applied sciences

MDPI

Letter

Small Object Detection in Optical Remote Sensing Images via Modified Faster R-CNN

Yun Ren [1,*], Changren Zhu [1] and Shunping Xiao [2]

[1] ATR National Lab, National University of Defense Technology, Changsha 410073, China; changrenzhu@nudt.edu.cn

[2] State Key Lab of Complex Electromagnetic Environment Effects on Electronics and Information System, National University of Defense Technology, Changsha 410073, China; shun_ping_xiao@163.com

* Correspondence: renyun_nudt@163.com; Tel.: +86-189-6715-8204

Received: 20 April 2018; Accepted: 16 May 2018; Published: 18 May 2018

Abstract: The PASCAL VOC Challenge performance has been significantly boosted by the prevalently CNN-based pipelines like Faster R-CNN. However, directly applying the Faster R-CNN to the small remote sensing objects usually renders poor performance. To address this issue, this paper investigates on how to modify Faster R-CNN for the task of small object detection in optical remote sensing images. First of all, we not only modify the RPN stage of Faster R-CNN by setting appropriate anchors but also leverage a single high-level feature map of a fine resolution by designing a similar architecture adopting top-down and skip connections. In addition, we incorporate context information to further boost small remote sensing object detection performance while we apply a simple sampling strategy to solve the issue about the imbalanced numbers of images between different classes. At last, we introduce a simple yet effective data augmentation method named 'random rotation' during training. Experimental results show that our modified Faster R-CNN algorithm improves the mean average precision by a large margin on detecting small remote sensing objects.

Keywords: object detection; modified faster R-CNN; remote sensing; feature pyramid

1. Introduction

With the development of remote sensing technology, the research of remote sensing images has been receiving remarkable attention. Meanwhile, ship and plane detection in the optical remote sensing images [1–3] plays an important role in a wild range of applications. Several breakthroughs have been witnessed in the area of large object detection with high resolution on the PASCAL VOC dataset in the past decade by the family of region-based convolutional neural networks (R-CNN) methods [4–7], especially Faster R-CNN [7]. However, they usually fail to detect very small objects, as rich representations are difficult to learn from their poor-quality appearance and structure. However, the object in the optical remote sensing images usually has the characteristic of small object size, which has posed much more challenges than normal object detection and good solutions are still rare so far.

Some efforts have been devoted to addressing small object detection problems. The common way [8,9] is to increase the feature resolution of small objects by simply magnifying the input images, which often results in heavy time consumption for training and testing. Another way [10,11] is centered on generating multi-scale representation which enhances high-level features by combining multiple lower-level features, which is to naively increase the feature dimension. This practice is not able to guarantee that the constructed features are interpretable and discriminative enough for small objects. FCN [12] combines coarse-to-fine predictions from multiple layers by averaging segmentation probabilities. SSD [9] and MS-CNN [13] predict objects at different layers of the feature hierarchy. Another category of approaches, including HyperNet [14], ION [15], PVANET [16], and FPN [17], combine outputs from multiple layers to extract more effective features. In fact, it is a critical role to elaborately design the scale of feature maps to recognize objects across multiple scales.

Empirically, the context information can conduce to improving the object detection performance in natural scenes. R*CNN [18] building on Fast R-CNN [6] uses more than one region for classification while still maintaining the ability to localize the action. MR-CNN [19] develop a multi-region CNN recognition model that yields an enriched object representation capable to capture a diversity of discriminative appearance factors. Mottaghi et al. [20] designed a novel deformable part-based model that exploits both global and local context around each candidate detection, which can help in detecting objects at all scales especially at tiny objects.

As we know, the PASCAL VOC dataset which contains 20 object categories is the most widely used benchmark dataset for generic object detection. The object instances in the PASCAL VOC are usually large because they occupy a major portion of the image. However, the concerned remote sensing object instances such as plane and ship usually have smaller object size in which the difficulty with small objects is intuitive. Current object detectors, like Faster R-CNN, always leverage the convolutional neural networks to extract increasingly abstract feature representations. During this process, the intermediate feature maps are usually down-sampled too many times by the convolutional layer or the pooling layer whose stride is greater than one. Obviously, it is expected that directly applying the Faster R-CNN to detecting the small remote sensing objects only obtains poor performance. To address this issue, we investigate how to modify Faster R-CNN for the task of small object detection in optical remote sensing images.

In this paper, we extend the prevailing Faster R-CNN for the small object detection in optical remote sensing images. Firstly, we elaborately modify anchors in the RPN stage of Faster R-CNN based on the statistics of our training set to generate the small object proposals. Secondly, an effective method is raised to produce higher-resolution feature maps, simultaneously utilizing low-level features and high level features, which is very critical to enable us to detect small remote sensing objects. Thirdly, we leverage the context information enclosing an object proposal during the training process to further boost the small object detection performance. Finally, we present a simple yet effective approach, called 'random rotation', to augment our available optical remote sensing data while applying a sampling strategy to solve the problem of non-uniform class distribution during training.

2. Modifying Faster R-CNN for Small Object Detection in Optical Remote Sensing Images

In the feature extraction process of Faster R-CNN, a region proposal network (RPN) shares convolutional layers with region-based detectors, especially Fast R-CNN, which can significantly reduce the proposal cost in comparison with the popular Selective Search. RPN is designed to efficiently predict region proposals with a wide range of scales and aspect ratios for the rather large object in PASCAL VOC. What is noticed is that the smallest RPN anchor boxes are much bigger than the most instances of our remote sensing object dataset.

It can be found from Figure 1 that the areas of most bounding boxes are between 10^2 and 100^2 pixels in our dataset. In the original paper, their anchor-based method is built on a pyramid of anchors with multiple scales and aspect ratio. The three aspect ratios used are 0.5, 1, 2, and the areas of the square shape bounding boxes at the three scales are 128^2, 256^2, and 512^2 pixels, respectively. However, the default setting of the anchor parameters, which is not able to cover the range of small object size in optical remote sensing images, enable to deliver good results on datasets such as PASCAL VOC where the objects are typically relatively large. In order to address this problem, several modifications are described as below.

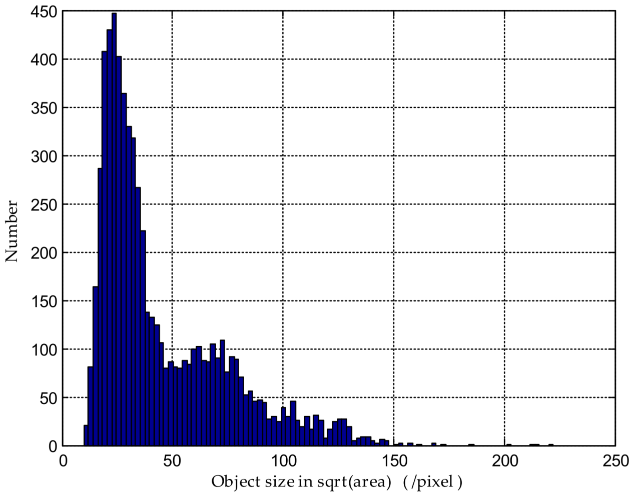


Figure 1. Histogram of object sizes in our remote sensing dataset.

It is noteworthy that the original RPN anchors are too large to cover the range of object sizes in our remote sensing dataset, which can be shown in the Figure 1. Based on this observation, we could either choose adequate anchors by experience or simply add additional anchors using the same powers-of-two scheme while keeping the default aspect ratios used in the original paper. Furthermore, we notice that the feature map size of the last shared convolutional layer is so small that the extracted features are only sensitive to the large objects. Because the intermediate feature maps are usually down-sampled four times by the stride of two pixels. Ren et al. [21] has experimentally proved that choosing the pre-trained ResNet-50 [22] model enable to obtain better performance than other pre-trained models like VGG [23] and Inception [24–26]. Following this, we choose the ResNet-50 model as the backbone in the Faster R-CNN by default. However, the intermediate feature maps are naturally down-sampled five times using the convolutional layer with the stride of 2.

Built on the successful FPN, the top-down pathway is adapted to generate higher resolution but semantically stronger feature maps for the shared feature extractor as shown in the Figure 2. These feature maps are then concatenated channel by channel with the feature maps from the bottom-up pathway via lateral connections. Furthermore, all the channel dimensions are beforehand adjusted to a fixed number by a 1×1 convolutional layer. Thus each lateral connection combines feature maps of the same spatial size from the bottom-up pathway and the top-down pathway. There are often many layers producing output maps of the same size and we say these layers are in the same network stage. Specifically, we use the feature maps output by the last residual block of second three stages which are denoted as res3d, res4f, and res5c in the Figure 2 respectively. Meanwhile, we notice that they have strides of $\{8, 16, 32\}$ pixels and the final feature output has the stride of 8 pixels. We simply up-sample the feature maps of the last layer of higher stages by the bilinear interpolation method. After merging the three feature outputs by element-wise addition, a 3×3 convolutional layer is appended to generate the final feature map, which is used to degrade the aliasing effect of up-sampling.

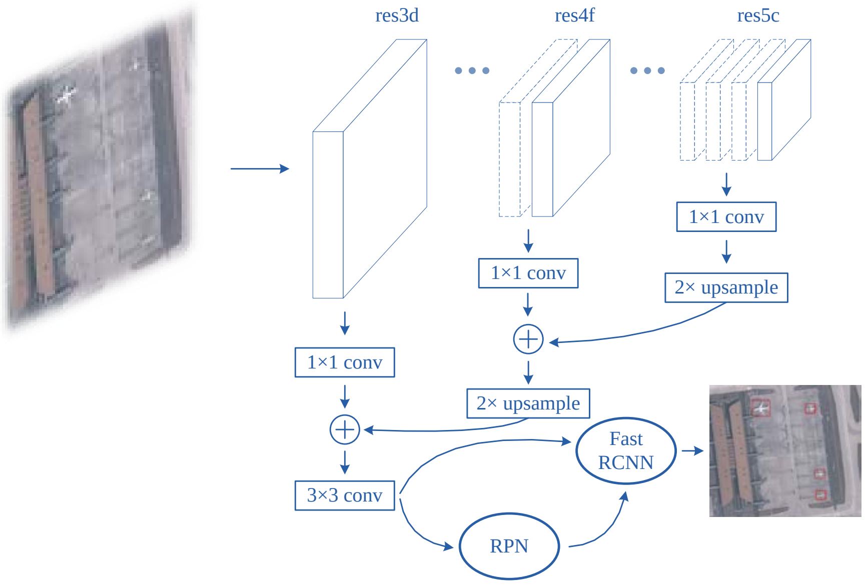


Figure 2. Illustration of the proposed modified Faster R-CNN for small remote sensing objects.

3. Contextual Detection Model

In this section, we design a new contextual model to better exploit contextual information for our remote sensing dataset. After generating the object proposals in the RPN stage, an ROI-Pooling layer is used to project each proposal onto the shared feature maps based on the strides of the network in the Fast R-CNN stage. The feature maps corresponding to each proposal are then encoded into a fixed-dimensional representation with a predefined spatial resolution. Following this, several fully connected layers are fed with these presentations for classification and class-specific bounding box regression.

Since context contributes to the object detection, we expect that it will help to effectively detect small remote sensing objects. Moreover, the feature maps corresponding to the small candidate proposals, whose spatial resolution may be less than 1×1 after enduring multiple down-sampling process, are less discriminable because small objects usually only occupy a small image area. Therefore, we focus on leveraging the context information to boost the performance of small remote sensing object detection.

As is demonstrated in the Figure 3, we incorporate the corresponding context region enclosing the proposal region with the candidate proposal after the ROI-Pooling layer in which the spatial resolution follows the default setting, namely 7×7. This is a simple and intuitive manner to merge the context information, which is proved to be effective in our experiments. Furthermore, the concatenated feature maps are fed into a 1×1 convolutional layer to reduce channel dimensions considering for the computation cost. Besides, we construct the shared feature maps with all the convolutional layers from the pre-trained ResNet-50 model which has no fully connected layers. So we attach two hidden 1024-d fully connected layers (each followed by Dropout [27] and ReLU layer), which are randomly initialized by the Xavier method, before the final classification and bounding box regression.

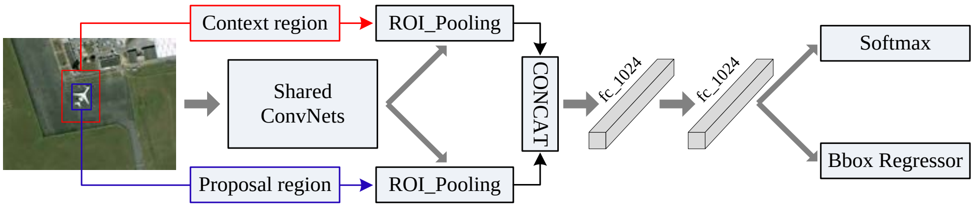


Figure 3. The contextual detection model.

4. Data Pre-Processing

In this section, we present a simple yet effective approach to augment our available optical remote sensing data while applying a sampling strategy to solve the problem of non-uniform class distribution during training.

As a matter of fact, the available optical remote sensing data are very scarce while they can be gathered difficultly. We only focus on two remote sensing object instances, ship and plane. Currently only the NWPU VHR-10 dataset [28], which contains totally 800 very-high-resolution (VHR) optical remote sensing images that were cropped from Google Earth and Vaihingen dataset and then manually annotated by experts, is available to the public. This dataset includes 10 categories—which are airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle—but there are only 90 images involved in the plane class and 57 images for the ship class, which is not enough for training. Getting more data is essential for our current situation.

We augment our training data by two methods: manual annotation and data augmentation. A set of 2608 images containing the 'ship' category in our dataset are collected from multiple sensors and platforms such as Google Earth with multiple resolutions. In addition, the flip operation is usually used for data augmentation in object detection. Built on Random Erasing (RE) [29], we introduce Random Rotation (RR) which is a simple yet effective data augmentation technique for training our modified Faster R-CNN. In particular, RR happens in a certain probability. An image within a mini-batch is randomly chosen to undergo either RR with probability p or kept unchanged with probability $1 - p$. RR randomly rotates an image by an angle θ. Notably, the four points of the ground truth, annotated by a rectangle region, are rotated by the angle θ. The original coordinates of these points are denoted anticlockwise as $\{(x_i, y_i), i = 1,2,3,4\}$, respectively. Hence, the rotated points $\{(x_i\prime, y_i\prime), i = 1,2,3,4\}$ can be calculated under the following Equation (1).

$$\begin{bmatrix} x_i\prime \\ y_i\prime \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad i = 1,2,3,4 \tag{1}$$

What is noteworthy is that the original bounding box of each ground truth becomes parallelogram after RR operation so that we employ their minimum enclosing rectangle (MER), whose upper left corner (x_{min}, y_{min}) and lower right corner (x_{max}, y_{max}) can be calculated by the following Equation (2), as the rotated bounding box. This has the advantage of leveraging the context information to some extent because the MER crops more regions than the original bounding box.

$$\begin{cases} x_{min} = \min(x_1\prime, x_2\prime, x_3\prime, x_4\prime) \\ y_{min} = \min(y_1\prime, y_2\prime, y_3\prime, y_4\prime) \\ x_{max} = \max(x_1\prime, x_2\prime, x_3\prime, x_4\prime) \\ y_{max} = \max(y_1\prime, y_2\prime, y_3\prime, y_4\prime) \end{cases} \tag{2}$$

Fortunately, in total we collected 5922 optical remote sensing images, 5216 for 'ship' and 706 for 'plane'. Obviously, the numbers of images in different classes are imbalanced, which pose great challenges for training. To address this, we apply a sampling strategy named 'balanced sampling' (BS) [30] during training. This strategy is aim at iterating as uniform as possible within an epoch with respect to classes. In reality, we use one type of list, namely a training list. A typical example for three classes is shown in the Figure 4. Firstly, we sort our training list class by class and count the largest category number denoted as K_3 in the Figure 4. Then we generate a random list of K_3 integers with the interval $[0, K_3 - 1]$ for each class. For each class, we leverage the mod operator to obtain a corresponding indexed-value list which the images are sampled according to. Finally, a new training list is generated by concatenating and shuffling the sampled image list. After an epoch, the foregoing operators are repeated again until the end of the whole model training. Apparently, our sampling strategy is not only cost effective but very easy to implement.

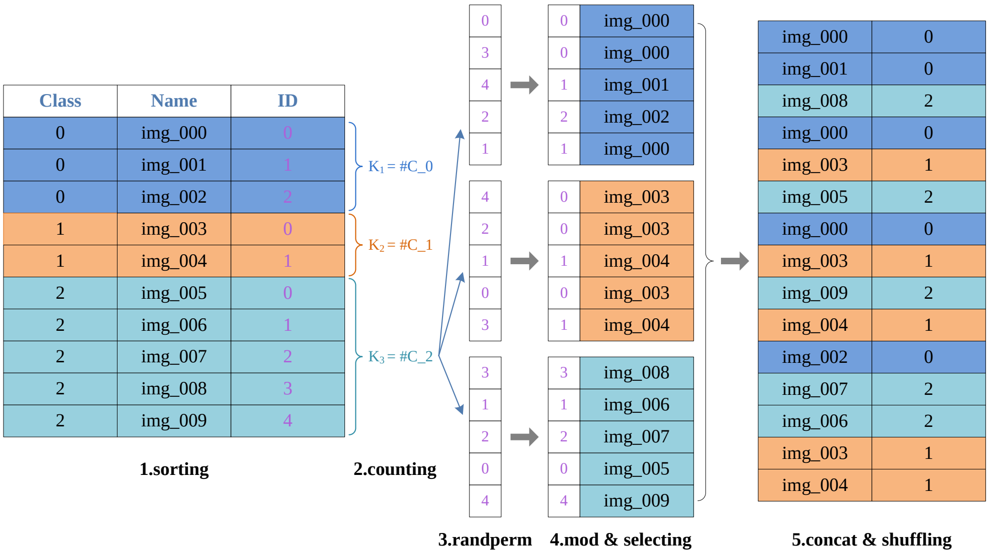


Figure 4. The pipeline of balanced sampling.

5. Experiments and Results

5.1. Implementation Details

Our experiments are conducted based on the modified Faster R-CNN detector which employs the ResNet-50 model if not specified. The model is initialized by the ImageNet classification model and then fine-turned on our optical remote sensing dataset which contains two-class remote sensing object instances, ship and plane. Some samples from our dataset, which have been resized to 128 × 128 pixels for viewing conveniently, are shown in Figure 5. It can be seen from Figure 5 that the images contain scenes of civilian ports, military bases, offshore areas, and far seas. In total, we collected 5922 optical remote sensing images named 'SORSI dataset', 5216 for ship and 706 for plane. It is noteworthy that the numbers of images in different classes are highly imbalanced, which poses great challenges for model training. It can be found from Figure 6 that the areas of most bounding boxes are between 10^2 and 100^2 pixels in our dataset. Besides, the areas of bounding boxes falling in the ship category dominate from 10^2 to 50^2 pixels while those in the plane category possess from 50^2 to 100^2 pixels. Obviously, it is far more difficult to detect ships than to detect planes. To make an evaluation, our dataset is randomly split into 80% for training and 20% for testing. In the training process, we flip all the training images while subtracting the mean value (103.939, 116.779, 123.68).

In all of the experiments, we trained and tested both RPN and Fast R-CNN on images of a single scale based on the deep learning framework, Caffe [31]. We resize the images such that their shorter side is 608 pixels under the premise of ensuring the longer side less than 1024 pixels. Meanwhile, we apply stochastic gradient descent (SGD) for 20K iterations to train the baseline model and the training rate starts with 0.01 and decreases to 0.0001 after 15K iterations. For anchors, we adopt three scales with box areas of 16^2, 40^2, and 100^2 pixels, and three aspect ratios of 1:1, 1:2, and 2:1, which are adjusted for better coverage of the size distribution of our optical remote sensing dataset.

The evaluation metric is average precision (AP) of each object instance and mean average precision (mAP) with Interception-of-Union (IoU) threshold as 0.5. To reduce redundancy, non-maximum suppression (NMS) is adopted on the proposal regions based on their box-classification scores. The IoU threshold is fixed for NMS at 0.7. All experiments were performed on Intel i7-6700K CPU and NVIDIA GTX1080 GPU.

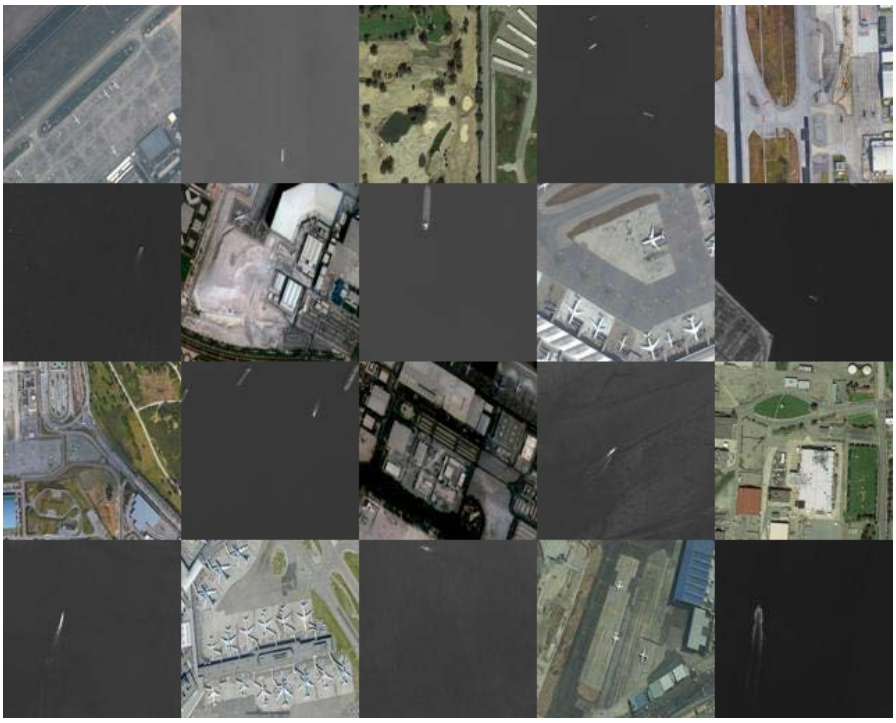


Figure 5. Some samples from SORSI dataset.

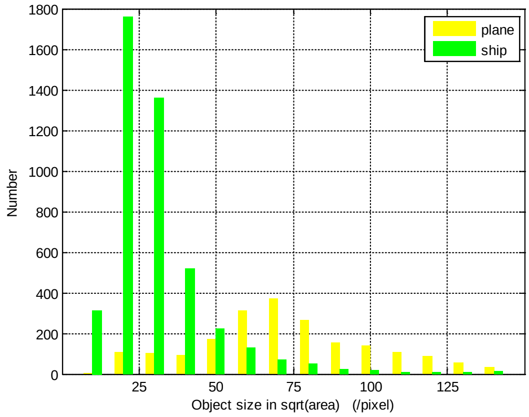


Figure 6. Histogram of object sizes class by class.

5.2. Comparative Experiment

As a baseline to validate the effectiveness of our method, we perform an experiment on our dataset, leaving the settings of the modified Faster R-CNN on the default parameters in which the pre-trained model is ResNet-50 networks except that the total stride on the last shared convolutional layer is 16 pixels in contrast. In this case, we only use the feature maps output by the last residual block of second two stages which are denoted as res4f and res5c in the Figure 2 respectively. Similarly, we up-sample the feature maps of the last layer of higher stages by a factor of 2 using the bilinear interpolation method for simplicity. Before that, we adopt a 1 × 1 convolutional layer to reduce channel dimensions. After merging the two feature outputs, a 3 × 3 convolutional layer is appended to generate the final feature map. Furthermore, a few experiments are separately performed to evaluate the impact of using appropriate anchor boxes by setting different anchor scales.

We report the results with using various strategies during training the modified Faster R-CNN on SORSI dataset in Table 1. The performance of the baseline is 66.6% mAP where the AP is 58.2% for ship and 75.0% for plane. Apparently, we find that the modified Faster R-CNN with the total stride of 8 achieves better performance than the baseline model with the total stride of 16 by a large margin, especially for the ship whose areas of bounding boxes are between 10^2 and 50^2 pixels. Besides, it is evidently shown that using appropriate anchor boxes can conduce to boosting the detection performance by almost two percentage points. This may bring in some insights about how to choose appropriate anchor boxes according to the existing dataset. Based on this conclusion, all of the follow-up experiments will adopt three scales with box areas of 16^2, 40^2, and 100^2 pixels.

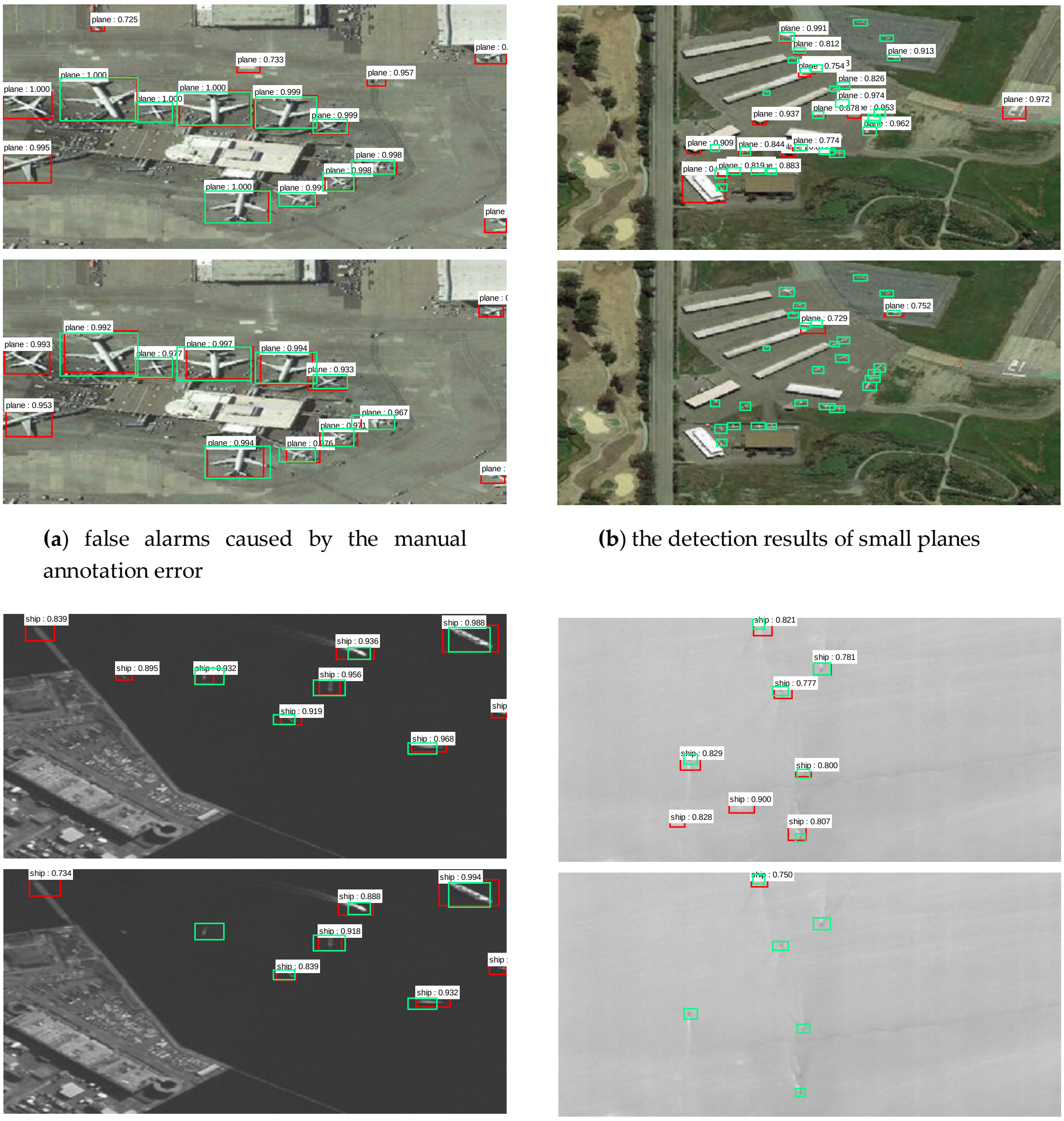
Table 1. The results of modified Faster R-CNN on SORSI dataset.

Method	**mAP (%)**	**AP (%)**		**Anchor Scale**	**Context**	**RR**	**BS**
		Ship	**Plane**				
Baseline (stride = 16)	66.6	58.2	75.0	$\{128^2\ 256^2\ 512^2\}$			
	67.1	59.5	74.6	$\{64^2\ 128^2\ 256^2\}$			
	68.1	60.1	76.0	$\{10^2\ 40^2\ 100^2\}$			
Modified Faster R-CNN (stride = 8)	73.5	71.0	76.0	$\{10^2\ 40^2\ 100^2\}$			
	74.1	71.7	76.5	$\{10^2\ 40^2\ 100^2\}$	√		
	75.8	69.7	81.8	$\{10^2\ 40^2\ 100^2\}$		√	
	76.7	69.8	83.6	$\{10^2\ 40^2\ 100^2\}$			√
	76.1	71.4	80.8	$\{10^2\ 40^2\ 100^2\}$	√	√	
	77.1	70.4	83.9	$\{10^2\ 40^2\ 100^2\}$		√	√
	78.3	72.3	84.3	$\{10^2\ 40^2\ 100^2\}$	√		√
	78.9	72.9	85.0	$\{10^2\ 40^2\ 100^2\}$	√	√	√

Due to the non-uniform class distribution, the plane class cannot be trained enough because the batch size has to be fixed to 1 for RPN stage. The relevant experimental results indicate that the balanced sampling strategy contributes to increasing the plane AP by 7.6% while the ship AP only decreases by 1.2%, which proves that the BS strategy is able to solve the effect of the non-uniform class distribution in some degree. Furthermore, we conduct several experiments to investigate the behavior of the proposed RR as a data augmentation technique. In the experiments, the probability p is set at 0.5 while the rotate angle θ is set at 10°. We observe that the behavior of the proposed RR is somewhat similar to that of BS which enlarges our training set equivalently. When using the contextual detection model during training, it is noted that we only obtain low-level improvements by 0.7% and 0.5%, respectively. Besides, when adopting any two kinds of the three strategies during training, we find that it can contribute to achieve better performance combining the balanced sampling strategy and the contextual detection model. At last, the best performance can be obtained by combining the three aforementioned strategies.

By comparing and analyzing the multiple groups of experiments, the validity of the proposed structure is verified. Our modified Faster R-CNN delivers very impressive performance on detecting small objects in optical remote sensing images. However, it can be seen from Table 1 that the mAP still has room for improvement. Through observation of the test results, we attribute this to two points: false alarm and misjudgment.

Some test results are shown in Figure 7 on the test set of our SORSI dataset. As shown in Figure 7b–d, more small objects are able to be detected in the case that the stride is equal to 8, which suggests that producing higher-resolution feature maps simultaneously utilizing low-level features and high level features is very critical to enable us to detect small remote sensing objects. However, ships tend to dock in a complex scene such as a port while planes always line up on the airfield. These scenes often contain objects with similar geometric construction, such as a long straight line. These disturbances will cause false alarms on the detector, as shown in Figure 7. Furthermore, some objects are too small to be detected, resulting in misjudgments as illustrated in Figure 7b. At the same time, a few ground truths may be not annotated due to the manual annotation error as indicated in Figure 7a, which leads to false alarms as well.



(**a**) false alarms caused by the manual annotation error

(**b**) the detection results of small planes

(**c**) false alarms caused by the similar geometric construction

(**d**) the detection results of small ships

Figure 7. Some test results on the test set of our SORSI dataset. The green boxes indicate the ground truths while the red boxes refer to the detected objects annotated by their scores. There are two detection results in every single image where the upper is the detection result in the case that the stride is equal to 8 and the bottom is the detection result in the case that the stride is equal to 16.

6. Conclusions

In this paper, we proposed a modified Faster R-CNN method to deal with the small object detection problem in optical remote sensing images. To address this, we designed a similar architecture adopting top-down and skip connections to produce a single high-level feature map of a fine resolution as the final shared feature output, which is very critical to enable us to detect small remote sensing objects. At the same time, we chose appropriate anchors to cover the size distribution of our optical remote sensing dataset. Furthermore, we leveraged the context information enclosing an object proposal to further improve the small object detection performance during training. We presented a simple yet effective approach, named 'random rotation', to augment our available optical remote sensing data while applying a sampling strategy to solve the problem of non-uniform class distribution during training. We conducted a wide range of experiments and provided a comprehensive analysis of the performance of our modified Faster R-CNN on the task of small object detection in optical remote sensing images. Our future work will focus on applying our approach to other remote sensing objects in complex scenes and detecting dense small optical remote sensing objects.

Author Contributions: Y.R. provided the original idea for the study; C.Z. and S.X. contributed to the discussion of the design; Y.R. conceived and designed the experiments; C.Z. supervised the research and contributed to the article's organization; and Y.R. drafted the manuscript, which was revised by all authors. All authors read and approved the final manuscript.

Acknowledgments: The authors would like to thank all the colleagues who generously provided their image dataset with the ground truth. The authors would also like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dong, C.; Liu, J.; Xu, F. Ship Detection in Optical Remote Sensing Images Based on Saliency and a Rotation-Invariant Descriptor. *Remote Sens.* **2018**, *10*, 400. [CrossRef]
2. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic Ship Detection in Remote Sensing Images from Google Earth of Complex Scenes Based on Multiscale Rotation Dense Feature Pyramid Networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
3. Xu, F.; Liu, J.; Sun, M.; Zeng, D.; Wang, X. A Hierarchical Maritime Target Detection Method for Optical Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 280. [CrossRef]
4. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
5. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014; pp. 346–361.
6. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
8. Chen, X.; Kundu, K.; Zhu, Y.; Berneshawi, A.G.; Ma, H.; Fidler, S.; Urtasun, R. 3D Object Proposals for Accurate Object Class Detection. *Lect. Notes Bus. Inf. Process.* **2015**, *122*, 34–45.
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
10. Li, H.; Lin, Z.; Shen, X.; Brandt, J.; Hua, G. A convolutional neural network cascade for face detection. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5325–5334.
11. Yang, F.; Choi, W.; Lin, Y. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2129–2137.
12. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef] [PubMed]

13. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 354–370.
14. Kong, T.; Yao, A.; Chen, Y.; Sun, F. HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 845–853.
15. Bell, S.; Lawrence Zitnick, C.; Bala, K.; Girshick, R. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2874–2883.
16. Hong, S.; Roh, B.; Kim, K.H.; Cheon, Y.; Park, M. PVANet: Lightweight Deep Neural Networks for Real-time Object Detection. *arXiv*, 2016.
17. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
18. Gkioxari, G.; Girshick, R.; Malik, J. Contextual Action Recognition with R*CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1080–1088.
19. Gidaris, S.; Komodakis, N. Object Detection via a Multi-region and Semantic Segmentation-Aware CNN Model. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1134–1142.
20. Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.G.; Lee, S.W.; Fidler, S.; Urtasun, R.; Yuille, A. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 891–898.
21. Ren, Y.; Zhu, C.; Xiao, S. Object Detection Based on Fast/Faster RCNN Employing Fully Convolutional Architectures. *Math. Prob. Eng.* **2018**, *2018*. [CrossRef]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
23. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, 2014.
24. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
25. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
26. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv*, 2016.
27. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
28. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
29. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. *arXiv*, 2017.
30. Shen, L.; Lin, Z.; Huang, Q. Relay Backpropagation for Effective Learning of Deep Convolutional Neural Networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 467–482.
31. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv*, 2014.