# A Sample Update-Based Convolutional Neural Network Framework for Object Detection in Large-Area Remote Sensing Images

Yuan Hu, Xiang Li, Nan Zhou, Lina Yang, Ling Peng, and Sha Xiao

***Abstract*— This letter addresses the issue of accurate object detection in large-area remote sensing images. Although many convolutional neural network (CNN)-based object detection models can achieve high accuracy in small image patches, the models perform poorly in large-area images due to the large quantity of false and missing detections that arise from complex backgrounds and diverse groundcover types. To address this challenge, this letter proposes a sample update-based CNN (SUCNN) framework for object detection in large-area remote sensing images. The proposed framework contains two stages. In the first stage, a base model—single-shot multibox detector—is trained with the training data set. In the second stage, artificial composite samples are generated to update the training set. The parameters of the first-stage model are fine-tuned with the updated data set to obtain the second-stage model. The first- and second-stage models are evaluated using the large-area remote sensing image test set. Comparison experiments show the effectiveness and superiority of the proposed SUCNN framework for object detection in large-area remote sensing images.**

***Index Terms*— Convolutional neural networks (CNNs), large-area remote sensing images, object detection, sample update.**

## I. Introduction

OBJECT detection in optical remote sensing images involves the identification of the locations and class labels of predicted objects in satellite or aerial images. Object detection has a vital role in an extensive range of remote sensing applications, such as urban planning, environmental monitoring, and other civil and military applications [1].

In recent years, an increasing number of methods has been developed for the detection of various geospatial objects in optical remote sensing images, such as airplanes, ships, and storage tanks. These methods fall into two main categories: traditional machine learning methods and deep learning methods.

Manuscript received August 14, 2018; revised November 30, 2018; accepted December 19, 2018. Date of publication January 16, 2019; date of current version May 21, 2019. This work was supported by the Jiangsu Provincial Surveying Mapping and Geoinformation Research Project under Grant JSCHKY201720. *(Corresponding author: Ling Peng.)*

Y. Hu, X. Li, L. Yang, and L. Peng are with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: penglingqiqi@126.com).

N. Zhou is with Image Sky International Co., Ltd, Jiangsu 215163, China.

S. Xiao is with the Key Laboratory of Spatial Data Mining and Information Sharing, Fuzhou University, Fujian 350116, China.

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

Digital Object Identifier 10.1109/LGRS.2018.2889247

Traditional machine learning methods regard object detection as a classification problem. The traditional machine learning method has two main steps: feature extraction and classification. In [2], a rotation-invariant histogram of oriented gradient feature was employed as an input to a support vector machine (SVM) classifier to obtain the corresponding predicted labels and locations. Moranduzzo and Melgani [3] applied the scale-invariant feature transform feature in the feature extraction process and applied an SVM classifier to discriminate between key points assigned to cars and the background. Machine learning methods are dominated by handcrafted features and shallow learning-based features. Although these features have been successfully utilized for certain specific object detection tasks, the involvement of human ingenuity in feature design or shallow structures significantly influences the representational power and effectiveness of object detection.

In recent years, deep learning has achieved considerable success in an extensive range of computer vision applications, particularly object detection. Most deep learning-based object detection methods adopt convolutional neural networks (CNNs) to extract features from input images. Several papers have attempted to detect geospatial objects in high-resolution remote sensing images using a deep learning method [4]–[10]. For instance, a unified deep CNN named DeepPlane was proposed to simultaneously detect the positions and classify the categories of aircrafts in remote sensing images [5]. Han *et al.* [7] proposed an integrated geospatial object detection framework based on the faster region-based CNN (Faster R-CNN) for multiclass geospatial object detection (e.g., airplanes, ships, and storage tanks).

Despite the breakthrough in object detection by deep learning methods, one problem remains unresolved. Although CNN-based object detection models can achieve high accuracy for test samples (typically small image patches), they cannot be directly applied to efficiently handle large-area remote sensing images. Large-area remote sensing images contain not only targets but also a large quantity of background objects that may not have fully appeared in training samples. Patterns or objects similar to the predicted targets in these areas may be falsely detected, and objects with new shapes, scales, and colors that are not in the training data set are prone to being undetected, which would dramatically decrease the accuracy.

To address the problem, this letter proposes a sample update-based CNN (SUCNN) framework for object detection

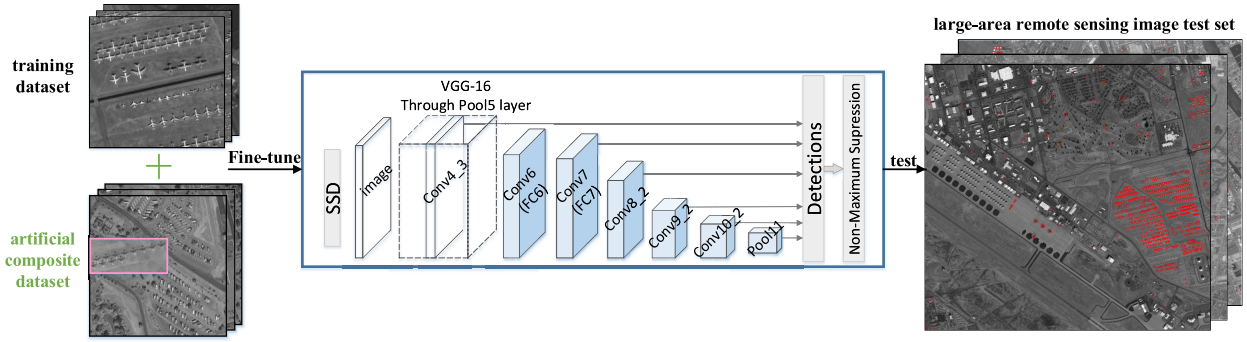


Fig. 1. Stage 2 pipeline of SUCNN framework for object detection in large-area remote sensing images.

in large-area remote sensing images. Some artificial composite samples are generated according to false and missing detections in the large-area remote sensing image validation set to combine the background images with some target objects. The background images are collected from the areas that are often falsely detected, especially areas that are located far from the predicted targets. The target objects are missing detections with new shapes, scales, and colors. The framework contains two stages of training, and the second stage of training absorbs the artificial composite samples, which dramatically improves the performance of the second-stage model for large-area remote sensing images compared with that of the first-stage model.

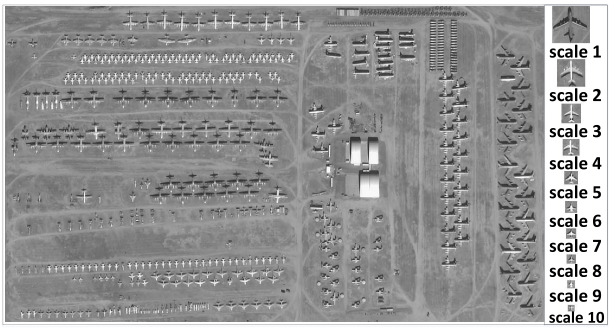


Fig. 2. Ten different scales of airplanes in an airport from a GF-2 satellite image.

## II. Methodology

### A. SUCNN Framework for Object Detection

Large-area remote sensing images generally cover an extensive range of ground types and a complex environment. Testing these images would cause numerous false and missing detections, which would dramatically reduce the detection accuracy. We propose a two-stage SUCNN framework to address this problem. In stage 1, the base model is trained with the training data set and the performance is evaluated using the large-area remote sensing image validation set. Many patterns are falsely detected because this background information is not trained as negative samples in the training data set. The background information is not trained because they are located far from the targets, which indicates that they cannot be included when collecting samples (such as 500 pixels $\times$ 500 pixels). For example, residential areas cannot be included when collecting airplane samples. Stage 2 training (as shown in Fig. 1) is needed to address this problem. Before starting stage 2 training, the data set is updated to introduce artificial composite samples using the proposed sample update method according to the false and missing detections in the large-area remote sensing image validation set. Details of the generation of the artificial composite samples are provided in Section II-C. Then, the first-stage model is fine-tuned using the updated data set to obtain the second-stage model. Finally, the second-stage model is evaluated using the large-area remote sensing image test set again.

### B. Base Model

As shown in Fig. 2, this airport has 10 different scales of airplanes. Many geospatial objects have multiple scales, such as airplanes, ships, and storage tanks. This characteristic complicates accurate object detection, and many deep learning-based object detection methods cannot achieve satisfactory results for multiscale objects. The majority of the methods employ the top-most layer of a convolutional layer in a CNN (ConvNet) to detect objects on different scales, such as Faster R-CNN [11] and R-FCN [12]. Although powerful, these methods impose a considerable burden on a single layer to model all possible object scales and shapes [13].

In this letter, we employ the single-shot multibox detector (SSD) as the base model to solve the multiscale object detection problem. The SSD is based on the notion that lower network layers have smaller receptive fields and finer details that are more suitable for detecting small objects. Conversely, higher layers are more suitable for detecting larger objects. Therefore, the SSD combines multiple layers within a ConvNet to form a strong multiscale detector. Each layer focuses on predicting objects of a certain scale.

The shape offsets and confidences for all object categories $(c_1, c_2, \cdots, c_p)$ are predicted for each default box. The total objective loss function is a weighted sum between the confidence loss (conf) and the localization loss (loc)

$$\mathrm{L}(x, c, l, g) = \frac{1}{N}(L_{\mathrm{conf}}(x, c) + \alpha L_{\mathrm{loc}}(x, l, g)) \qquad (1)$$

where $N$ is the number of default boxes, the localization loss is the smooth $L1$ loss between the ground-truth box $(g)$ parameter and the predicted box $(l)$ parameter, and the confidence loss is the softmax loss over multiple classes confidences $(c)$. Refer to [14] for details.

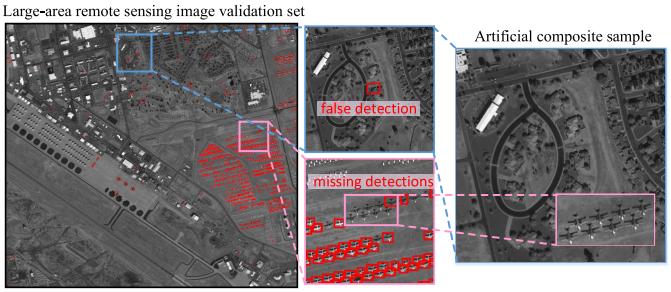


Fig. 3. Generation of artificial composite samples.

### C. Sample Update

To decrease the false detections that are located far from the predicted targets, additional background information must be encoded. To enrich the data set and reduce the number of missing detections, additional objects with different shapes, scales, and colors must be collected. An image, either a training sample or a test sample, must contain at least one positive sample to be fed into a deep learning object detection model. (Training samples and test samples are the images used for training and testing, respectively; positive samples and negative samples are the proposal boxes for targets and backgrounds in a training or test sample.) Therefore, our intuition is to combine the background images with missing detection objects.

According to this intuition, we propose a method of generating artificial composite samples. First, the first-stage model is tested using the large-area remote sensing image validation set; numerous false and missing detections would occur. Subimages (500 pixels $\times$ 500 pixels in this letter) are cropped as the background images centered on the false detections. Second, missing detections are chosen as the positive samples. Finally, positive samples are combined with the background images to generate the artificial composite samples (see Fig. 3), and the artificial composite samples are combined with the original data set to form the new training data set for stage 2 training. In this case, missing detections are trained as positive samples, and the false detections are trained as negative samples. Both false detections and missing detections will be reduced after the model is fine-tuned with the updated data set. The first-stage model evaluation using the validation set provides specific and effective guidance to achieve better performance with a minimum of artificial composite samples.

The number of positive samples in the artificial composite samples must be approximately equal to the number of positive samples in the samples of the training data set. If the former number is substantially less than the latter number, the accuracy will be reduced because samples with fewer positive samples will dilute the number of positive samples in each iteration of training, which reduces the likelihood that positive samples will be trained.

## III. Experiments and Discussion

### A. Data Set Description

In the training data set, 500 images are collected and labeled with three object categories: airplanes, ships, and storage tanks. The size of each image is 500 pixels $\times$ 500 pixels. These images contain 7561 airplanes, 1312 ships, and 2338 storage tanks. Artificial composite samples are generated in stage 2 training, including 48 composite airplane samples, 20 composite ship samples, and 36 composite storage tank samples. Three large-area remote sensing images in the large-area remote sensing image test set were employed to test each object category. The size of each image is 4000 $\times$ 4000 pixels. The images have a large area, complex backgrounds, and an extensive variety of ground cover types. In addition to the three object categories, there are residential areas, industrial areas, grasslands, and woodlands in the images.

### B. Evaluation Metrics

We adopt the precision–recall curve (PRC) and average precision (AP) to quantitatively evaluate the performance of the object detection method. A certain number of indexes are needed, namely, true positive (TP), false positive (FP), and false negative (FN). TP denotes the number of correct detections, FP denotes the number of false detections, and FN denotes the number of missing detections.

PRC: The precision metric measures the fraction of detections that are TP. The recall metric measures the fraction of positives that are correctly retrieved. The precision and recall metrics are defined as

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{3}$$

A detection is considered TP if the intersection-over-union overlap between the ground-truth box and the predicted box exceeds 0.5. Otherwise, the detection is an FP.

AP: The AP metric computes the area under the PRC. A higher AP value indicates better performance, and vice versa.

### C. Base Model Selection Experiments

To select the base model for the SUCNN framework, we compared the SSD model with the Faster R-CNN and R-FCN. We adopted the same training and test data sets for the three models. Parameter optimization experiments were conducted for all three models to obtain their best performance on the data set.

Fig. 4 shows the quantitative comparison results of the three different methods measured by the AP values and PRCs. The SSD outperforms the Faster R-CNN and R-FCN for all three object classes in terms of AP by a large margin. This result demonstrates the superiority of SSD for detecting multiscale and dense objects. Thus, the SSD is chosen as the base model for the SUCNN.

### D. Comparison Experiments for the SUCNN

Two stages of training exist for the SUCNN framework. For the first stage of training, the base model—SSD—is trained using the training set. For the second stage of training, the artificial composite samples are generated according to the

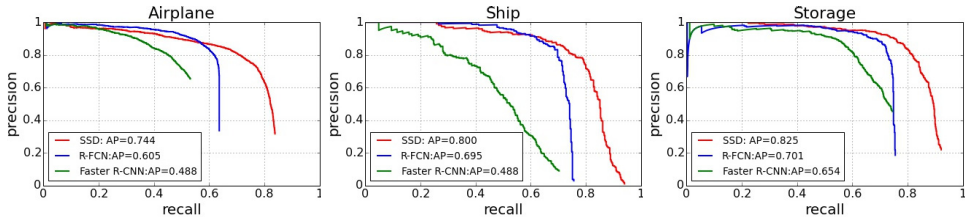


Fig. 4. PRCs of the SSD, Faster R-CNN, and R-FCN for airplanes, ships, and storage tanks, respectively.

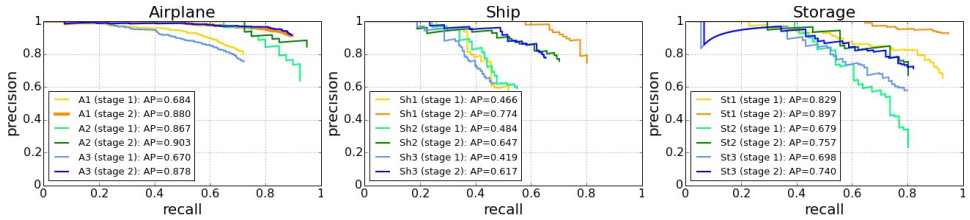


Fig. 5. PRCs of the two-stage models of the SUCNN on large-area remote sensing images for airplanes, ships, and storage tanks. A1–A3 for Airplane Images 1–3, respectively; Sh1–3 for Ship Images 1–3, respectively; and St1–3 for Storage Images 1–3, respectively.

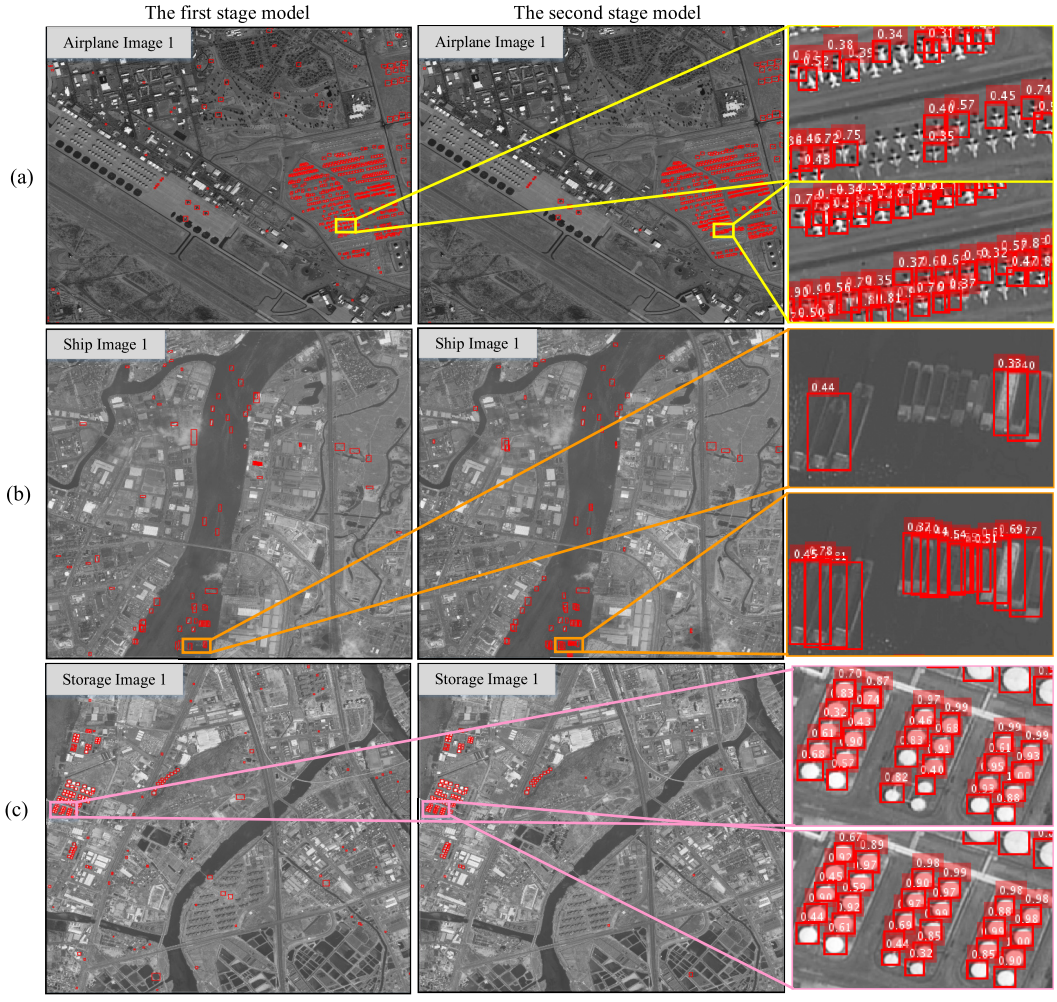


Fig. 6. Detection results using the two-stage models of the SUCNN for large-area images that contain (a) airplanes, (b) ships, and (c) storage tanks.

false and missing detections on the large-area remote sensing image validation set. The training set is updated using the artificial composite samples, and the first-stage model is fine-tuned to obtain the second-stage model.

The performance of the first- and second-stage models on the large-area remote sensing image test set are compared in Fig. 5. The second-stage model of the SUCNN obtained a remarkable boost over all three object classes in terms

TABLE I
PERFORMANCE COMPARISONS OF THE TWO-STAGE MODELS OF THE SUCNN FOR NINE LARGE-AREA REMOTE SENSING IMAGES

| Test image | Models | Average precision (AP) |
|---|---|---|
| Airplane Image 1 | stage 1 | 0.684 |
| | stage 2 | 0.88 |
| Airplane Image 2 | stage 1 | 0.867 |
| | stage 2 | 0.903 |
| Airplane Image 3 | stage 1 | 0.67 |
| | stage 2 | 0.878 |
| Ship Image 1 | stage 1 | 0.466 |
| | stage 2 | 0.774 |
| Ship Image 2 | stage 1 | 0.419 |
| | stage 2 | 0.617 |
| Ship Image 3 | stage 1 | 0.484 |
| | stage 2 | 0.647 |
| Storage Image 1 | stage 1 | 0.829 |
| | stage 2 | 0.897 |
| Storage Image 2 | stage 1 | 0.679 |
| | stage 2 | 0.757 |
| Storage Image 3 | stage 1 | 0.698 |
| | stage 2 | 0.74 |

of AP and PRC. Detailed performance comparisons of the two-stage models of the SUCNN on nine large-area images are shown in Table I.

Fig. 6 shows the object detection results of the two-stage models of the SUCNN on three example images. For Airplane Image 1, numerous false detections appear in residential areas, woodlands, and certain parking lots in the detection results of the first-stage model. However, these areas are clean on the corresponding resulting image from the second-stage model. For Ship Image 1, numerous similar patterns in industrial areas, farmlands, and bridges are falsely detected using the first-stage model. However, only a few false detections occur in the resulting second-stage model detection image. For Storage Image 1, several round homogeneous areas of vegetation on both sides of the road and bare lands in residential and industrial areas are falsely detected using the first-stage model, whereas few areas are detected using the second-stage model. The first two columns in Fig. 6 show the effectiveness of the SUCNN in suppressing false detections on large-area images. The magnification of a small partial area in the last column shows the superiority of the SUCNN in detecting extremely dense objects. The first-stage model misses numerous boxes in dense objects. Conversely, the second-stage model of the SUCNN has an impressive performance in the same areas.

### *E. Discussion*

The comparison experiments demonstrate the effectiveness and superiority of the SUCNN in large-area remote sensing image testing, which stems from the sample update mechanism. First, the updated data set includes more background information that is located far from the objects, which enables the model to learn more diverse negative samples and distinguish more difficult patterns, which are highly similar to the predicted targets. Therefore, these patterns or areas will not be falsely detected. Second, the positive samples in the artificial composite samples are critical to the reduction of missing detections. We crop the missing detections on large-area images as the positive samples, such that the model can identify more diverse samples.

## IV. CONCLUSION

This letter proposed a novel and effective object detection framework—SUCNN—to improve the object detection accuracy in large-area remote sensing images. Artificial composite samples are generated using the proposed sample update method. The sample update mechanism substantially improves the second-stage model by combining the background images with target objects according to the false and missing detections. The quantitative comparison results over three object categories demonstrate the effectiveness and superiority of the proposed method for object detection in large-area remote sensing images.

## REFERENCES

[1] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.
[2] 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 Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 74–78, Jan. 2014.
[3] T. Moranduzzo and F. Melgani, "A SIFT-SVM method for detecting cars in UAV images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, New York, NY, USA, Jul. 2012, pp. 6868–6871.
[4] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
[5] H. Wang, Y. Gong, Y. Wang, L. F. Wang, and C. H. Pan, "DeepPlane: A unified deep model for aircraft detection and recognition in remote sensing images," *J. Appl. Remote Sens.*, vol. 11, p. 10, Sep. 2017.
[6] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3652–3664, Aug. 2017.
[7] X. Han, Y. Zhong, and L. Zhang, "An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery," *Remote Sens.*, vol. 9, p. 666, Jul. 2017.
[8] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.
[9] G. S. Xia *et al.* (2018). "DOTA: A large-scale dataset for object detection in aerial images." [Online]. Available: https://arxiv.org/abs/1711.10398
[10] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.
[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[12] J. Dai, Y. Li, K. He, and J. Sun. (2016). "R-FCN: Object detection via region-based fully convolutional networks." [Online]. Available: https://arxiv.org/abs/1605.06409
[13] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. (2017). "DSSD: Deconvolutional single shot detector." [Online]. Available: https://arxiv.org/abs/1701.06659
[14] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Computer Vision—ECCV*, vol. 9905, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.