

Received August 21, 2020, accepted September 10, 2020, date of publication September 28, 2020, date of current version October 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027044

Patch-Based Three-Stage Aggregation Network for Object Detection in High Resolution Remote Sensing Images

BING SUI¹, MENG XU², AND FENG GAO³

¹Hunan Key Laboratory of Meteorological Disaster Prevention and Reduction, Changsha 410007, China

²Jiangsu Climate Center, Nanjing 210009, China

³Hunan Glonavin Information Technology Company Ltd., Changsha 410205, China

Corresponding author: Feng Gao (38492913@qq.com)

This work was supported by the National Natural Science Foundation of China (NSFC) (No. 61902420).

ABSTRACT High-resolution remote sensing image object detection plays an increasingly important role in image processing and interpretation. The application of region-based convolutional neural network (R-CNN) greatly enhances the performance of object detection. However, the attributes of remote sensing images such as overlage image size, similar background, disequilibrium distribution of categories make this task more challenging. The previous works have focused on extracting multi-scale features of region proposals, often ignoring the quality of region of interest (ROI). In this work, we proposed a patch-based three-stage aggregation network (PTAN) for object detection in high-resolution remote sensing images. It consists of a three-stage cascade structure that sequentially improves the quality of candidate regions by increasing the IoU threshold stage by stage, and adopts a resampling strategy to obtain sufficient region proposals. At the same time, we also proposed patch-based strategy and applied it to the framework during training and inference. Ablation experiments and comprehensive evaluations on a communal remote sensing image object detection dataset DOTA demonstrate the effectiveness and robustness of the proposed framework, which obtained a mean average precision (mAP) value of 0.7958 on validation dataset and a front-rank mAP of 0.7858 on testing dataset. On another remote sensing image object detection dataset NWPU VHR-10, the proposed PTAN obtained a mAP value of 0.9187, outperforming other five object detectors.


INDEX TERMS High-resolution remote sensing image, object detection, cascade network, aggregation structure, region proposal quality.

I. INTRODUCTION

Object detection in high resolution remote sensing image is a core issue of image analysis and interpretation, which mainly includes two tasks: classification and regression [1]. The classification task labels the category of each predicted object while the regression task locates the coordinates of objects in the image [2]. In recent years, as a rapidly developing machine learning method in the environment of big data and high-performance computing, deep learning technology shows the attributes of the powerful ability for feature extraction and expression. It has been widely applied in the fields of natural image scene classification, semantic segmentation, object detection [3]–[5]. Convolutional neural

network (CNN) is a kind of feedforward neural network that contains convolutional computation with deep structure. It is one of the representative algorithms of deep learning. CNN extracts the deep features of the input image layer by layer, possessing the attributes of local receptive fields, shared weights, and pooling [6]. Therefore, how to apply CNN to remote sensing image object detection has attracted more and more attention.

In the field of natural image processing, the application of CNN has greatly improved the performance of object detection, and many outstanding algorithms have emerged. These methods are mainly divided into the region-based methods and the region-free methods [7]. The region-based methods [8]–[11] consist of two stages: candidate region generation, precise object classification and location. They first adopt region proposal algorithms to generate a mass of

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin .

candidate boxes, labeling as positive and negative samples. Then these class-agnostic candidate regions are sent to subsequent detectors for accurate classification and positioning. These methods possess higher detection accuracy but slower speed. The region-free methods [12]–[14] consider object detection as a regression problem, they do not generate region proposals but predict the class confidence and coordinates directly. They greatly improve the detection speed, although damaging some precision. Object detection in natural image has made significant progress, however, compared with natural images, remote sensing images show the following attributes [15], which make the classical object detection algorithms inapplicable to remote sensing images:

- 1) Overlarge image size. The size of remote sensing image may larger than 10,000 pixels, making processing more time-consuming and memory-consuming.
- 2) Disequilibrium distribution of target size and quantity between different categories. In remote sensing images, some objects densely appear such as ships while some objects scatteredly appear such as ground track fields. Meanwhile, the size of these two categories are not of the same magnitude.
- 3) Imaging perspective. Remote sensing images are obtained by satellite-borne or space-borne sensors from a top-down view, hence they easily influenced by weather and illumination conditions. Besides, the occlusion, shadow, semblable background and border sharpness factors also affect remote sensing images.

In order to address the aforementioned problems, numerous remote sensing image object detection algorithms have been proposed. Guo *et al.* [2] presented a unified multi-scale CNN, the base network can produce feature maps with different receptive fields to be responsible for objects with different scales. Zhang *et al.* [16] proposed a hierarchical robust CNN, which first adopts multi-scale convolutional features to represent the hierarchical spatial semantic information, then multiple fully connected layer features are stacked together to enhance the detection performance. Zhang *et al.* [17] utilizes inherent multi-scale pyramidal features and combines the strong-semantic, coarse-resolution features and the weak-semantic, high-resolution features simultaneously for remote sensing image object detection.

The above algorithms aim to explore the effectiveness of multi-scale features and achieve benign results, but they ignore the fact that in object detection issues, the higher the intersection over union (IoU) between the candidate boxes obtained by region proposal algorithm and ground truth, the higher the accuracy of detection results obtained by the detector [18]. The IoU threshold is used to label whether a sample is positive or negative. Obviously, when the threshold value u of positive samples is higher, the detector has a stronger ability to distinguish positive samples from negative samples, so the detection performance will be better. Therefore, the starting point of this work is to obtain positive candidate boxes with higher IoU threshold and keep the quantity sufficient at the same time

To obtain sufficient positive region proposals with high IoU threshold, and achieve high precision remote sensing image object detection, we propose an end-to-end framework, namely, Patch-based Three-stage Aggregation Network (PTAN) in this work. It first adopts the region proposal network (RPN) [10] to generate numerous candidate boxes, and then obtain positive samples with high IoU threshold through a three-stage aggregation network. Finally, these positive samples are sent to the subsequent detectors to achieve accurate object detection. The main contributions of this paper are summarized as follows:

- 1) We propose an effective three-stage aggregation network. During training, the proposed PTAN takes the region proposals with lower IoU threshold as input and outputs region proposals with higher IoU threshold at each stage, while the number of candidate boxes remained unchanged.
- 2) During inference, we also adopt three-stage cascade structure. The quality of the region proposals are sequentially improved.
- 3) Considering the overlarge size of remote sensing image, we divided the image into patches with a certain degree of overlap both during training and inference. The patch-based mechanism can effectively improve the performance of object detection.

The ablation experiments were conducted on two remote sensing image object detection datasets DOTA [19] and NWPU VHR-10 [20]. The rest of this paper is organized as follows. Section II introduces works related to this paper. Section III expounds the proposed framework in detail. Section IV elaborates the experimental data, evaluation criteria, settings and ablation experiments. Finally, the conclusions are drawn in Section V.

II. RELATED WORKS

We adopt region-based object detection method as our infrastructure. In this section, we first describe the workflow of region-based method, and then analyze the influence of IoU threshold value of region proposals in object detection task.

A. REGION-BASED OBJECT DETECTION METHODS

The region-based methods, including R-CNN [8], SPP-Net [21], Fast R-CNN [9], Faster R-CNN [10], FPN [11] and Cascade R-CNN [18], first generate abundant candidate boxes by region proposal algorithms [10], [22], [23]. These candidate boxes are labeled as positive or negative samples according to their IoU value with corresponding ground truth. They are sent to the subsequent detector, which employs CNN for feature extraction and perform refined category classification and coordinate regression.

Faster R-CNN [10] integrates the above steps into a unified framework and realizes end-to-end object detection. It consists of two modules, namely, RPN and Fast R-CNN, and they share convolutional features. The general technological

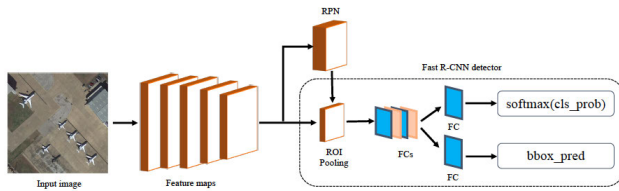


FIGURE 1. The architecture of Faster R-CNN, which consists of RPN and Fast R-CNN modules. These two modules share convolutional features to realize end-to-end object detection.

process of Faster R-CNN is shown in Figure 1. For a given image, the shared convolutional layers are utilized to extract features. The class-agnostic candidate regions with fixed scales and aspect ratios are obtained at each location on the last convolutional layer, also known as “anchor”. The mapped anchors can traverse every corner of the original image thus getting abundant region proposals. The positive and negative samples are selected by calculating the IoU thresholds between these candidate regions and the ground truth, and they are sent to Fast R-CNN [9] for further object detection. The shared convolution attribute of Faster R-CNN greatly speeds up the object detection, thus achieving real-time process.

B. THE INFLUENCE OF IOU THRESHOLD BETWEEN REGION PROPOSAL AND GROUND-TRUTH

In the object detection task, how to efficiently extract the features from region proposal is of great significance. In the previous remote sensing image object detection works [17], [24], [25], the focuses are to extract multi-scale or cross-scale features, and try to adopt syncretic features to extract the object. One thing we all know is that the positive samples of higher IoU threshold value with ground-truth are send into the object detector, the higher the final detection accuracy. In order to balance the quality and quantity of candidate boxes, the threshold values u is quite loose, typically $u = 0.5$. If simply increasing the threshold value, the quality of candidate box will be promoted, but the quantity will be reduced at the same time, which will lead to overfitting problem during training. Besides, the threshold values used for training are not exactly the same with those used for inference. We can not calculate the IoU threshold value because of the lack of ground truth during inference, so we can only treat all the region proposals as positive samples, which will bring about the problem of mismatch. Based on the above analysis, we constructed a three-stage cascade network with training sequentially, using the output of one stage to train the next. [18] has revealed that the output IoU of a regressor is almost invariably better than the input IoU. Therefore, the focus of this work is to gradually increase the IoU threshold value of positive samples while keeping the quantity unchanged, that is to say, to improve the quality of region proposals and extract more efficient features for accurate object detection in remote sensing images.

III. METHODOLOGY

In this section, we first elaborate the architecture of the proposed PTAN in detail, and then introduce how the cascade network works during training and inference. Finally, a patch-based training and inference strategy is proposed to enhance the performance of PTAN.

A. OVERVIEW OF THE PROPOSED FRAMEWORK

The overview structure of the proposed PTAN is shown in Figure 2, which is built on the work of Faster R-CNN [10], consisting of an RPN module and a three-stage aggregation network module. RPN is a kind of fully convolutional network [26], which can deal with the arbitrary-size image through the “anchor” mechanism, and numerous class-agnostic region proposals with an objectness score are generated. These candidate boxes will be sent to the subsequent detectors for accurate category classification and coordinate regression.

To improve detection accuracy, in other words, to generate region proposals of higher overlap with ground truth, we can sequentially increase the IoU threshold value stage-by-stage to produce positive samples. The subsequent detector receives proposals with higher confidence and will naturally obtain high-precision detection results. As shown in Figure 2, We first sent the region proposals generated by RPN into the first-stage regressor, which was provided with the lowest IoU threshold. After that, the output candidate boxes obtain larger IoU threshold values than the input candidate boxes. We adopt the resampling strategy to select same number of candidate boxes in the previous stage and sent them to the next stage. Iterating through this sequence until the candidate boxes with relatively highest IoU threshold are used in the third-stage for object detection.

We employ the form similar to Faster R-CNN [10] to minimize the loss function. The difference is that it relies on a cascade of specialized regressors, that is, the regressor of the current state is closely related to the regressor of the previous stage. For a region proposal x^t at t stage, we define its class label y^t as follow:

$$y^t = \begin{cases} g_{y^t}, & \text{IoU}(x^t, g) \geq u^t \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where g_{y^t} is the class label of the ground truth object g , u^t is the IoU threshold value at stage t . If the IoU between x^t and the corresponding ground truth g is greater than the specified threshold u^t , the region proposal x^t is considered an example of the class. At each stage t , the detector includes a classifier h_t and a regressor f_t optimized for IoU threshold u^t , noting that $u^t > u^{t-1}$. With these analyses, the loss function at each stage t is defined as:

$$L(x^t, g) = L_{cls}(h_t(x^t), y^t) + \lambda[y^t \geq 1]L_{loc}(f_t(x^t, b^t), g)p \quad (2)$$

where the L_{cls} item represents the classification loss, which adopts cross-entropy loss function. $h_t(x^t)$ is a $K+1$ -dimensional (including background) estimate of the

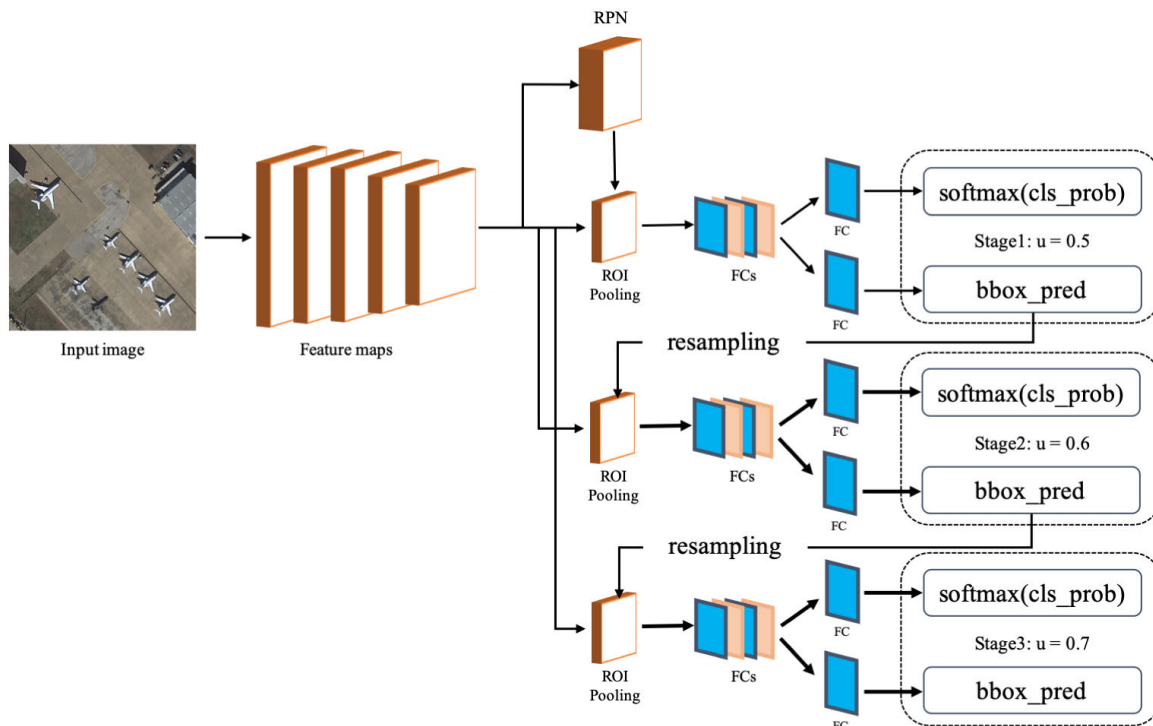


FIGURE 2. The overview of PTAN, which consists of an RPN module and a three-stage aggregation network module for producing region proposals with higher IoU threshold.

posterior distribution over classes, which assigns an image sample x^t to one of $K+1$ classes. y^t is the corresponding class label defined by equation (1). λ is a balancing parameter while $[y^t \geq 1]$ is the class identification function. Only samples with class label participate in the calculation of coordinate regression. We set $\lambda = 1$ in this work. The L_{loc} item represents regression loss, which adopts a smoothed L1 loss function as in Fast-RCNN [9]. The task of bounding box regression is to regress a candidate bounding box b^t into its corresponding ground truth \mathbf{g} for a candidate bounding box x^t , utilizing a regressor $f_t(x^t, b^t)$. Attentionally, a candidate box b^t can be represented by its upper-left coordinates (x, y) as well as its width w and height h , formally, $b^t = (b_x, b_y, b_w, b_h)$. We note that the regressor of the current state is closely related to the regressor of the previous stage, thus $b^t = f_{t-1}(x^{t-1}, b^{t-1})$, which guarantees that the network was trained sequentially with increasing IoU thresholds and enables high quality object detection in remote sensing images.

B. HOW AGGREGATION NETWORK WORKS DURING TRAINING AND INFERENCE

RPN generates about 2000 proposals during training, and these proposals are sent into the first stage of three-stage aggregation network, which calculates the IoU between each proposal and corresponding ground truth according to IoU threshold specified at the current stage. These proposals

are divided into positive samples (foreground) and negative samples (background), which are sampled so that the ratio between positive and negative is 1:3, and the total number of the two is usually 128. After that, the proposals are sent into ROI Pooling layer for feature extraction and the following classification and box regression. Through the first stage, we can obtain the candidate boxes with higher IoU, then we adopt resampling strategy to adjust the distribution of the proposals and select the same number of candidate boxes to send to the next stage for training. Through three stages of cascading operations, high quality proposals are available for fine-grained object detection.

We still utilize the same cascading operation during inference. However, we are unable to calculate the IoU due to the lack of corresponding ground truth, so we directly sent the generated region proposals to ROI Pooling layer for processing. Because of our aggregation structure, the quality of the candidate boxes is sequentially improved, and they are sent into higher quality detector, thus achieving high precision object detection.

C. A PATCH-BASED TRAINING AND INFERENCE MECHANISM

Due to the overlage size of remote sensing images, semantics information will be lost if they are directly sent to the network for processing. Therefore, we adopt patch-based strategy during training and prediction, and the flowchart is shown

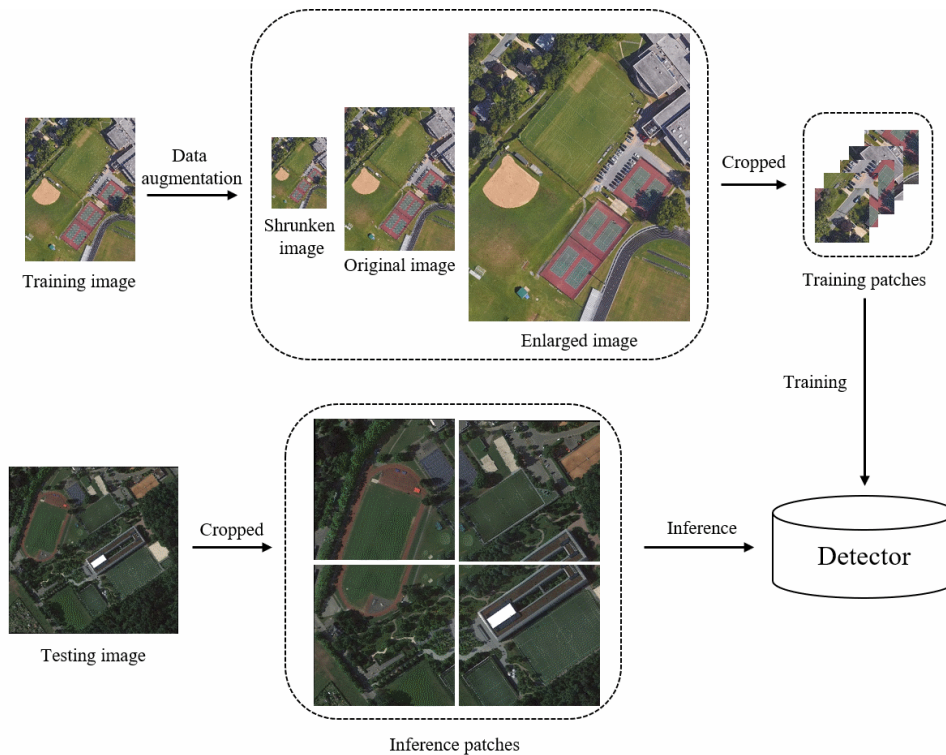


FIGURE 3. The flowchart of patch-based training and inference strategy.

in Figure 3. Considering the uneven distribution of objects in remote sensing images and the difference of size between different categories, we enlarge and shrink remote sensing images by a factor of 2 and 0.5 respectively. The enlarged samples enhance the resolution features of the small objects while the shrunken samples integrally clip the large objects into a single patch for processing. At the same time, through such operations we effectively enrich the training samples. We cropped the set of original images, enlarged images and shrunken images into patches, and the size of each piece is 1000×1000 pixels with an overlap of 500 pixels. Finally, 30,000 samples are randomly selected for training. During the prediction process, we sliced the test images into patches of 1300×1300 pixels with an overlap of 200 pixels for speeding up object detection.

IV. EXPERIMENTS AND RESULT ANALYSIS

A. EXPERIMENTAL DATASET

We validated the effectiveness of the proposed PTAN framework on DOTA [19] and NWPU VHR-10 [20] datasets. DOTA consists of 2806 remote sensing images from different sensors and platforms. Each image ranges in size from 800×800 to 4000×4000 pixels and contains objects of a wide variety of scales. The composition of DOTA dataset includes 1,411 training images, 458 validation images and 937 testing images. DOTA dataset covers 15 categories, namely, plane, ship, storage tank, baseball diamond, tennis court,

basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field and swimming pool, with a total of 188,282 annotated instances. DOTA dataset includes two forms of object detection. Task 1 adopts an oriented bounding box as ground truth. Task 2 utilizes a horizontal bounding box as ground truth. In this work, we only focus on the horizontal bounding box detection task with the form of $(x_{min}, y_{min}, x_{max}, y_{max})$.

The NWPU VHR-10 dataset is a multi-class, multi-source and multi-resolution remote sensing image object detection dataset. It contains a total of 800 images covering 10 categories, namely, airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge and vehicle. The average image size of NWPU VHR-10 dataset is about 600×800 pixels. We did not adopt patch-based strategy during training and prediction on this dataset. The NWPU VHR-10 dataset includes two subsets, formally, a positive subset with 650 images and a negative subset with 150 images. Each instance in the positive subset is manually annotated with horizontal bounding box. We conducted ablation experiments on the positive subset and randomly selected 323 images for training and 327 images for testing.

B. EVALUATION CRITERIA

We utilized Precision-Recall Curve (PRC) and Average Precision (AP) as evaluation criteria in our experiments, which are

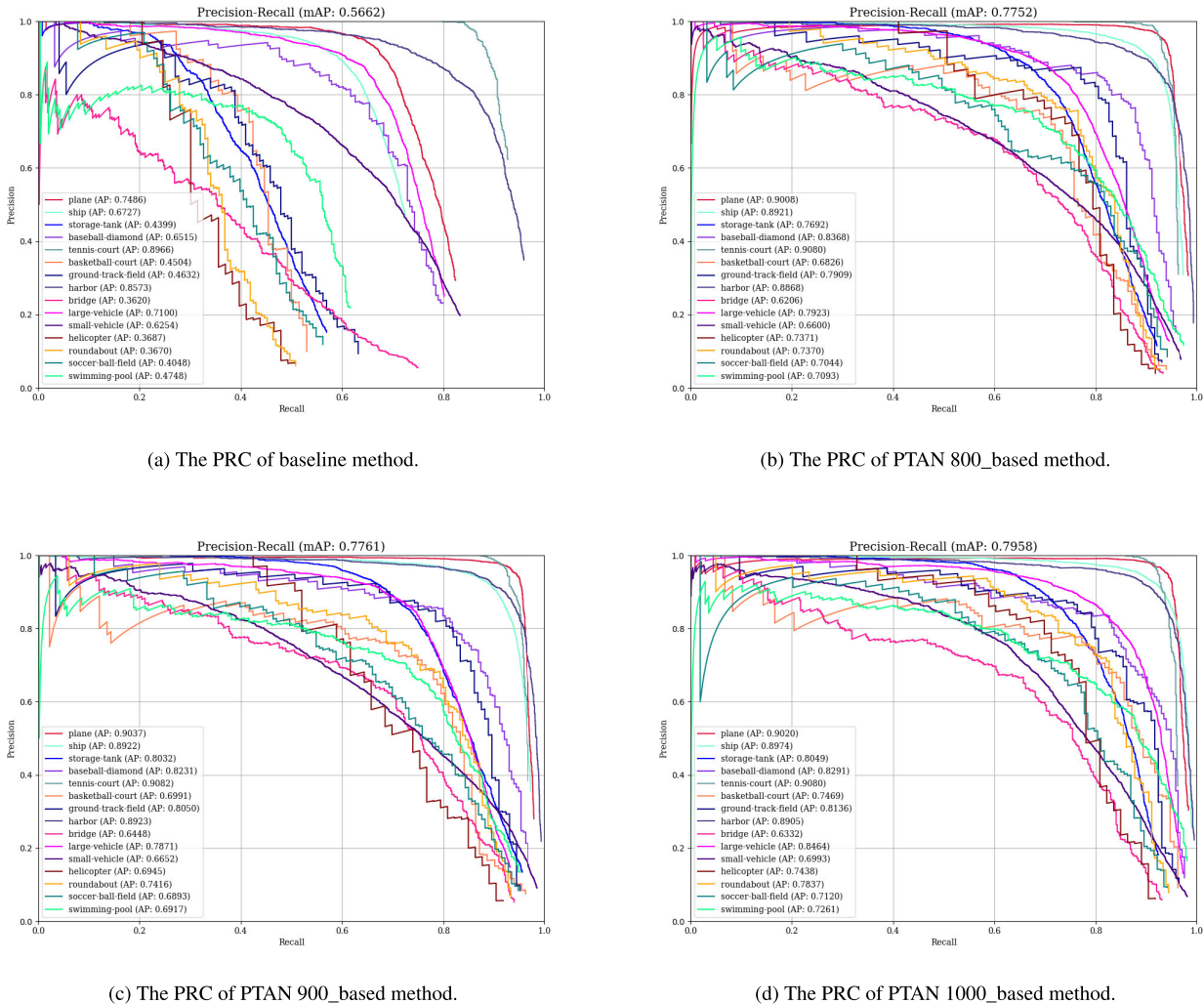


FIGURE 4. The PRC curves of internal ablation experiments.

related to precision metric and recall metric. The precision metric represents the proportion of the number of correctly detected objects to the total number of detected objects. The recall metric represents the proportion of the number of correctly detected objects to the total number of labeled objects. The PRC indicates the relationship between the precision criterion and the recall criterion. For a certain category, whose detector is considered outstanding if its precision stays high as recall increases. Average Precision (AP) represents the averaged precision ranges all recall values with the interval of [0, 1]. From the perspective of numerical quantification, AP is equals to the area under the PRC. In general, a splendid detector is accompanied by a high AP value. Mean average precision (mAP) is the mean AP over all classes.

C. EXPERIMENTAL SETTINGS

During training, the IoU threshold values for each stage are 0.5,0.6 and 0.7, respectively. We selected cascade R-CNN

[18] as the baseline method and ResNet50 as the backbone. For DOTA dataset, we trained a total of 600k iterations with a learning rate of 0.0025 for the first 300k iterations, 0.00025 for the next 100k iterations, and 0.000025 for the remaining 200k iterations. For NWPU VHR-10 dataset, we trained a total of 6000 iterations and the learning rate decay strategy is the same as in the DOTA dataset. The patch-based three-stage aggregation network was trained by stochastic gradient descent (SGD) algorithm with a mini-batch of 1 image and 128 region proposals in each iteration. Weight decay and momentum are 0.0001 and 0.9 respectively.

D. ABLATION EXPERIMENTS ON THE DOTA DATASET

We demonstrated the effectiveness of the proposed network and the patch-based strategy through two experiments on the DOAT dataset, namely, internal ablation experiments and external ablation experiments. The internal ablation



FIGURE 5. The detection results of PTAN on DOTA dataset.

experiments constantly change the parameters of the framework and compare with the baseline method. The external ablation experiments compare with state-of-the-art methods. We note that the contrasted methods are implemented in their original environments without any additions.

1) INTERNAL ABLATION EXPERIMENTS

We implemented four sets of ablation experiments and the results are shown in Table 1. We adopted (a), (b), (c), (d) to represent each method and the bold font to indicate the

highest value in each row. The baseline method in Table 1(a) is Cascade R-CNN [18]. The numbers 800, 900 and 1000 in 1(b), 1(c) and 1(d) respectively represent that we resized the input image to the corresponding size. Comparing Table 1(a) with 1(b), (c), (d), we found that the performance of the proposed framework in this paper is far better than that of the baseline method and the maximum mAP improvement can achieve 22.96% (0.5662 to 0.7958), which indicates that the proposed PTAN is suitable for remote sensing image object detection. Comparing Table 1(b) with 1(c), (d), we saw

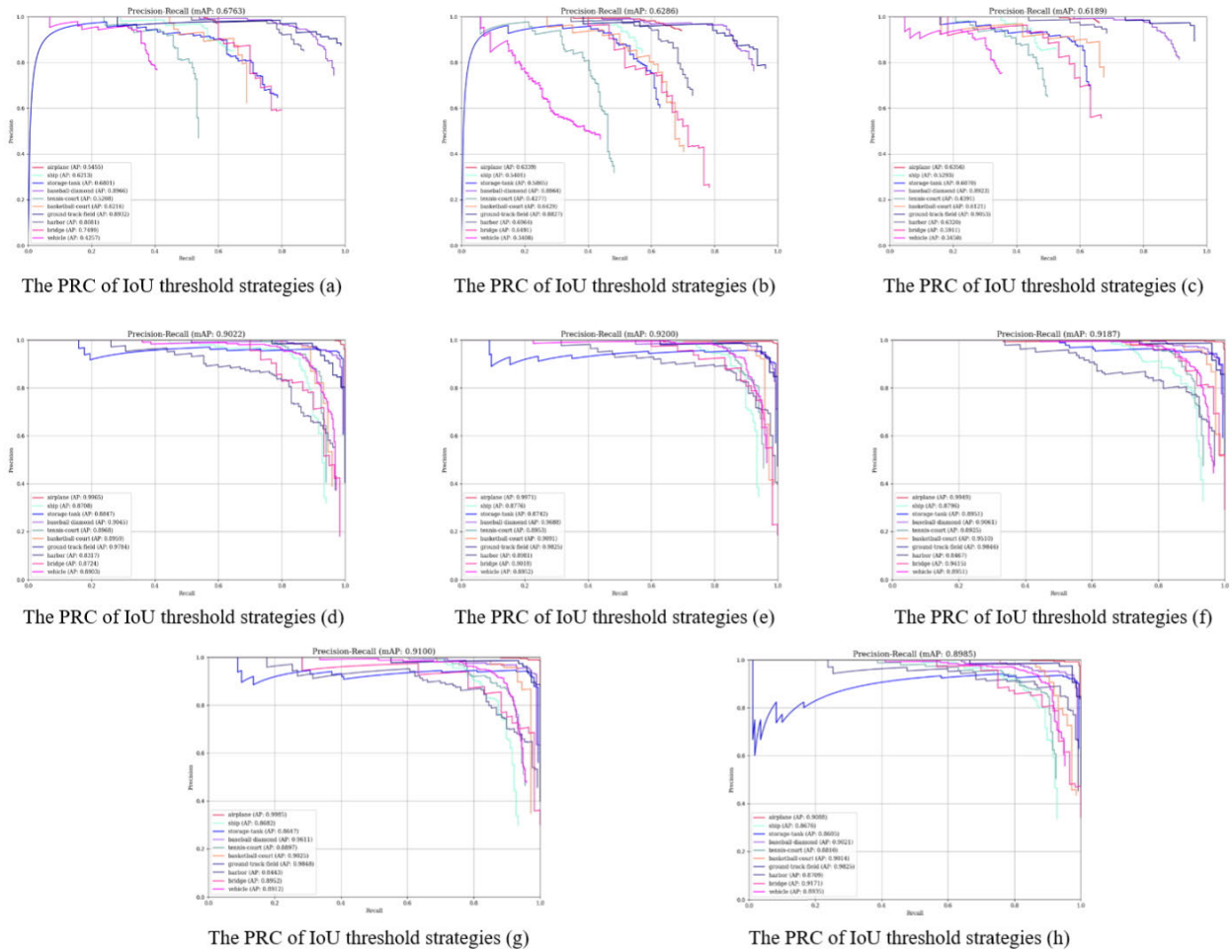


FIGURE 6. The PRCs of different network structures and IoU threshold strategies on NWPU VHR-10 dataset.

TABLE 1. The AP values of internal ablation experiments.

Method	baseline(a)	PTAN 800_based(b)	PTAN 900_based(c)	PTAN 1000_based(d)
plane	0.7486	0.9008	0.9037	0.9020
ship	0.6727	0.8921	0.8922	0.8974
storage tank	0.4399	0.7692	0.8032	0.8049
baseball diamond	0.6515	0.8368	0.8231	0.8291
tennis court	0.8966	0.9080	0.9082	0.9080
basketball court	0.4504	0.6826	0.6991	0.7469
ground track field	0.4632	0.7909	0.8050	0.8136
harbor	0.8573	0.8868	0.8923	0.8905
bridge	0.3620	0.6206	0.6448	0.6332
large vehicle	0.7100	0.7923	0.7871	0.8464
small vehicle	0.6254	0.6600	0.6652	0.6993
helicopter	0.3687	0.7371	0.6945	0.7438
roundabout	0.3670	0.7370	0.7416	0.7837
soccer ball field	0.4048	0.7044	0.6893	0.7120
swimming pool	0.4748	0.7093	0.6917	0.7261
mAP	0.5662	0.7752	0.7761	0.7958

that as the size of the input sample increases, so does the detection accuracy. It is worth noting that the accuracy of 1(d) has been greatly improved, achieving a mAP of 0.7958. We deem that on the one hand, the input image size increases, and on the other hand, the image size of 1(d) matches the size of the sample itself.

The quantified PRCs over four ablation experiments are plotted in Figure 6. We also visualized some detection results as shown in Figure 5.

TABLE 2. The AP values of external ablation experiments.

Method	FPN	Cascade R-CNN	Ours
plane	0.7072	0.7486	0.9020
ship	0.6073	0.6727	0.8974
storage tank	0.3864	0.4399	0.8049
baseball diamond	0.6043	0.6515	0.8291
tennis court	0.8155	0.8966	0.9080
basketball court	0.4193	0.4504	0.7469
ground track field	0.4190	0.4632	0.8136
harbor	0.7699	0.8573	0.8905
bridge	0.2631	0.3620	0.6332
large vehicle	0.6706	0.7100	0.8464
small vehicle	0.5276	0.6254	0.6993
helicopter	0.2976	0.3687	0.7438
roundabout	0.2589	0.3670	0.7837
soccer ball field	0.3784	0.4048	0.7120
swimming pool	0.3235	0.4748	0.7261
mAP	0.4966	0.5662	0.7958

2) EXTERNAL ABLATION EXPERIMENTS

We compared our framework with other region-based object detection networks mainly including FPN [11] and Cascade R-CNN [18]. We made no additional changes to these methods and the comparison results are shown in Table 2.

TABLE 3. Ablation experiments of different network structures and IoU threshold strategies on NWPU VHR-10 dataset.

networks	one stage					three stages			
	IoU thresholds	0.5(a)	0.6(b)	0.7(c)	(0.3, 0.4, 0.5)(d)	(0.4, 0.5, 0.6)(e)	(0.5, 0.6, 0.7)(f)	(0.6, 0.7, 0.8)(g)	(0.7, 0.8, 0.9)(h)
airplane	0.5455	0.6339	0.6356	0.9965	0.9971	0.9949	0.9985	0.9088	
ship	0.6213	0.5401	0.5293	0.8708	0.8776	0.8796	0.8682	0.8676	
storage tank	0.6801	0.5865	0.607	0.8847	0.8742	0.8951	0.8647	0.8605	
baseball diamond	0.8966	0.8864	0.8923	0.9045	0.9688	0.9061	0.9611	0.9021	
tennis court	0.5208	0.4277	0.4391	0.8968	0.8953	0.8925	0.8897	0.881	
basketball court	0.6216	0.6429	0.6121	0.8959	0.9091	0.951	0.9025	0.9014	
ground track field	0.8932	0.8827	0.9053	0.9784	0.9825	0.9846	0.9848	0.9825	
harbor	0.8081	0.6964	0.632	0.8317	0.8981	0.8467	0.8443	0.8709	
bridge	0.7499	0.6491	0.5911	0.8724	0.9018	0.9415	0.8952	0.9171	
vehicle	0.4257	0.3408	0.345	0.8903	0.8952	0.8951	0.8912	0.8935	
mAP	0.6763	0.6286	0.6189	0.9022	0.92	0.9187	0.91	0.8985	

TABLE 4. Performance comparison with other five remote sensing image object detectors.

Method	RICNN	R-P-Faster RCNN	CBFF-SSD	MIF-CNN	DFCCNN-VGG	Ours
airplane	0.8835	0.906	0.9693	0.967	0.9058	0.9949
ship	0.7734	0.762	0.9426	0.654	0.9011	0.8796
storage tank	0.8527	0.403	0.8095	0.813	0.8768	0.8951
baseball diamond	0.8812	0.908	0.9909	0.854	0.9882	0.9061
tennis court	0.4083	0.797	0.915	0.683	0.895	0.8925
basketball court	0.5845	0.774	0.9264	0.545	0.9078	0.951
ground track field	0.8673	0.88	0.9882	0.657	0.9062	0.9846
harbor	0.686	0.762	0.9159	0.737	0.8872	0.8467
bridge	0.6151	0.575	0.8968	0.477	0.9034	0.9415
vehicle	0.711	0.666	0.7878	0.631	0.8773	0.8951
mAP	0.7263	0.743	0.9142	0.702	0.9052	0.9187

The results in Table 2 show that our framework has obtained an overwhelming advantage, which once again proves that the proposed framework and strategy are suitable for remote sensing image object detection. We note that we only extract features on a single feature map. We believe that multi-scale features fusion like FPN [11] will inevitably improve the accuracy of object detection, which is also our future research work.

We also submitted the inference results based on the testing dataset to DOTA Evaluation Server ¹ (Our result is named of ‘‘Sui’’ in Task 2) to verify the effectiveness of the proposed framework and achieve a front-rank mAP of 0.7858, which demonstrates that the proposed framework possess outstanding robustness.

E. ABLATION EXPERIMENTS ON THE NWPU VHR-10 DATASET

To verify the effectiveness of the three-stage aggregation structure, we designed experiments about different network structures and IoU threshold strategies on the NWPU VHR-10 dataset. The aggregation structures consist of one stage and three stages with IoU thresholds ranging from 0.3 to 0.9. The results are shown in Table 3. The highest value in each row is shown in bold.

As can be seen from Table 3, the results of three-stage aggregated network are overwhelmingly superior to those of

the one-stage network on the whole. When using one-stage network structure, the detection accuracy decreases as the IoU threshold increasing. This is due to the reduced number of candidate boxes available for refined detection task. The three-stage aggregation structure not only improves the quality of the candidate boxes, but also ensures the quantity unchanged thus enhancing the detection performance.

In this ablation experiments, the highest mAP of 0.92 was obtained when the threshold values of the three stages were 0.4, 0.5 and 0.6, respectively. With the same IoU threshold values of 0.5, 0.6 and 0.7 as in Section IV-D, a comparable mAP of 0.9187 was obtained. Besides, too small or too large IoU threshold values will reduce the accuracy of detection results. A small IoU threshold value will introduce low-quality candidate boxes, while a high IoU threshold value will reduce the number of high-quality candidate boxes. Therefore, an appropriate IoU threshold interval such as [0.4, 0.7], can ensure the effectiveness of the detection results in the three-stage aggregation network. The quantified PRCs of different network structures and IoU threshold strategies are plotted in Figure 6.

We also compared the proposed PTAN framework with other five object detectors including RICNN [20], R-P-Faster RCNN [15], CBFF-SSD [27], MIF-CNN [28] and DFCCNN-VGG [29]. These methods were proposed for remote sensing image object detection. We used the native results of these methods without any additions. The comparison results are shown in Table 4.

¹(<http://captain.whu.edu.cn/DOTAweb/results.html>)



FIGURE 7. The detection results of PTAN on the NWPU VHR-10 dataset.

It can be seen from the comparison of the six methods in Table 4 that the proposed PTAN had an advantage in mAP criterion. Our PTAN also outperformed other object detectors in five categories including airplane, storage tank, basketball court, bridge and vehicle. We note that CBFF-SSD [27] obtained a comparable mAP of 0.9142. However, it used two feature fusion units and seven feature maps, and analyzed the calculation of each layer in the framework. These operations complicated the algorithm. Our PTAN sequentially improved the quality of candidate regions by increasing the IoU threshold stage by stage and obtained the highest mAP value of 0.9187, which demonstrated the effectiveness of the framework. The visualization results on the NWPU VHR-10 dataset are shown in Figure 7.

V. CONCLUSION

In this work, we proposed an effective patch-based three-stage aggregation network for object detection in high resolution remote sensing images. It sequentially improves the quality of the region proposals by increasing the IoU threshold stage by stage to achieve high-precision object detection. To overcome the size restrictions of the input images, we also put forward a patch-based strategy. Experimental results on the DOTA and NWPU VHR-10 datasets attest that the

proposed framework could significantly improve the accuracy for the task of remote sensing image object detection.

ACKNOWLEDGMENT

The authors would like to thank Prof. Gui-Song Xia from State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University for providing the unexceptionable remote sensing image object detection dataset DOTA.

REFERENCES

- [1] Z. Chen, T. Zhang, and C. Ouyang, "End-to-end airplane detection using transfer learning in remote sensing images," *Remote Sens.*, vol. 10, no. 1, pp. 1–15, 2018.
- [2] W. Guo, W. Yang, H. Zhang, and G. Hua, "Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network," *Remote Sens.*, vol. 10, no. 1, p. 131, Jan. 2018. [Online]. Available: <https://www.mdpi.com/2072-4292/10/1/131>
- [3] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0031320317304120>
- [4] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
- [5] X. Zhang, G. Chen, W. Wang, Q. Wang, and F. Dai, "Object-based land-cover supervised classification for very-high-resolution UAV images using stacked denoising autoencoders," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 7, pp. 3373–3385, Jul. 2017.

- [6] M. A. Nielsen, "Neural networks and deep learning," Tech. Rep., 2018. [Online]. Available: <http://neuralnetworksanddeeplearning.com/>
- [7] S. Chen, R. Zhan, and J. Zhang, "Geospatial object detection in remote sensing imagery based on multiscale single-shot detector with activated semantics," *Remote Sens.*, vol. 10, no. 6, p. 820, May 2018. [Online]. Available: <https://www.mdpi.com/2072-4292/10/6/820>
- [8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.
- [9] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
- [10] 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.
- [11] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.
- [12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [13] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf., Amsterdam, The Netherlands*, vol. 9905. Springer, Oct. 2016, pp. 21–37.
- [14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
- [15] 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, no. 7, p. 666, Jun. 2017.
- [16] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.
- [17] X. Zhang, K. Zhu, G. Chen, X. Tan, L. Zhang, F. Dai, P. Liao, and Y. Gong, "Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network," *Remote Sens.*, vol. 11, no. 7, p. 755, Mar. 2019. [Online]. Available: <https://www.mdpi.com/2072-4292/11/7/755>
- [18] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
- [19] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
- [20] 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.
- [21] 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.
- [22] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.
- [23] C. L. Zitnick and P. Dollár, "Edge Boxes: Locating Object Proposals from Edges," in *Proc. ECCV*, 2014, pp. 391–405.
- [24] X. Yang, H. Sun, K. Fu, J. Yang, X. Sun, M. Yan, and Z. Guo, "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, Jan. 2018. [Online]. Available: <https://www.mdpi.com/2072-4292/10/1/132>
- [25] X. Ying, Q. Wang, X. Li, M. Yu, H. Jiang, J. Gao, Z. Liu, and R. Yu, "Multi-attention object detection model in remote sensing images based on multi-scale," *IEEE Access*, vol. 7, pp. 94508–94519, 2019.
- [26] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
- [27] L. Li, S. Zhang, and J. Wu, "Efficient object detection framework and hardware architecture for remote sensing images," *Remote Sens.*, vol. 11, no. 20, p. 2376, 2019. [Online]. Available: <https://www.mdpi.com/2072-4292/11/20/2376>
- [28] W. Zhao, W. Ma, L. Jiao, P. Chen, S. Yang, and B. Hou, "Multi-scale image block-level F-CNN for remote sensing images object detection," *IEEE Access*, vol. 7, pp. 43607–43621, 2019.
- [29] B. Cheng, Z. Li, Q. Wu, B. Li, H. Yang, L. Qing, and B. Qi, "Multi-class objects detection method in remote sensing image based on direct feedback control for convolutional neural network," *IEEE Access*, vol. 7, pp. 144691–144709, 2019.



BING SUI received the M.S. degree in signal and information processing from the University of Science and Technology of China, Hefei, China, in 2008. He is currently with the Hunan Key Laboratory of Meteorological Disaster Prevention and Reduction, Changsha, China. His research interests include remote sensing information processing, the remote sensing monitoring of ecological environment, and the remote sensing monitoring and assessment of meteorological disasters.



MENG XU received the M.S. degree from Nanjing University, Nanjing, China, in 2008. He is currently with the Jiangsu Climate Center, Nanjing, China. His research interests include remote sensing information processing, the remote sensing monitoring of ecological environment, and the remote sensing monitoring and assessment of meteorological disasters.



FENG GAO received the M.S. and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 2006 and 2011, respectively. He is currently with Hunan Glonavin Information Technology Company Ltd. His research interests include remote sensing and photogrammetry.

• • •