to optimize the function.

$$f(x_i) = \text{soft}\max(x_i) = \frac{e^{x_i}}{\sum_{i=1}^{N} e^{x_i}} \tag{3}$$

$$\text{loss} = -\sum_{i=1}^{m} y_i \log f(x_i) + \lambda \sum_{l=1}^{L} \text{sum}\left(\|w_l\|^2\right). \tag{4}$$

Another important parameter is the learning rate. The parameters of the pretrained model are fine; thus, the network layer learning rate should be set to a small number. Therefore, a small learning rate is used to ensure that both the features of the new sample can be learned and the excellent parameters of the pretrained model can be maintained. Taking Faster R-CNN, for example, the parameters that must be set are the following. (1) By initializing the training network and selecting the pretrained model, the "train_net" is set to VGG16.v2.caffemodel. (2) Set the number of categories. In this paper, all aircraft types are considered the same type; thus, there are only two categories, aircraft and background. In the data layer, num_classes is set to two; in the cls_score layer, num_output is set to two; and in the bbox_pred layer, num_output is set to eight. (3) In setting the learning rate, the base learning rate, base_ly, is set to 0.001. The learning rate ly_mult of the cls_score layer and the bbox_pred layer is set to five and ten, respectively. The learning rates of other convolutional layers are not modified. (4) In setting the number of iterations of the backpropagation function, max_iters is set to 40,000.

3.3. Geometric Feature Constraint. In a remote sensing image, the aircraft covers only a small area; the remaining area is the background. The region proposals are mostly background, not aircraft. Reducing the number of region proposals can make object detection more targeted, which helps improve efficiency [16]. The special imaging method of the remote sensing image determines that only the top features of the aircraft can be viewed in the image [36]. Thus, the geometric features of aircraft in the image have high stability [37]. This feature can be exploited when designing a detection method. When the resolution of the image is fixed, the geometric parameters of each aircraft are constant. Therefore, the geometric features can be used as important conditions for aircraft detection.

The geometric features mentioned in this paper mainly refer to external contour sizes of aircraft in remote sensing images. It is possible to use geometric signature constraints for detection because the following three conditions are met. First, conventional remote sensing imaging employs nadir images, that is, the sensor gets an image from the sky that is perpendicular (or near-perpendicular) to the ground. Thus, the top contour features of aircraft are seen in these remote sensing images; these features (length, width, and aspect ratio) are relatively stable constants. Although the length and width are affected by image resolution, they can be calculated from the resolution of the image, whereas the aspect ratio is not affected by resolution and is more fixed. Second, the length, width, wingspan, and body length ratio of different types of aircraft are all within a certain range. These geometric features give us an exact basis for calculating region proposals. Third, statistical methods are used to count the geometric features of many aircraft samples. Samples have been collected in the world's major countries from large, medium, and small airports, covering almost all types of aircraft, including large passenger aircraft, fighters, small private aircraft, and rotor helicopters. These statistical results are used as a reference for geometric feature constraints.

The image resolution and the result of aircraft geometry statistics are used to calculate the size of the sliding window. The conditions that need to be met are as follows. The minimum overlap between the windows must be greater than the maximum size of the aircraft to ensure that no aircraft will be missed. The number of aircraft in a sliding window is very small; a window contains about 2000–3000 region proposals, but the aircraft are probably in only a few. The number of target region proposals is much smaller than the number of nontarget region proposals, and having many nontarget areas results in increased overall time consumption. With GFCs, deleting a candidate box that does not satisfy the condition can greatly improve the time efficiency of target detection.

Taking the ground truth of the aircraft in the sample as an example, their lengths and widths are within a fixed range. The original 2,511 sample images were used to analyze these parameters, and the results are shown in Figure 4.

Figure 4(a) shows that the aspect ratio of the circumscribed rectangle is primarily around 1.0, and the range is (0.6, 1.6). Figure 4(b) shows that the width and height of the circumscribed rectangle are mainly around 140 pixels, and the range is (50, 300). These values should be appropriately enlarged during aircraft detection to ensure that all aircraft are not filtered out on account of statistical errors. The geometric features of the aircraft in the remote sensing image have rotation invariance. The aspect ratio is within a fixed range, even at different resolutions; thus, it can be used as a constraint to roughly distinguish between target and background. (r_1, r_2), (w_1, w_2), and (h_1, h_2) represent the ranges of aspect ratio, width, and height, respectively.

For the same image, reducing the number of nontarget region proposals has no effect on the target detection accuracy (in deep learning, this accuracy means the classification score). However, it can improve the detection efficiency. Based on the GFC, a region proposal filter method is proposed. (r_1, r_2), (w_1, w_2), and (h_1, h_2) are set as thresholds to filter the region proposals. The main algorithm is shown in (5), where Region Proposal = 1 means the region proposal is preserved, and Region Proposal = 0 means it is deleted. Ratio represents the target aspect ratio, width represents the target width, height denotes the target height, and α is the image resolution coefficient, which is used to adjust the threshold at different resolutions.

$$\text{Region Proposal} = \begin{cases} 1, & r_1 \le \text{ratio} \le r_2 \\ 0, & \text{ratio} < r_1 \text{ or ratio} > r_2 \end{cases}$$

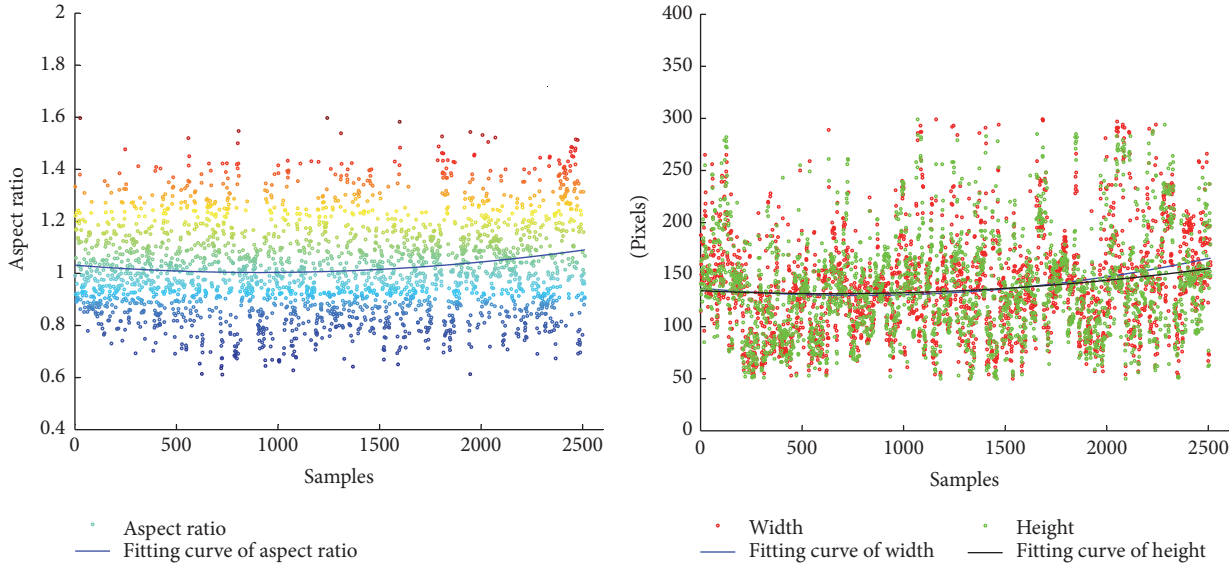


FIGURE 4: *Geometric features all of the sample aircraft.* (a) Aspect ratio of each sample aircraft. They are approximately 1.0, and the range is (0.6, 1.6). The color in Figure 4(a) represents the value of the aspect ratio; the blue is the smallest value; the red is the largest value; and the middle is gradually transitions. (b) Size of each sample aircraft; the red is the width, and the green is the height. They are approximately 140 pixels, and the range is (50, 300).

$$\text{Region Proposal} = \begin{cases} 1, & \alpha \times w_1 \le \text{width} \le \alpha \times w_2 \\ 0, & \text{width} < \alpha \times w_1 \text{ or width} > \alpha \times w_2 \end{cases}$$

$$\text{Region Proposal} = \begin{cases} 1, & \alpha \times h_1 \le \text{height} \le \alpha \times h_2 \\ 0, & \text{height} < \alpha \times h_1 \text{ or height} > \alpha \times h_2. \end{cases} \quad (5)$$

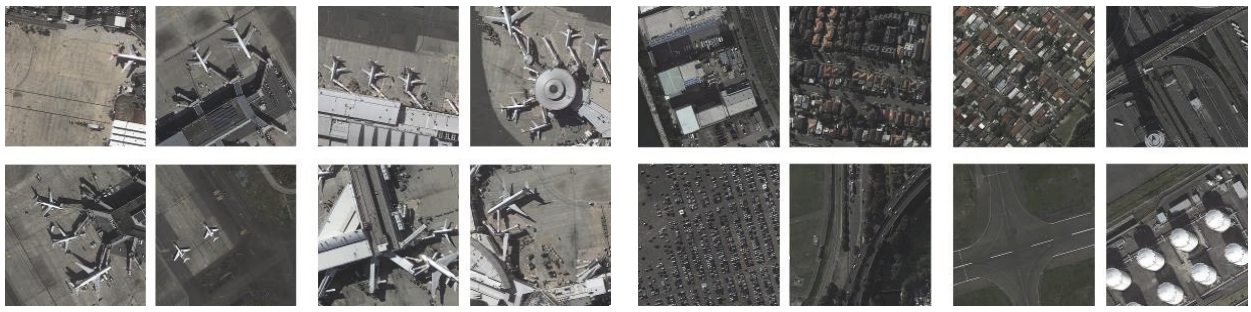
4. Materials

In this study, we constructed three kinds of datasets: scene, aircraft sample, and test datasets. All data were obtained from Google Earth. The scene dataset was from the 18th level, and its spatial resolution was 1.2 m. The remaining two datasets were from the 20th level, and the spatial resolution was 0.3 m. From a professional point of view, compared with the real remote sensing images, the data obtained from Google Earth contained insufficient information. However, the deep learning method is different from the traditional remote sensing image processing method in that it can obtain rich information from this image type. We advocate the use of real remote sensing images; nonetheless, one could attempt use of Google Earth images.

4.1. Training Sample Datasets. Creating training samples is a key step in object detection in the field of deep learning. The sample quality directly affects the test results. In the CCNN structure, the first-level network uses the CNN image classification model, and the second-level network uses the Faster R-CNN object detection model. The two networks have different structures and functions; therefore, we created two sample datasets, one for scene classification and the other for aircraft detection.

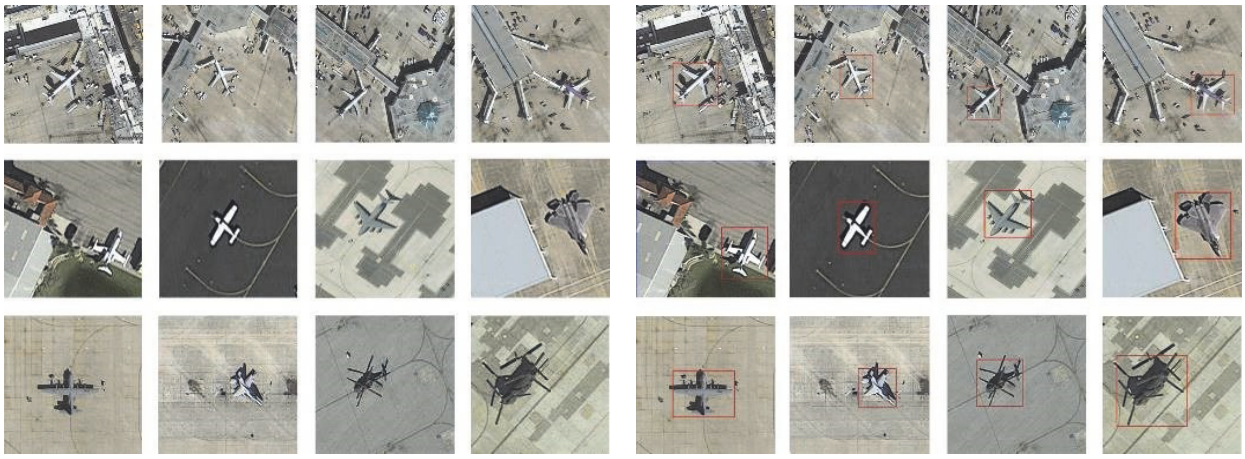
The first dataset was the remote sensing image scene dataset, which was used to train the CNN classification model. We used the "UC Merced Land Use Dataset" as the basic dataset [38]. It contained 21 scene categories and included many typical background categories, such as agriculture, airplanes, beaches, buildings, bushes, and parking lots. This dataset was suitable for scene classification; however, there were only 100 images in each category. To obtain a high-accuracy training model, it was necessary to increase the sample number of airports and other scenes that are easily confused with airports. These scenes include industrial areas, residential areas, harbors, overpasses, parking lots, highways, runways, and storage tanks. Another 300 samples of airport images and 700 samples of other scene categories were added to the dataset. The number of scene categories remained 21. The sample number could be expanded to be eight times the original by image transformation. Finally, the total number of samples was 24,800. Some added images are shown in Figure 5, where (a) depicts an airport scene, and (b) depicts scenes that are easily confused with airports.

The second dataset was an aircraft sample library of remote sensing images. It was used to train the Fast R-CNN network model. It contained 2,511 original aircraft images, some of which are shown in Figure 6. The sample contained not only all aircraft types, but also as much background as possible. The diversity of the samples could help the network



(a) (b)

Figure 5: *Sample images of different scenes*. (a) Airport scene, including a parking apron, runway, and general terminal. (b) Scenes easily confused with an airport, including industrial areas, residential areas, harbors, overpasses, parking lots, highways, runways, and storage tanks.



(a) (b)

Figure 6: *Training sample dataset of aircraft in remote sensing images*. (a) Original training sample images. (b) Corresponding labeled ground truth images.

learn different features and better distinguish the target and background. As shown in Figure 6(a), the sample covers almost all types of aircraft in the existing remote sensing image, including an airliner, small plane, military aircraft, and helicopters.

We obtained the image from the zenith point to the nadir point. Thus, the unique imaging perspective results of the aircraft in the remote sensing images showed a fixed geometric form. We could only view the top of the aircraft. Therefore, the direction and displacement of the aircraft were two important factors that affected the expression of the target. To ensure that the target had a higher recognition rate in different directions and displacements, we created transformations on the sample, such as a horizontal flip, vertical flip, and rotations of 90°, 180°, and 270°. These transformations not only could increase the number of samples but also adjust the direction and structure of the target to increase the sample diversity and make the model more robust. After transformation, the sample contained a total of 20,088 images. These images were divided into three groups with different numbers. Plane_S contained 2,511 images, Plane_M contained 10,044 images, and Plane_L contained 20,088 images.

Labeling the image is an important part of creating a training sample dataset. The label is the basis for distinguishing between the aircraft and background. The conventional sample producing method is to independently create a positive sample and negative sample. In this paper, we produce a sample in a different approach; the difference is that the positive and negative samples are included in the same image. The advantage to this approach is that the target is better distinguished from the background, which can improve the detection accuracy. For example, the trestle that connects the airplane and terminal is prone to false detection. If the airplane and trestle are contained in the same image, the target detection network can better distinguish their differences. As shown in Figure 6(b), the aircraft position is manually

Table 1: Detailed information of the three airports.

Test image	Image size (pixel)	Aircraft number	Covered area (km^2)
Berlin Tegel Airport	22016 × 9728	35	19.1
Sydney International Airport	13568 × 22272	78	26.9
Tokyo Haneda Airport	20480 × 17920	94	32.7

Table 2: Results of transfer-learning on different train networks and different sample sizes.

Training network	Datasets	Time (min)	Per Iteration (s)	Loss_bbox	Loss_cls
CaffeNet	Plane_S	64	0.061	0.055	0.087
	Plane_M	241	0.061	0.051	0.076
	Plane_L	837	0.062	0.047	0.025
VGG16	Plane_S	320	0.446	0.032	0.084
	Plane_M	497	0.446	0.029	0.044
	Plane_L	976	0.453	0.019	0.028

marked in the image. An XML annotation file is generated in which basic information about the image and the aircraft is automatically saved.

4.2. Test Dataset. Three large-area images were employed for the test. Each image contained a complete airport, and the area was 19.1 km^2, 26.9 km^2, and 32.7 km^2, respectively. The test images had the characteristics of a large area, a diversity of ground cover types, and complex backgrounds. In addition to the airport, there were residential areas, harbors, industrial areas, woodlands, and grasslands in the image. The types of objects in the image were very rich and the environment was complex. Details of the three images are shown in Figure 7 and outlined in Table 1.

Aircraft and airport detection are two different research topics [39]. The focus of this paper is aircraft detection of high-resolution remote sensing images. Our objective is not to detect the airport; rather, it is to directly detect the aircraft. The test images were three airports. Nevertheless, the experimental data did not cover only the airport area. In fact, when the images were enlarged, it was evident that they contained harbors, industrial zones, and residential areas.

5. Results and Discussion

In this paper, the full implementation was based on the open-source deep learning library Caffe [40]. The experiments were run using an Nvidia Titan X 12GD5 graphics-processing unit (GPU). The network models used in the experiments were VGG16 pretrained with ImageNet.

5.1. Training Detection Model. The two VGG16 pretrained models were fine-tuned using the two sample datasets presented in Section 4.1. The evaluation training results are shown in Figure 8. As shown in Figure 8(a), the image classification accuracy quickly improves during the first 10,000 iterations and reaches a steady value of 0.9 at 40,000 iterations. Figure 8(b) depicts the total loss of the object detection model, whereby it reaches 0.1 at 40,000 iterations. Figures 8(c) and 8(d) are the location and classification losses, respectively, with both reaching a value of less than 0.1. Excellent transfer-learning results provided a good experimental basis and were used for the next step.

The transfer-learning results are directly related to the accuracy of the next aircraft detection process. Using transfer-learning, a high-accuracy aircraft detection model could be obtained using a small number of training samples and by requiring a brief training time. Here, we use CaffeNet and VGG16 as examples. The number of iterations was 40,000. Four indicators were used to analyze the results: total training time (Time), a single iteration time (Per Iteration), position accuracy (Loss_bbox), and classification accuracy (Loss_cls). The transfer-learning results are shown in Table 2. The results show that each combination of transfer-learning has good training results. The model with more training samples has higher accuracy than the model with fewer samples, and the complex model VGG16 has higher accuracy than the simple model, CaffeNet.

Table 2 shows that the accuracy of target classification (Loss_cls) increases with the number of samples. Plane2_L increases by approximately 60% compared to Plane2_S. The position accuracy (Loss_bbox) has the same trend as the former; however, it is not obvious. This result proves that the method of producing samples proposed in Section 4.1 is effective. The image transformation method can increase the sample number and enhance the robustness of the samples. It can additionally improve the ability of transfer-learning and enhance the robustness and accuracy of the detection model. However, the target detection accuracy remains affected by the number of samples. When the number of samples is large, the target detection accuracy is relatively high, and the training time is correspondingly increased.

CaffeNet is a small network, whereas VGG16 is a large network, and its structure is more complex than CaffeNet. As the complexity of the network increases, the position accuracy also increases. Large networks increase by approximately 40% over small networks. In the same network, the training time increases with the number of samples and the network complexity. The results show that transfer-learning is an effective high-accuracy model training method, and it can

Table 3: Large-area image aircraft detection results.

Test image	True aircraft	Detected aircraft	Missing detection	False detection
Berlin Tegel Airport	35	35	0	2
Sydney International Airport	78	76	2	5
Tokyo Haneda Airport	94	91	3	4

Table 4: Detailed evaluation criteria and comparison for the four detection results.

Test image	Method	FDR (%)	MR (%)	AC (%)	ER (%)
Berlin Tegel Airport	CCNN	5.71	0	100	5.71
	LOGNet-C	7.69	3.23	96.77	10.92
Sydney International Airport	CCNN	6.58	2.56	97.44	9.14
	LOGNet-C	48.54	10.87	89.13	59.41
Tokyo Haneda Airport	CCNN	4.39	3.19	96.81	7.58
	LOGNet-C	20.80	1.54	98.46	22.34

realize the task of training a high-accuracy target detection model with fewer samples.

5.2. Aircraft Detection. The three large-area test images were used to verify the aircraft detection accuracy. The results are shown in Figure 9. Based on the global and local details in the figure, the overall detection result is good. Most of the aircraft in the airport area were detected. The detection-missed aircraft were small aircraft; there was no missed detection of large aircraft. Moreover, there was no falsely detected aircraft in industrial areas, residential areas, woodlands, grasslands, or other areas beyond the airports. However, there were individual falsely detected targets within the airports, primarily airport trestles, and airport terminals. The quantitative indicators are shown in Table 3, in which "True" indicates the number of real aircraft in the original image, "Detected" indicates the number of aircraft detected, "Missing" is the number of detection-missed aircraft, and "False" is the number of falsely detected aircraft.

The two-level CCNN aircraft detection structure could rapidly process large-area remote sensing images with high accuracy; however, the commission errors and errors remained. In a complex environment, many objects have features similar to those of aircraft, especially long-shaped buildings and facilities. As listed in Table 3, the false detection rate of each airport is higher than the missing detection rate. Through analysis, we determined that the detection-missed aircraft were not actually lost. In the detection procedure, the classification threshold was set to 0.6. When the classification score of a target was less than 0.6, the target was not marked. Then, the missing detection appeared. When the threshold was lowered to 0.5, the detection-missed targets were marked, and the detection miss rate was zero. Nevertheless, there were more falsely detected targets. Airport terminals and several buildings had features similar to those of aircraft. These were considered aircraft and marked. When the threshold was raised, the false detection rate decreased, while the detection miss rate increased. When the threshold was lowered, the false detection rate increased, while the detection miss rate decreased. According to the needs of the detection task, it is a better choice to set a suitable threshold and to balance the false detection rate and detection miss rate.

The experimental results in Table 3 show that the proposed method of CCNN is feasible. A high-accuracy aircraft detection model can be obtained by using the transfer-learning method to fine-tune a pretrained model. The GFC method greatly reduces the number of region proposals to directly improve the detection efficiency. The two-level cascade network architecture can detect aircraft in large-area remote sensing images with high accuracy and high efficiency. However, some shortcomings remain. Detailed analysis of transfer-learning and detection accuracy is outlined below.

The aircraft detection method performance was evaluated by four criteria: falsely detected ratio (FDR), miss ratio (MR), accuracy (AC), and error ratio (ER). Their meanings are given in (6). A remote sensing image aircraft detection method named LOGNet-C based on weakly supervised learning is proposed in article [20]. The results of the CCNN method were compared with those of the LOGNet-C method. The four evaluation indicators and comparison results are shown in Table 4.

$$\begin{aligned} \text{FDR} &= \frac{\text{Number of falsely detected aircraft}}{\text{Number of detected aircraft}} \times 100\% \\ \text{MR} &= \frac{\text{Number of missing aircraft}}{\text{Number of aircraft}} \times 100\% \\ \text{AC} &= \frac{\text{Number of detected aircraft}}{\text{Number of aircraft}} \times 100\% \\ \text{ER} &= \text{FDR} + \text{MR}. \end{aligned} \tag{6}$$

The comparison results show that the false detection rate of the CCNN method is reduced by an average of 64% more than the method of LOGNet-C. The missing detection rate decreases by an average of 23.1%. Overall, the detection accuracy increases by an average of 3.66%.

The test results showed that small aircraft are more prone to missed detection, while long-shaped buildings and facilities are more prone to false detection, as shown in Figure 10. This phenomenon is related to the choice of training

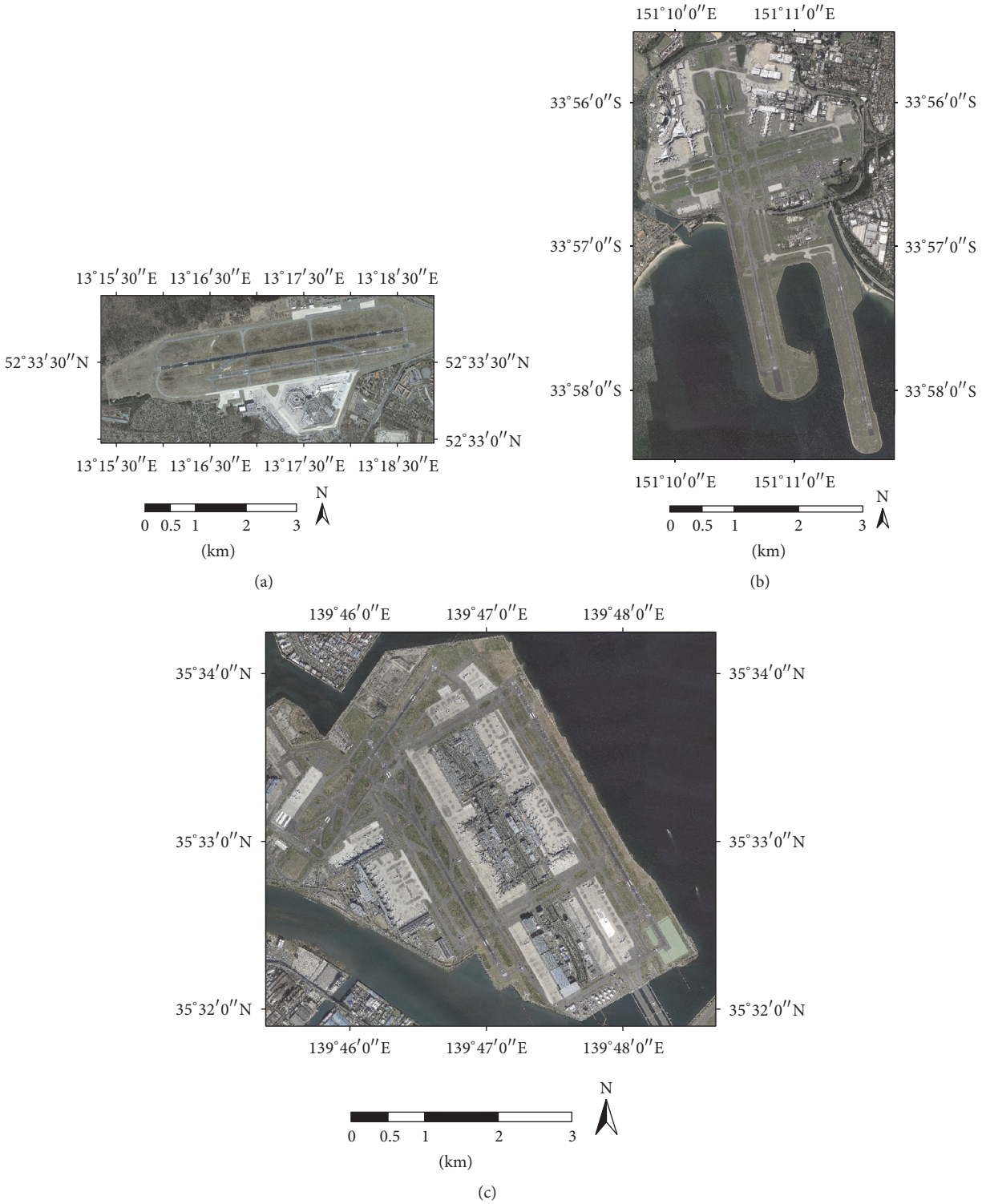


Figure 7: *Test remote sensing images for aircraft detection.* (a) Berlin Tegel Airport. (b) Sydney International Airport. (c) Tokyo Haneda Airport.

samples. The small aircraft in the sample are only a minority. In the transfer-learning stage, the detection model did not learn more features of small aircraft. Long-shaped buildings were detected as aircraft. The first reason is that these buildings had similar contour features as aircraft. The second reason is that there were an inadequate number of similar buildings in the negative sample. Consequently, the features contained in the negative sample were insufficient. When

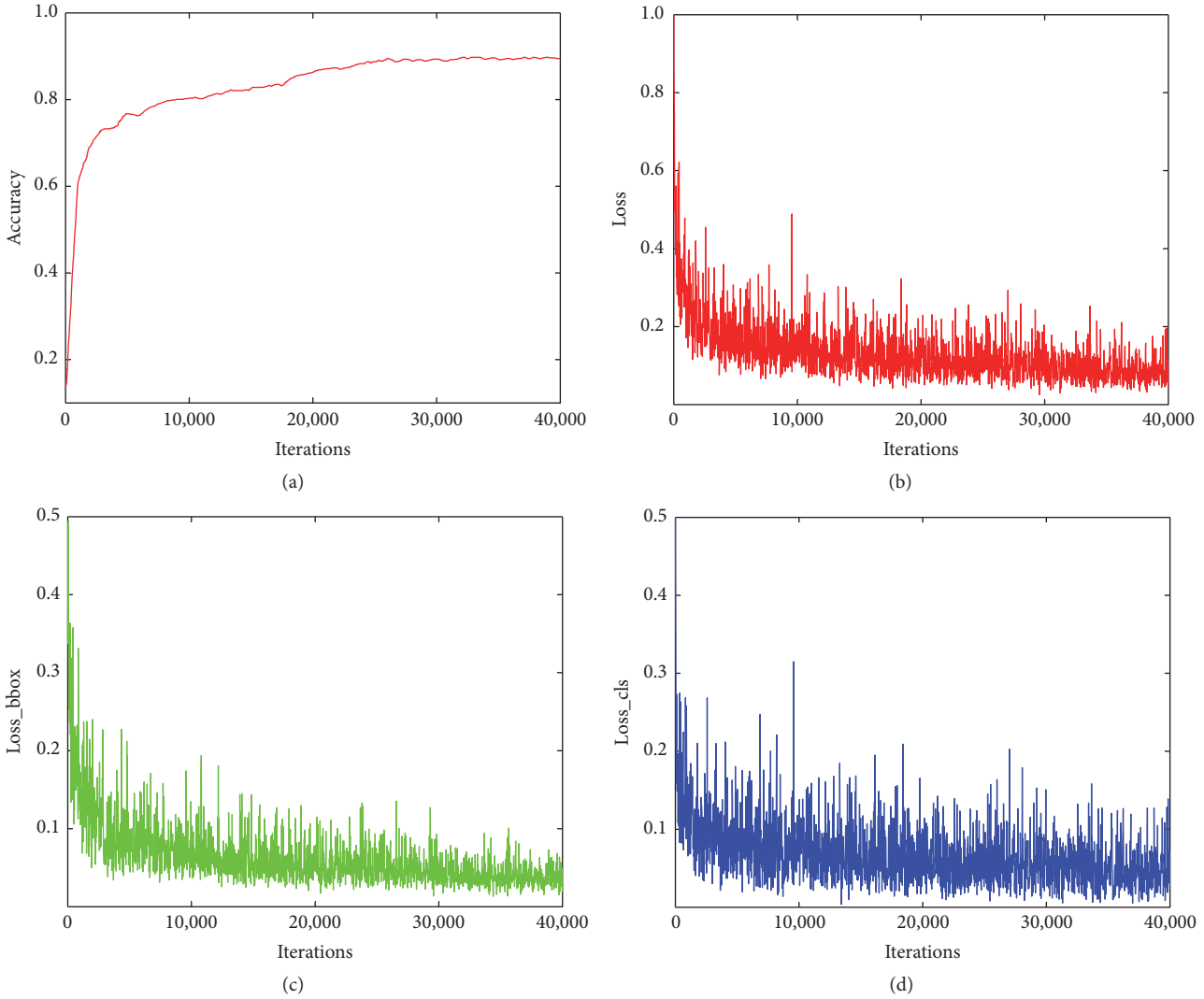


Figure 8: *Accuracy and loss changes of the two models during the fine-tuning process of image classification and object detection.* (a) Accuracy of the image classification model. (b) Total loss of the object detection model, which was the weighted sum of loss_cls and loss_bbox. (c) Location loss of the object detection model. (d) Classification loss of the object detection model.

increasing the number of training samples of these types and retraining the model, the target detection accuracy was increased. The analysis results show that when producing the training sample dataset, it should not only contain all types of targets, but the number of each target type should be balanced. The negative samples should contain as much background information as possible, especially for objects that have a texture, color, geometry, or other features similar to those of aircraft.

All false detection occurred in the airport. Building-intensive areas were prone to false detection; however, they did not appear. This was on account of the contribution of the first-level cascade network, which filtered out nontarget areas and did not feed these areas into the aircraft detection network. Missed detection primarily occurred in the airport interior, and no aircraft were filtered out by the first-level network. The main cause of the missed detection was that the classification score was not high, and the network did not extract an adequate number of features.

6. Conclusions

In the framework of deep learning, a rapid large-area remote sensing image aircraft detection method of CCNN was proposed. Using the transfer-learning method, the parameters of the existing target detection model were fine-tuned with a small number of samples, and a high-accuracy aircraft detection model was obtained. Based on the geometric feature invariance of the aircraft in the remote sensing image, a region proposal processing algorithm, GFC, was proposed to improve the efficiency. Based on the above methods, a cascade network architecture was designed for large-area remote sensing images for aircraft detection. This method not only can directly handle large-area remote sensing images for

(a)

(b)

(c)

Figure 9: *Aircraft detection results of the three test images, including the global images and local enlarged images*. Detection results of (a) Berlin Tegel Airport; (b) Sydney International Airport; (c) Tokyo Haneda Airport.

(a) (b) (c) (d)

Figure 10: *Commission and omission errors*. The purple rectangle is the correctly detected aircraft, the green rectangle is the falsely detected aircraft, and aircraft without a rectangle are missed detected aircraft. (a) and (b) are falsely detected aircraft, that is, predominantly long-shaped buildings and facilities. (c) and (d) are detection-missed aircraft, predominantly small aircraft.

aircraft detection, but it also overcomes the difficulty of training a high-accuracy model with a small number of samples.

The experimental results showed that the detection accuracy increased by an average of 3.66%. The false detection rate decreased by an average of 64%, and the missed detection rate decreased by an average of 23.1%. However, some of the algorithms warrant improvement. The target detection threshold is experientially set and has some limitations. In future work, we intend to address the needs of the detection task to make the system automatically set the threshold value. In the present work, we focused on aircraft detection. If the target is extended to other typical targets, such as ships, tanks, and vehicles, then this method will have a larger application scope.

The failures in target detection in this experiment may have occurred for the following reasons. First, the number of small aircraft in our sample is insufficient, and the sample number of various types of aircraft is not balanced, which has resulted in inadequate learning of the features of small aircraft during training. Second, the choice of target threshold must be appropriate. When detection is nearly complete, each region proposal is scored. When this score is greater than the threshold, the target is marked and displayed; however, when the score is less than the threshold, the target is not marked and is not displayed but is instead hidden as background. There is a close relationship between the threshold value and detection accuracy, especially false positives and negatives. When the threshold value is too small, false positives decrease. However, when the threshold value is too large, false negatives increase. Currently, we set our threshold value empirically and plan to adopt an adaptive threshold method in the future.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank Dr. Changjian Qiao for his excellent technical support. Without his instructive guidance, impressive kindness, and patience, they could not have completed this thesis. They would also like to thank Dr. Huijin Yang and Yuan Zhao for critically reviewing the manuscript.

References

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

[2] Y. Li, X. Sun, H. Wang, H. Sun, and X. Li, "Automatic target detection in high-resolution remote sensing images using a contour-based spatial model," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 5, pp. 886–890, 2012.

[3] X. Bai, H. Zhang, and J. Zhou, "VHR object detection based on structural feature extraction and query expansion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 10, pp. 6508–6520, 2014.

[4] L. Gao, B. Yang, Q. Du, and B. Zhang, "Adjusted spectral matched filter for target detection in hyperspectral imagery," *Remote Sensing*, vol. 7, no. 6, pp. 6611–6634, 2015.

[5] E. Fakiris, G. Papatheodorou, M. Geraga, and G. Ferentinos, "An automatic target detection algorithm for swath Sonar backscatter imagery, using image texture and independent component analysis," *Remote Sensing*, vol. 8, no. 5, article no. 373, 2016.

[6] W. Zhang, X. Sun, K. Fu, C. Wang, and H. Wang, "Object detection in high-resolution remote sensing images using rotation invariant parts based model," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 74–78, 2014.

[7] H. Sun, X. Sun, H. Wang, Y. Li, and X. Li, "Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 1, pp. 109–113, 2012.

[8] X. Sun, H. Wang, and K. Fu, "Automatic detection of geospatial objects using taxonomic semantics," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 23–27, 2010.

[9] R. M. Willett, M. F. Duarte, M. A. Davenport, and R. G. Baraniuk, "Sparsity and structure in hyperspectral imaging: sensing, reconstruction, and target detection," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 116–126, 2014.

[10] F. Gao, Q. Xu, and B. Li, "Aircraft detection from VHR images based on circle-frequency filter and multilevel features," *The Scientific World Journal*, vol. 2013, Article ID 917928, 7 pages, 2013.

[11] I. Sevo and A. Avramovic, "Convolutional neural network based automatic object detection on aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 5, pp. 740–744, 2016.

[12] L. Zhang, G.-S. Xia, T. Wu, L. Lin, and X. C. Tai, "Deep learning for remote sensing image understanding," *Journal of Sensors*, vol. 2016, Article ID 7954154, 2 pages, 2016.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[14] E. Pasolli, F. Melgani, N. Alajlan, and N. Conci, "Optical image classification: a ground-truth design framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 6, pp. 3580–3597, 2013.

[15] R. Girshick, "Fast R-CNN," in *Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV '15)*, pp. 1440–1448, Santiago, Chile, December 2015.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[17] W. Liao, A. Pižurica, P. Scheunders, W. Philips, and Y. Pi, "Semisupervised local discriminant analysis for feature extraction in hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 184–198, 2013.

[18] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, "Weakly supervised learning for target detection in remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 4, pp. 701–705, 2014.

[19] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *International Journal of Computer Vision*, vol. 100, no. 3, pp. 275–293, 2012.

[20] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 5553–5563, 2016.

[21] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2015.

[22] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 10, pp. 1797–1801, 2014.

[23] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: benchmark and state of the art," *Proceedings of the IEEE*, vol. 99, pp. 1–19, 2017.

[24] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '15)*, pp. 36–43, June 2015.

[25] O. A. B. Penatti, K. Nogueira, and J. A. Dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '15)*, pp. 44–51, June 2015.

[26] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 6, pp. 3660–3671, 2016.

[27] C. Tao, Y. Tan, H. Cai, and J. Tian, "Airport detection from large IKONOS images using clustered sift keypoints and region information," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 1, pp. 128–132, 2011.

[28] Ö. Aytekin, U. Zongur, and U. Halici, "Texture-based airport runway detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 3, pp. 471–475, 2013.

[29] X. Yao, J. Han, L. Guo, S. Bu, and Z. Liu, "A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and CRF," *Neurocomputing*, vol. 164, pp. 162–172, 2015.

[30] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[31] X. Yao, J. Han, D. Zhang, and F. Nie, "Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3196–3209, 2017.

[32] L. Cao, C. Wang, and J. Li, "Vehicle detection from highway satellite images via transfer learning," *Information Sciences*, vol. 366, pp. 177–187, 2016.

[33] S. Gupta, S. Rana, B. Saha, D. Phung, and S. Venkatesh, "A new transfer learning framework with application to model-agnostic multi-task learning," *Knowledge and Information Systems*, vol. 49, no. 3, pp. 933–973, 2016.

[34] A. Van Opbroek, M. A. Ikram, M. W. Vernooij, and M. De Bruijne, "Transfer learning improves supervised image segmentation across imaging protocols," *IEEE Transactions on Medical Imaging*, vol. 34, no. 5, pp. 1018–1030, 2015.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[36] W. Zhao, Z. Guo, J. Yue, X. Zhang, and L. Luo, "On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery," *International Journal of Remote Sensing*, vol. 36, no. 13, pp. 3368–3379, 2015.

[37] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 11, pp. 2321–2325, 2015.

[38] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '10)*, pp. 270–279, ACM, San Jose, Calif, USA, November 2010.

[39] X. Wang, Q. Lv, B. Wang, and L. Zhang, "Airport detection in remote sensing images: a method based on saliency map," *Cognitive Neurodynamics*, vol. 7, no. 2, pp. 143–154, 2013.

[40] Y. Jia, E. Shelhamer, J. Donahue et al., "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678, ACM, Orlando, Fla, USA, November 2014.