remote sensing

MDPI

Article

Object Detection in Remote Sensing Images via Multi-Feature Pyramid Network with Receptive Field Block

Zhichao Yuan [1,2], Ziming Liu [1,2,3], Chunbo Zhu [1,2], Jing Qi [4] and Danpei Zhao [1,2,*]

[1] Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China; zhichaoyuan@buaa.edu.cn (Z.Y.); ziming.liu@connect.polyu.hk (Z.L.); 15151122@buaa.edu.cn (C.Z.)
[2] Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China
[3] Department of Computing, The Hong Kong Polytechnic University, Hongkong 999077, China
[4] DFH Satellite Co., Ltd., Beijing 100094, China; yuanbochn@buaa.edu.cn
* Correspondence: zhaodanpei@buaa.edu.cn

Abstract: Object detection in optical remote sensing images (ORSIs) remains a difficult task because ORSIs always have some specific characteristics such as scale-differences between classes, numerous instances in one image and complex background texture. To address these problems, we propose a new Multi-Feature Pyramid Network (MFPNet) with Receptive Field Block (RFB) that integrates both local and global features to detect scattered objects and targets with scale-differences in ORSIs. We build a Multi-Feature Pyramid Module (M-FPM) with two cascaded convolution pyramids as the main structure of MFPNet, which handles object detection of different scales very well. RFB is designed to construct local context information, which makes the network more suitable for the objects detection around complex background. Asymmetric convolution kernel is introduced to RFB to improve the ability of feature attraction by adding nonlinear transformation. Then, a two-step detection network is constructed to combine the M-FPM and RFB to obtain more accurate results. Through a comprehensive evaluation of the experimental results on two publicly available remote sensing datasets Levir and DIOR, we demonstrate that our method outperforms state-of-the-art networks for about 1.3% mAP in Levir dataset and 4.1% mAP in DIOR dataset. Experimental results prove the effectiveness of our method in ORSIs of complex environments.

Keywords: object detection; convolutional neural networks (CNNs); remote sensing image (ORSIs); receptive field; multi-feature pyramid

check for updates

Citation: Yuan, Z.; Liu, Z.; Zhu, C.; Qi, J.; Zhao, D. Object Detection in Remote Sensing Images via Multi-Feature Pyramid Network with Receptive Field Block. *Remote Sens.* **2021**, *13*, 862. https://doi.org/10.3390/rs13050862

Academic Editor: Jon Atli Benediktsson

Received: 23 January 2021
Accepted: 18 February 2021
Published: 25 February 2021

1. Introduction

As a challenging problem in the field of aerial image analysis, object detection in optical remote sensing images (ORSIs) has attracted much more attention in recent years. Meanwhile, optical remote sensing object detection is also an important part of the object detection field because it has strong practical value. Nowadays, object detection models have made great progress due to the development of convolutional neural networks (CNNs). A large number of one-stage [1–4] and two-stage networks [5–10] have been put forward. Large-scale natural image datasets for object detection tasks such as Pascal VOC [11,12] and MS COCO [13] have also emerged. One-stage networks have high computational efficiency, but the accuracy is often inferior to two-stage networks. Besides, some anchor-free methods [14–16] also emerged in recent years. These approaches abandon anchors, and are more intelligent when facing detection and classification problem. However, they are not much efficient as one-stage networks. These methods are then applied to various fields, such as face recognition [17,18], semantic segmentation [19–21] and autonomous driving [22,23].

ORSIs possess abundant information and have a wide application prospect in both military and civilian fields. However, object detection in ORSIs is a challenging task. As shown in Figure 1, there are problems such as massive co-existing instances in one image, different scales between classes, and complex background textures. Due to these characteristics, generic object detection approaches cannot perform well on ORSIs. However, trapped by the absence of public datasets, there are not much detection methods for ORSIs. In the past two years, object detection for ORSIs has got a chance to return to the public with the appearance of large remote sensing datasets such as Levir [24], DOTA [25] and DIOR [26]. Some networks such as [27–29] have emerged. These methods are designed for the rotation cases of remote sensing objects, and they combined these with local context features. However, these networks do not perform well in images that contain massive instances. Some other methods [30–32] improve the speed of detection, but they are not well optimized for small instances in ORSIs.

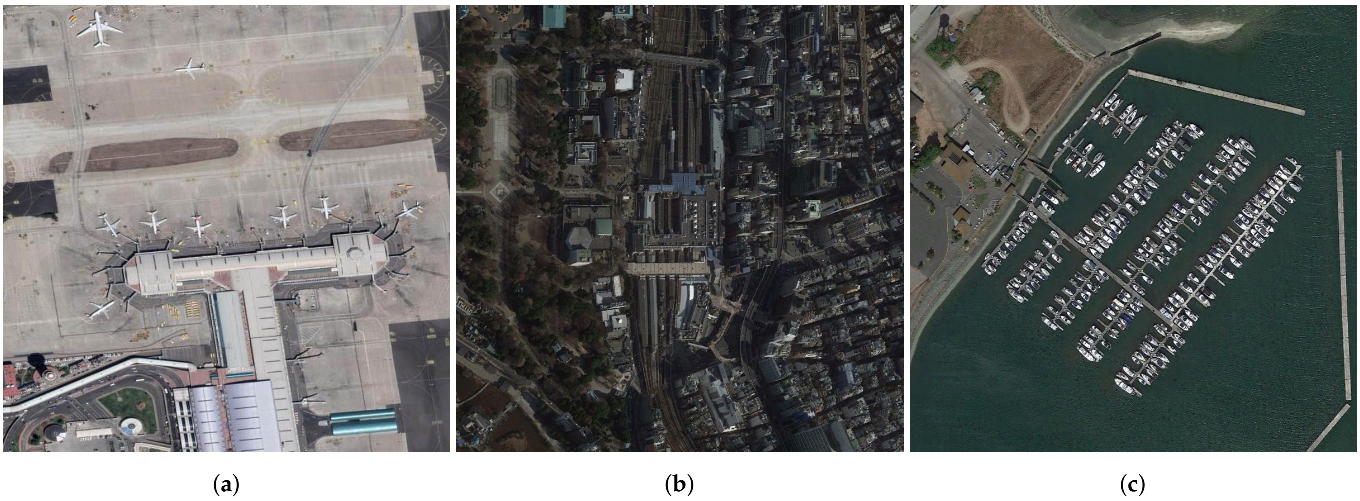

Figure 1. Three main problems in detection of ORSIs. (**a**) refers to the situation of scattered objects in one image. (**b**) has a railway station in the middle, however it is overwhelmed by the complicated background. (**c**) is a scene with two categories of objects with different scale.

To solve the problems above, we propose a Multi-Feature Pyramid Network (MFPNet) with the Receptive Field Block (RFB). Different from the classical pyramid-based networks, we design a cascaded pyramid structure called Multi-Feature Pyramid Module (M-FPM), which can better balance the local and global feature extraction. The structure firstly integrates features with different scales and make them compatible with local details and global semantic information. Then the mixed features flow into the second stage of the cascaded pyramid and generate balanced multi-scale features. RFB are performed to extract local context features, strengthen the mining of local information, and enhance the ability to detect instances with complex texture. Our method reaches state-of-the-art performance on the Levir and DIOR benchmark. The main contributions of this work are summarized as follows:

- We design a Multi-Feature Pyramid Network which obtains multi-scale features from a cascaded pyramid structure. M-FPM is proposed to integrates both local and global features, increasing the robustness of detecting scattered objects with scale differences.
- RFB is constructed to keep detailed information while increasing the receptive field. By generating high resolution feature maps with local context information, the accuracy of detection can be raised.
- We have improved the network's ability for multi-scale objects detection. The model is also more responsive to ORSIs with numerous and scattered instances or complex backgrounds. The network has achieved state-of-the-art results in the largest remote sensing dataset currently.

2. Related Works

In this section, we will briefly introduce object detection networks in ORSIs, methods based on feature pyramids, and approaches using receptive fields. These studies have greatly contributed to our method.

2.1. Object Detection Networks in ORSIs

CNN-based object detection networks are mainly divided into one-stage and two-stage ones. Two-stage networks have a step of extracting Regions of Interest (ROIs), so the detection accuracy of the network is improved, but the speed is also reduced. Faster R-CNN [10] proposes a Region Proposal Network (RPN), which uses anchors to generate candidate bounding boxes on the generated feature map. Cascade R-CNN [33] trains multiple cascaded networks with various Intersection over Union (IoU) thresholds and combines these networks together. PANet [34] makes full use of feature fusion to make network detection results more accurate. One-stage networks omit the step of generating ROIs, which can effectively speed up detection with the sacrifice on accuracy. RetinaNet [1] introduces Focal Loss, which allows network to assign attention on hard negative samples. CornerNet [35] is a keypoint-based method which uses corners of the bounding box to detect objects. CenterNet [16] handles the center point of the bounding boxes for detection.

CNN-based object detection methods have achieved great improvement in natural image target location task, which provides a practical option for ORSIs object detection. Many optical remote sensing object detection methods today use R-CNN based networks [27,29]. Ref. [29] seeks a breakthrough on the region proposal network (RPN). In order to solve the problem of appearance ambiguity, Ref. [29] adds additional multiangle anchors. Meanwhile, a hybrid restricted Boltzmann machine is introduced to provide local context features. Ref. [27] imposes a rotation-invariant regularization function and a Fisher discriminant regularization function to remove the similarity between classes. However, the quality of region proposals affects the object detection performance. In order to achieve real-time object detection, regression-based networks are also applied to ORSIs. Ref. [36] uses YOLO-v2 [37] structure as a basic network to detect rotated objects. In response to the lack of datasets, Refs. [38,39] focused on the transfer learning from natural images to ORSIs to improve the detection accuracy. Ref. [39] uses multi-feature fusion. It combines two independent CNNs to detect small objects from coarse to fine and achieves good results. Except for this, some methods also use hard example mining [32,40] to make the model more robust.

2.2. Object Detection Networks Based on Feature Pyramids

The feature pyramid is widely used for detecting objects with various scales. SSD [2] obtain multi-scale features by extracting anchors in different stages. It achieves high detection speed, but its performance for small objects is generally not as good as two-stage networks. DSSD [41] uses deconvolution layers to extract both low-level and high-level features, making contextual information much richer. However, the deconvolution module slows down the speed of detection, and the deeper backbone also decreases the inference speed. RefineDet [4] combines the advantages of one-stage and two-stage networks by performing a correction step towards anchors to improve detection accuracy. M2Det [42] uses Multi-Level Feature Pyramid Network (MLFPN) to fuse multiple features to detect objects with different scales. However, the network structure is too complicated which lead to a huge amount of parameters. Though there are still problems to be solved, the above methods have verified that the feature pyramid has played an important role in facing multi-scale and multi-class objects.

2.3. Object Detection Networks Using Receptive Fields

In order to reduce the depth of the network while improving the speed and accuracy of detection, various works have been done to investigate the receptive field. According to these research, the application of the receptive field enhances the feature representation of the area. The early application of the receptive field came from the Inception [43]. After that, both Deformable CNN [44] and ASPP [45] have made their own understanding of receptive fields. Inception [43] combines different sizes of convolution kernels to obtain multi-scale information. However, the center point of each convolution kernel is in the same position, which does not have a good effect on the extraction of edge information. Meanwhile, Inception also needs to ensure that the kernel is large enough to get adequate information, which increases the amount of parameters. ASPP [45] uses dilated convolution to extract the characteristic information of the receptive fields in local areas. However, there are some gaps between the convolution kernels of ASPP, and the collection of context information is not very well. Deformable CNN [44] uses deformable convolution, which improves the ability of CNNs for spatial modeling. However, the model can only operate on the target area, and lacks the processing of context information. In ORSIs object detection, each object and its surrounding background information are closely related, so this method is not applicable. According to these shortcomings, RFBNet [46] improves the problem of lacking receptive field information around the object. However, this model does not integrate global and local features, so its feature extraction has limitations. All in all, the receptive field makes it possible for fast object detection network.

3. Multi-Feature Pyramid Network

Figure 2 shows the structure of the Multi-Feature Pyramid Network (MFPNet). The MFPNet is proposed to solve the problems of scattered distributed objects, complex backgrounds, and scale difference between classes. These problems may also exist in general object detection, but they are more significant in ORSIs. The size of the same object at different resolutions may vary greatly, so most approaches are only effective on a specific dataset. Our target is to promote the robustness of the method. As concluded in Section 2, feature pyramids and receptive fields can make significant promotion for optical remote sensing object detection. Thus, multi-scale features are adopted in our method. In order to ensure that the network can cover semantic information from shallow layers to deep ones, a cascaded feature pyramid is constructed in the VGG16 backbone and a cascade method is used to merge the features of different layers to highlight the features of small objects. Then, the RFB is used to extract the local context information from the cascaded feature layer. Dilated convolutions are used to ensure the rotation-invariance. Finally, the second convolution feature pyramid is built to propose multi-level anchors. Two detection modules are used to improve the regression and classification of the final bounding box. VGG16 is adopted as the backbone to ensure the real-time performance. In the following parts, we will introduce the network from three aspects: (1) Multi-Feature Pyramid Module; (2) Receptive Field Block; (3) Double-Check Detection Network Module.

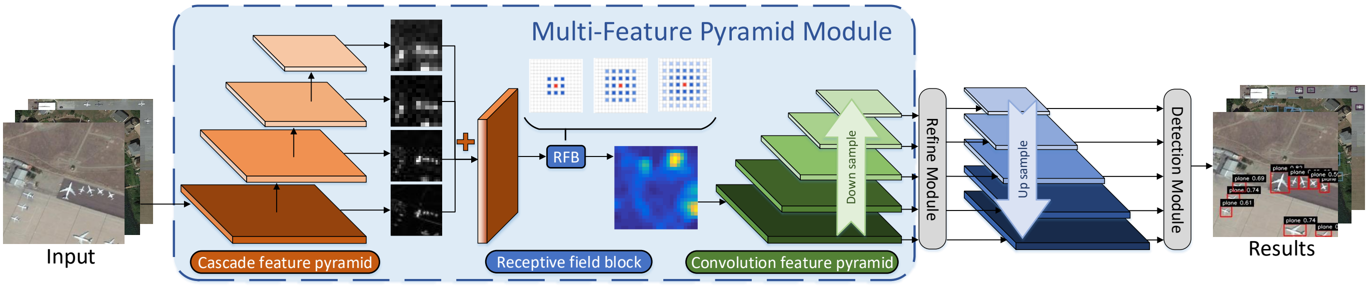

Figure 2. Architecture of MFPNet.

3.1. *Multi-Feature Pyramid Module*

Since the scale of objects in the ORSIs varies greatly, it is necessary to combine features of different levels for detection. To solve the stated problems of ORSIs, the multi-scale method is adopted, which lead to the concept of the Multi-Feature Pyramid Module (M-FPM). The main architecture of the M-FPM consists of two parts. A cascaded feature pyramid and a convolution feature pyramid. In the first cascaded feature pyramid, Convolution Layer (Conv)3_3, Conv4_3, Conv5_3 and Conv_fc7 in VGG16 are extracted as four source feature layers. The extraction of the feature layers refers to YOLO-v3 [47]. Though the max pooling strategy reduces the number of parameters and enhances the global feature, the local features are also suppressed. The cascade strategy can make up for this loss. Most objects in ORSIs are small and scattered, so local features need to be preserved as much as possible. Thus, the pyramid start with Conv3_3 (of 80 × 80 pixels) with rich local features. Conv4_3 (40 × 40), Conv5_3 (20 × 20) and Conv_fc7 (10 × 10) are, respectively, upsampled to the same size as Conv3_3 to obtain the cascaded feature map. This feature map contains both the global and local features, which allows us to extract features using the RFB. It also enriches the features on different scales. Details of the local and global features extracted by the cascaded feature pyramid are shown in Table 1. After obtaining the cascaded feature map, we use stride convolution to construct the convolution feature pyramid. The structure of the M-FPM is shown in Figure 3.

Table 1. The details of local and global features in M-FPM.

Layer	Output	Feature
Conv3_3	80 × 80 × 256	local
Conv4_3	40 × 40 × 512	⇓
Conv5_3	20 × 20 × 512	
Conv_fc7	10 × 10 × 1024	global

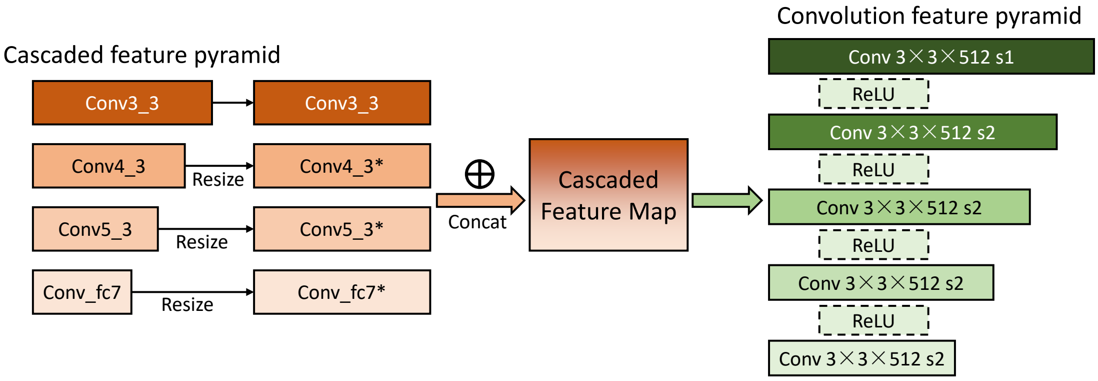

Figure 3. The structure of Multi-Feature Pyramid Module. Mark * refers to the resized feature maps.

3.2. *Receptive Field Block*

Receptive field refers to the area where the convolution kernel can extract information in the feature map. Receptive Field Block is a module that can obtain local context information. In order to simulate the various receptive field sizes of the human eye, the block is designed with a variety dilated convolution with different expansion rate, which can increase the receptive field of convolution while reducing the information loss caused by the operation of max pooling. As shown in Figure 4, for the incoming cascaded feature map, the block first constructs a bottleneck structure using 1 × 1 convolution to reduce features, and then performs convolutions with different kernel size. Asymmetrical convolution kernels like 1 × 3 and 3 × 1 are used to introduce more nonlinear transformation as well as save computational consumption. Dilated convolutions with different expansion rate are then performed to simulate multi-scale receptive field. Finally, the results of the four different branches are integrated and sent into the next layer. The Receptive Field Block is created to help the network generating higher resolution feature maps for local context information and improving the accuracy of object detection. Details of the local context features in RFB are shown in Table 2. The resolution and dimension of the feature map remains the same after concatenating the output of these four branches. Compared with methods using classical receptive field, our operations on the receptive field act on the cascaded feature map, including the context information in different scales.

Table 2. The details of local and global features in RFB.

Branch	Kernel	Input	Output	Receptive Field
1	1 × 1 3 × 3, $R = 1$	80 × 80 × 512 80 × 80 × 128	80 × 80 × 128 80 × 80 × 128	12 × 12
2	1 × 1 3 × 1 3 × 3, $R = 3$	80 × 80 × 512 80 × 80 × 128 80 × 80 × 128	80 × 80 × 128 80 × 80 × 128 80 × 80 × 128	36 × 36
3	1 × 1 1 × 3 3 × 3, $R = 3$	80 × 80 × 512 80 × 80 × 128 80 × 80 × 128	80 × 80 × 128 80 × 80 × 128 80 × 80 × 128	36 × 36
4	1 × 1 1 × 3 3 × 1 3 × 3, $R = 5$	80 × 80 × 512 80 × 80 × 64 80 × 80 × 96 80 × 80 × 128	80 × 80 × 64 80 × 80 × 96 80 × 80 × 128 80 × 80 × 128	60 × 60

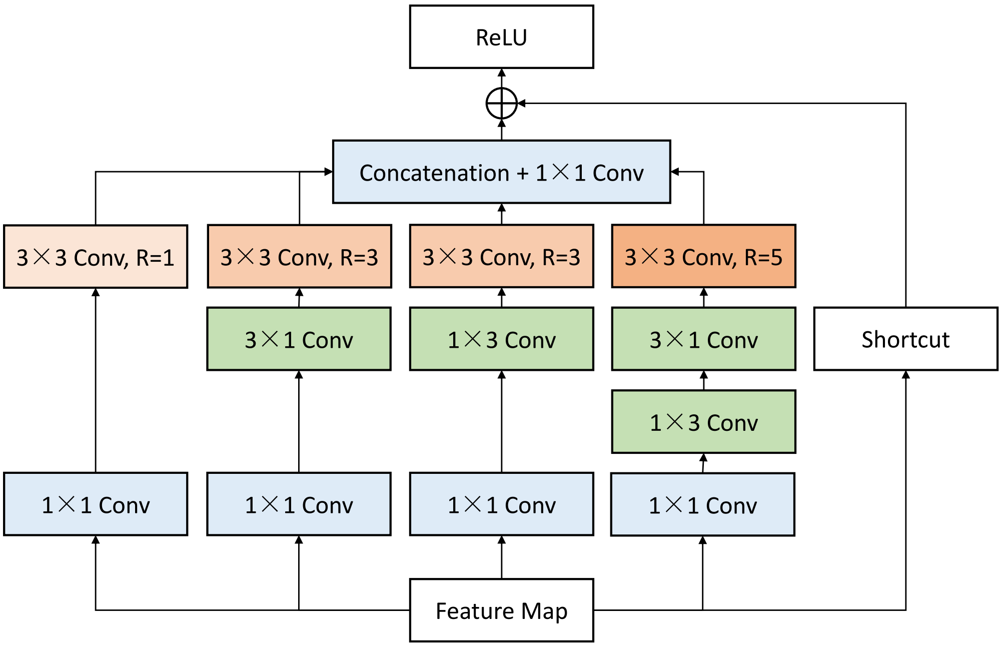

Figure 4. The structure of Receptive Field Block. *R* refers to the expansion rate in the dilated convolution.

3.3. Double-Check Detection Network Module

For object detection networks, the accuracy of bounding box regression is important. Although a one-stage detection network is much faster, it is not good for some challenging conditions, especially for small objects in ORSIs. So, a one-stage double-check network is used to perform better regression and classification of the bounding boxes. Bounding boxes are initiated by anchors. In the first step of detection, a first-time regression is made on the size and position of the bounding box. Specifically, the initial position of the anchor box is fixed. We predict each confidence score from n candidate anchors and determine whether it is foreground or not. At the same time, the network predicts the offset of each anchor box to get the initial accurate position and size of the anchor box. This refined anchor box is then sent to a second detection network for double-check. In the second step, the location and scale of the refined anchors will be further optimized, and a confidence score will be assigned for each category. In order to reduce the impact of easy negative samples on the detection and to make the network focus on hard negative samples, the threshold is set 0.99 to filter the easy negative samples. The link between the two-steps is a Transfer Connection Block (TCB), which is the same as described by RefineDet [4]. The double-check detection network corresponds to the same feature map size, all from the feature pyramid generated by the stride convolution. This design further improves the detection accuracy and is more suitable for detecting small objects.

In order to deal with the various scales of objects, different kinds of anchors are required. The input image resolution is 320 × 320. With the steps of 4, 8, 16, 32, and 64 pixels, the network can obtain five feature layers of different sizes. At the same time, three anchor boxes with aspect ratios of 0.5, 1.0 and 2.0 are designed and placed in different feature layers. In the training process, we determine the relationship between the anchor box and the ground truth box according to the IoU size and match the anchor box with the ground truth box with IoU greater than 0.5.

3.4. Loss Function

Our loss function consists of two parts, the first step detection network loss $\mathcal{L}_d$, and the second step detection network loss $\mathcal{L}_u$. Similar to [4], combining the two-step losses as the total loss function can achieve better performance. The loss function is designed as:

$$\mathcal{L}_d(p_i, x_i) = \frac{1}{N_d}\left(\sum_i \mathcal{L}_b(p_i, [l_i^* \geq 1]) + \sum_i [l_i^* \geq 1]\mathcal{L}_r(x_i, g_i^*)\right) \tag{1}$$

$$\mathcal{L}_u(c_i, t_i) = \frac{1}{N_u}\left(\sum_i \mathcal{L}_m(c_i, l_i^*) + \sum_i [l_i^* \geq 1]\mathcal{L}_r(t_i, g_i^*)\right) \tag{2}$$

$$\mathcal{L}(p_i, x_i, c_i, t_i) = \mathcal{L}_d(p_i, x_i) + \mathcal{L}_u(c_i, t_i) \tag{3}$$

where i is the information of the anchor box, p_i is the confidence level of being the object measured by the first step detection network, x_i is the position of the anchor box obtained by the first step detection network, and c_i is the category of the anchor box in the second step detection network, t_i is the final position of the anchor box in the second step detection network. l_i^* is the category label of the true value, and g_i^* is the location of the true value. N_d and N_u represent the number of anchor boxes that are true in the first and second steps detection network, respectively. The loss function $\mathcal{L}_b$ is a cross-entropy loss function for classifying the anchor box, the loss function $\mathcal{L}_m$ is a softmax loss function for multi-class classification, and the regression loss function $\mathcal{L}_r$ is a Smooth-L1 function.

4. Experiments

In this section, we compare the proposed MFPNet with the state-of-the-art methods. Adequate experiments are conducted on two optical remote sensing datasets, the Levir [24] dataset and the DIOR [26] dataset. Ablation studies are also conducted to prove the effectiveness of each module. We use PyTorch to implement MFPNet, and the experimental environment is CUDA 9.0, cuDNN v6, and NVIDIA GeForce 1080Ti.

4.1. Datasets and Evaluation Metric

The two datasets we adopted have differences in categories and data distributions.

Levir: Levir [24] is a small-scale remote sensing dataset, but it contains small objects in ORSIs, so it is suitable for small and medium scale object detection. Levir contains 3791 high-resolution Google Earth images. The image size of the dataset is 800 × 600. Levir covers most types of environments in ORSIs, such as cities, villages, mountains and oceans. Three types of objects are labeled in the dataset, including aircrafts, ships and oil tanks. There are 11,000 labeled independent objects in total. Although the number instances is relatively small, the categories of targets are common in detection of ORSIs. So Levir is a dataset suitable for verifying the feasibility of the method. The scales of the instances range from smaller than 30 × 30 pixels to larger than 150 × 150 pixels, which is a hard task for a classical detector. The number of instances in the dataset is shown in Figure 5. Numbers of objects with various scales are shown in different color.

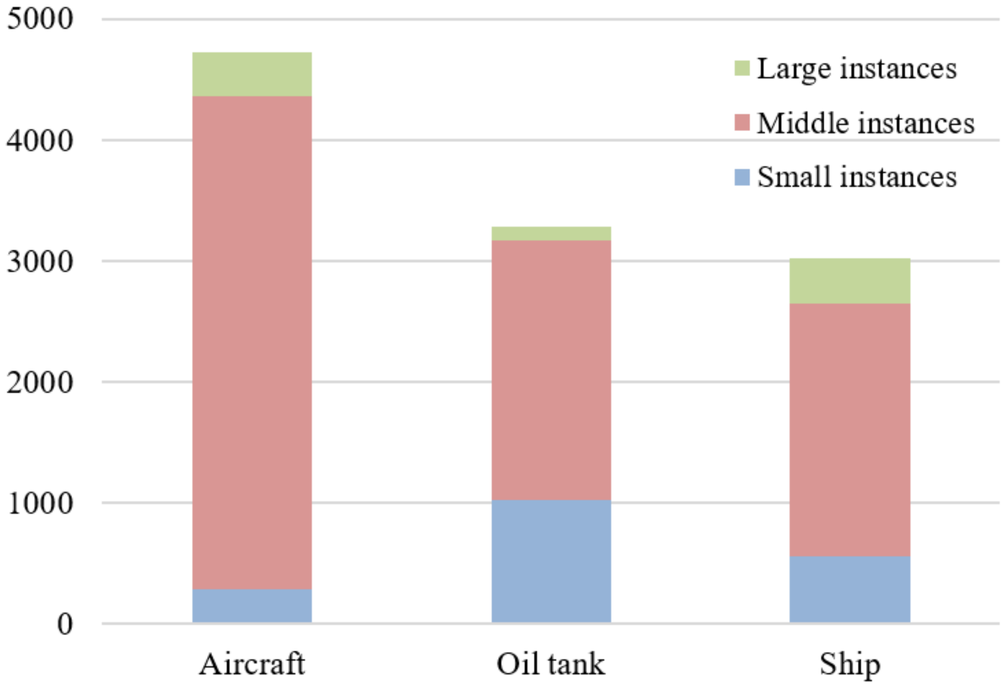

Figure 5. The instance distribution of each category in Levir. Small instances refer to those smaller than 30 × 30 pixels. Middle ones are those between 30 × 30 and 150 × 150 pixels. Large instances are bigger than 150 × 150 pixels.

DIOR: DIOR [26] is currently the largest optical remote sensing dataset. As shown in Figure 6, it contains 23,463 images and 192,472 instances, covering 20 object classes. The size of images in the dataset is 800 × 800 pixels and the spatial resolutions range from 0.5 m to 30 m. The dataset has a large range of object size variations, not only in terms of spatial resolution, but also in the aspect of inter- and intra-class size variability across objects. For example, most instances of ships are small object whose size is smaller than 30 × 30 pixels, while most instances of train stations are large that have a size bigger than 150 × 150 pixels. For small objects such as oil tanks or ships, there are great numbers of scenes that have numerous instances. There are images with complex background in DIOR, making the dataset even harder. Moreover, the data show a long-tailed distribution, that is, some classes such as *Ship* have mass instances, while other classes such as *Train station* have few targets. This is a challenge to the generalization performance of the detection method.

As shown in Figure 7, the first line are images from Levir and the others are from DIOR. Images in DIOR have more complex background. Besides, the problems of scale-difference and scattered objects are more significant in DIOR. In general, DIOR is a more complex dataset comparing with Levir.

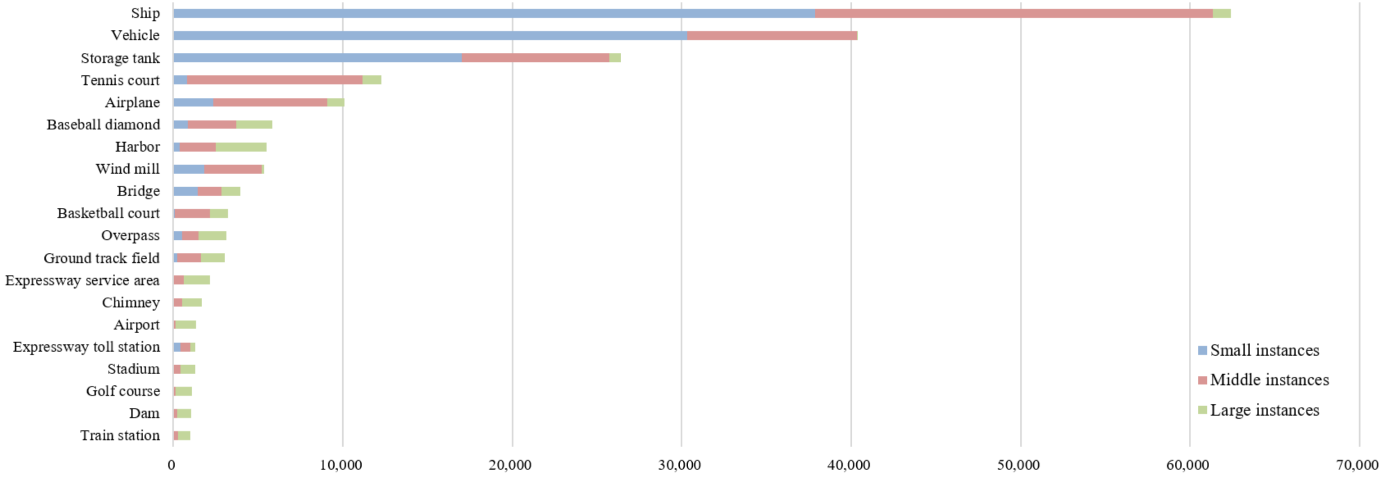

Figure 6. Total number of instances of each category in DIOR.

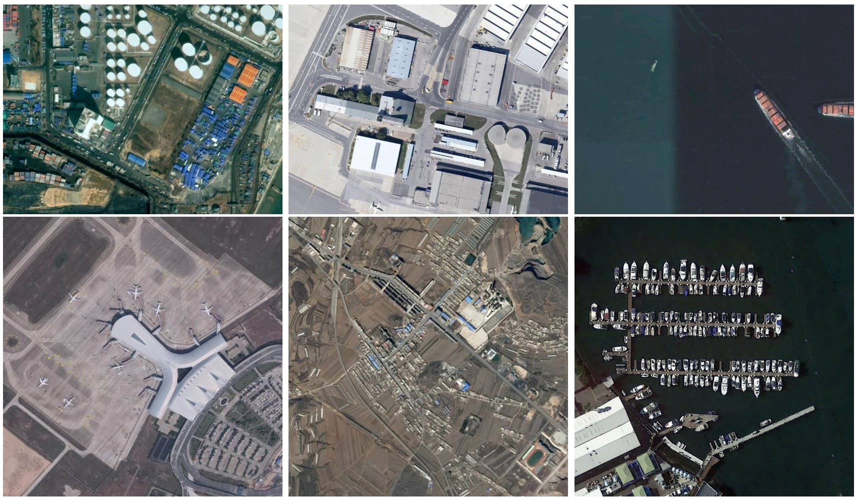

Figure 7. Schematic images of the datasets. The first row are images from Levir and the others are from DIOR. Both datasets have scenes of the three problems pointed out, (i) scattered instances in one image, (ii) complex background, and (iii) scale-differences between targets.

Evaluation Metric: The same as most object detection approaches, mean average precision (mAP) is adopted as the evaluation metric of our method. The mAP is defined as:

$$mAP = \frac{1}{K}\sum_{n=1}^{K}\int(P_n(R_n))dRn \tag{4}$$

where R_n is the recall for a given class n, and $P_n(R_n)$ represents the precision as the recall of this class is R_n. K represents the total number of classes. The recall and precision are defined as follows:

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$P = \frac{TP}{TP + FP} \tag{6}$$

where TP, FN and FP represent the number of true positives, false negatives and false positives, respectively.

4.2. Implementation Details

In this part, we explain some details during training and inference, including the strategy of data augmentation, hard negative mining, and optimization.

Data Augmentation: For the input data, augmentation operations similar to SSD [2] are performed, including mirror inversion, random cropping, random mirror flipping, image size scaling and other steps. The significance of data expansion is to make the network more robust and adapt to different situations.

Hard Negative Mining: In order to alleviate the imbalance between positive and negative anchors, and make the training of the network more effective, it is necessary to remove those easy negative anchors and make the network more concentrated on learning hard ones. We introduced hard negative mining, selecting the negative anchor boxes with the highest loss value to learn, and keeping the ratio of foreground and background anchors to 1:3.

Optimization and Inference: During the training process, an ImageNet pre-trained model is loaded to the backbone. The parameters of the additional layers are randomly initialized using Kaiming method [48]. The batch size during training is set to 32. We use the Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a weight decay of 0.0005 to fine-tune the entire network. For Levir dataset, the learning rate is initialized to 0.001 for the first 40,000 iterations. In 40k to 80k iterations, the learning rate is attenuated to 10^{-4}. In the last 40k iterations, we attenuate the learning rate to 10^{-5}.

During the inference process, the first step of the double-check detection network deletes anchors with negative confidence scores T larger than 0.99 and then refine the positions of other anchors. The second step of the network uses the refined anchors for classification and further regression. Non-maximum suppression is applied to obtain the final result.

4.3. Experimental Results

For experiments, we split the test data into different cases such as images with incomplete or scattered objects, so as to prove that our model can achieve promising results in various complex situations.

4.3.1. Levir Dataset

We first conduct experiments on Levir dataset. VGG16 is adopted as the backbone. The proposed MFPNet is compared with several state-of-the-art method, including Faster R-CNN [10], SSD [2], RetinaNet [1], RefineDet [4] and CenterNet [16]. The experimental results are shown in Table 3.

Table 3. Detection results on the Levir test set. All models are trained on Levir trainval set. The best performance are shown in red.

Method	Backbone	Airplane	Ship	Oil Tank	mAP
Faster R-CNN [10]	VGG16	90.1	89.5	71.0	83.6
Faster R-CNN [10]	ResNet-50 [49]	87.6	81.6	71.9	80.4
SSD300 [2]	ResNet-50	87.7	81.5	68.7	79.3
RetinaNet500 [1]	ResNet-50	87.6	80.2	74.0	80.6
RefineDet320 [4]	VGG16	90.7	89.3	85.4	88.5
CenterNet-DLA [16]	DLA-34	81.7	79.0	75.7	78.8
Ours	VGG16	90.7	88.8	89.8	89.8

The input size of images will affect the detection accuracy. Many researchers [1–3] have proved that the greater the size of the input image, the detection of small objects will be more friendly, and the detection results will tend to be better, especially for the large number of small objects in the ORSIs. Using a lower input size 320 × 320 (due to insufficient computing resources, we are unable to test 512 × 512 inputs. However, through experience, we believe that the detection accuracy of the larger input will be much higher), MFPNet is capable of generating 89.8% mAP, exceeding all current one-stage and two-stage methods. The input size of these methods is greater than or equal to ours, which proves that our network has favorable performance on the detection of small objects in ORSIs. The results on Levir dataset is shown in Figures 8–10.

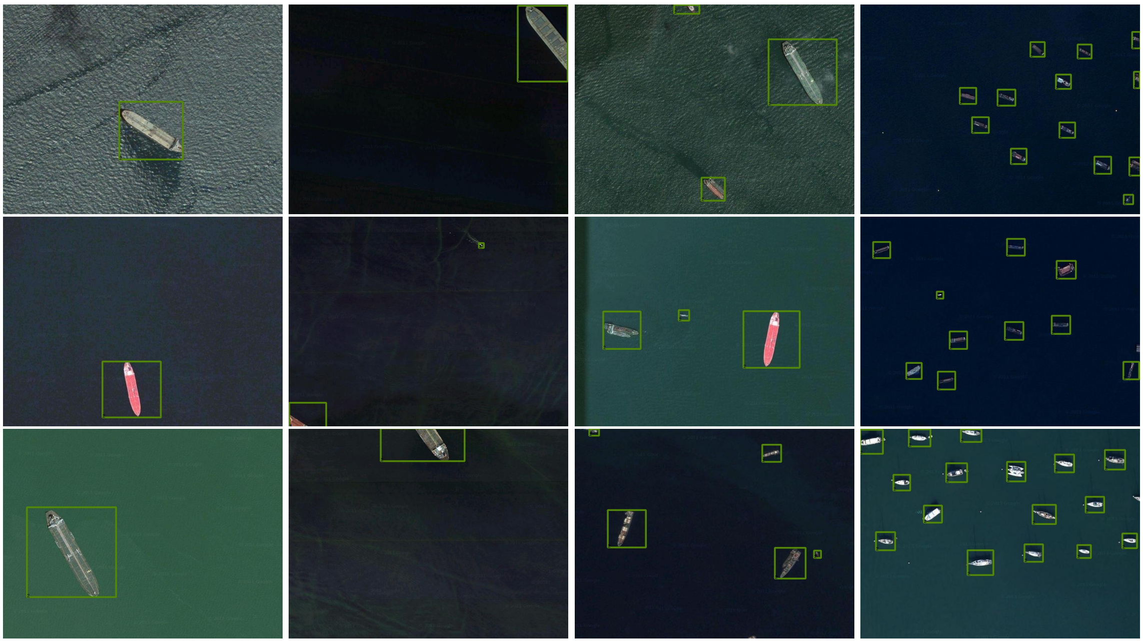

Figure 8. Results of ships on the Levir dataset. For cases of incomplete objects, size differences and huge number of targets, our method provides promising results.

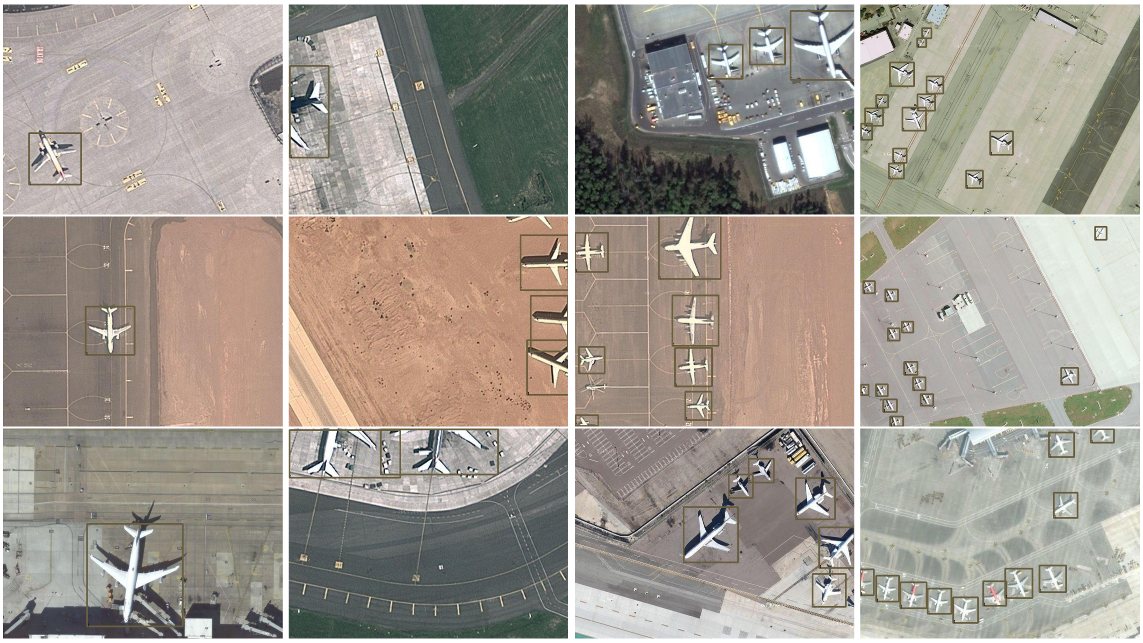

Figure 9. Results of aircrafts on the Levir dataset. For cases of incomplete objects, size differences and huge number of targets, our method provides promising results.

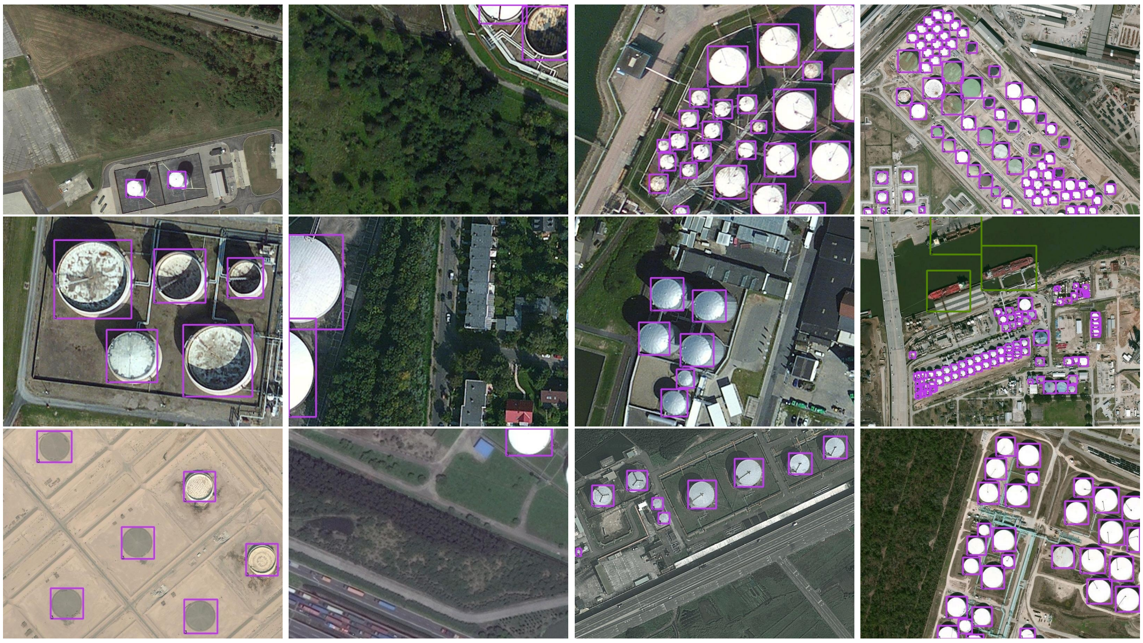

Figure 10. Results of oil tanks on the Levir dataset. For cases of incomplete objects, size differences and huge number of targets, our method provides promising results.

Figure 8 shows the results of ships. The first column shows the most common cases. The second column contains some broken objects. MFPNet detects these fragmentary cases correctly. The third column contains instances with different sizes. The M-FPM mechanism ensure the accuracy of these cases. The last column are images that contain huge numbers of targets. Though the objects in these images are small, our method can accurately detect them. Figure 9 shows the results of aircrafts. Similar to Figure 8, the first column are easy cases and the second column contains broken objects cases. The third and the last column refer to images with different size objects and mass instances. MPFNet provides accurate detection. Figure 10 shows the results of oil tanks. It also contains four cases similar to Figures 8 and 9. It is worth mentioning that though the instances in the last column is very small, our network still have high detection quality.

4.3.2. DIOR Dataset

DIOR [26] is the latest large-scale optical remote sensing dataset for target location task. The advantage of this dataset is that it contains objects with different scales. As shown in Table 4, in order to better display the results, we numbered each category in the dataset.

Table 4. Categories in the DIOR dataset and their corresponding numbers.

c1	Airplane	c6	Chimney	c11	Ground track field	c16	Storage tank
c2	Airport	c7	Dam	c12	Harbor	c17	Tennis court
c3	Baseball field	c8	Expressway service area	c13	Overpass	c18	Train station
c4	Basketball court	c9	Expressway toll station	c14	Ship	c19	Vehicle
c5	Bridge	c10	Golf court	c15	Stadium	c20	Wind mill

The specific experimental results are shown in Table 5. The input size is also set as 320 × 320 and the learning rate is the same size as the experiment in Levir. MFPNet is capable of generating 71.2% of mAP, which is 4.1% higher than the RefineDet method, exceeding all current one and two-stage methods. Among the 20 classes of objects, most of the classes have reached the top two highest accuracies. It means the model proposed in this paper can show robustness against the imbalanced data distribution. At the same time, experiments are also conducted by using various backbone networks including VGG16, ResNet-50, etc. Furthermore, we exploit the potential capability of the proposed MFPNet. ResNet-50 has fewer network parameters than VGG16, but the network structure is much deeper. From the results in Table 5, VGG16 performs better in the overall results, but for objects with rich texture information, the detection results of ResNet-50 are much better. The overall performance of our network excels all the comparison methods. Even if we replace the backbone, it maintains favorable results and proves the effectiveness of the network structure. The performance of our proposed MFPNet on multi-scale optical remote sensing objects is also outstanding. In the first place, both the airplane and the ship are small objects. MFPNet obtains high mAP on these categories. For objects such as airports, stadiums and basketball courts, which have special texture features, our results are all in the top two.

As shown in Figure 11, the first row shows the results of numerous objects. MFPNet can detect almost all the objects even they are small and scattered. The second row shows the conditions that the texture of the background is complex. It is hard for the network to identify the objects from the background. However, our proposed model can find the objects in most cases. The third row shows the results when objects with different scale exist in one image. The M-FPM ensure our network to find all of the objects. The last row shows several results of other categories. According to the results in Table 5, YOLO-v3 [47] is very good for small object detection. However, for larger objects, such as airports and golf courses, the test result is the worst. It means that YOLO-v3 has poor adaptability to multi-scale objects. RICNN [28], RICAOD [29] and RIFD-CNN [27], which are designed for optical remote sensing objects, achieve worse results on the DIOR dataset than other methods because they are not adapt to multi-scale objects. Comparing to all of these methods, the proposed MFPNet is more capable in detection tasks of ORSIs.

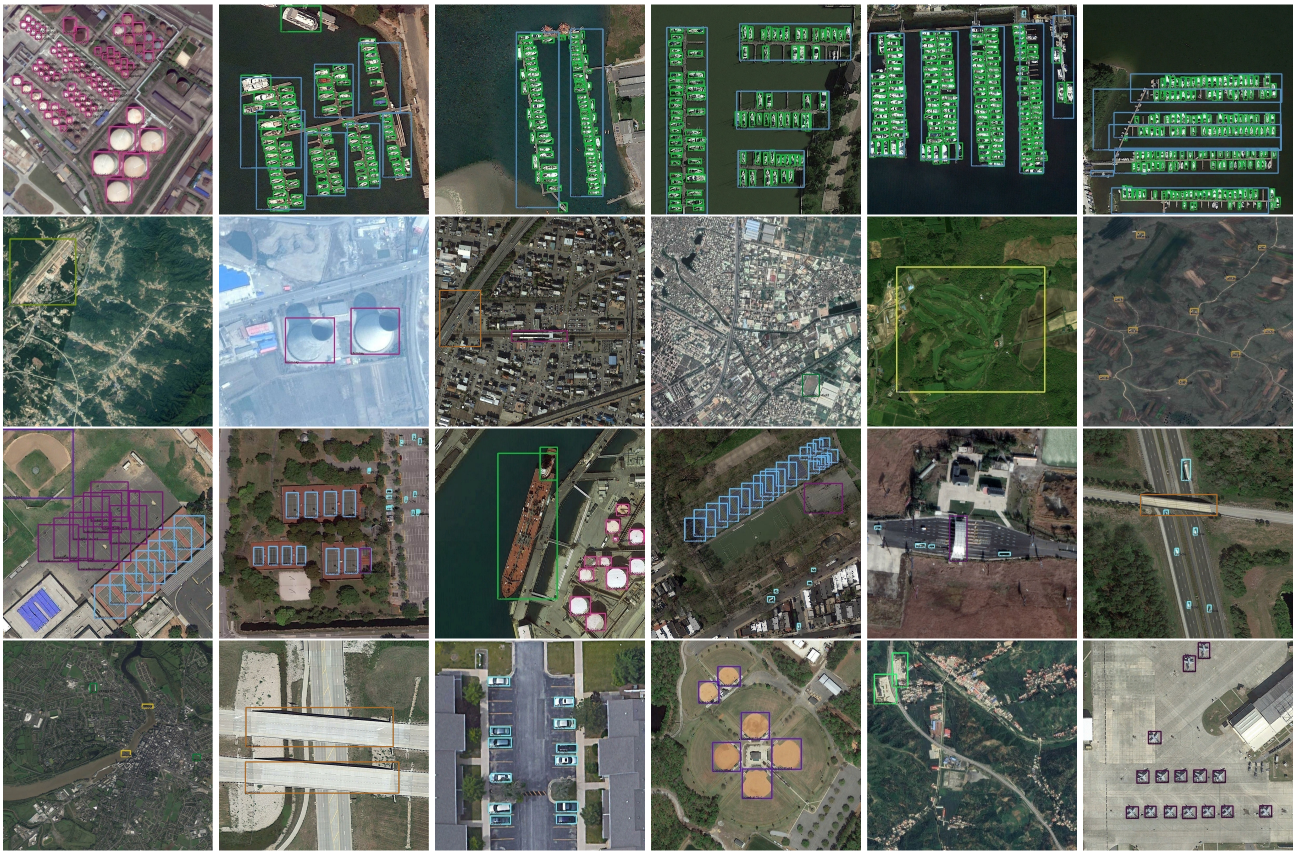

Figure 11. Qualitative results of MFPNet320 on the DIOR test set. The first row is the detection results when a large number of scattered objects. The second row shows the results when the background texture is complex. The third row contains the detection results when scale difference exists in and between classes. The fourth row are some results of other cases.

Table 5. Detection results on the DIOR test set. All models are trained on the DIOR trainval set. * refers to adding FPN to the method. The best performances are marked in red, and the second ones are marked in blue. Green indicates the third best performances.

Method	Backbone	mAP	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16	c17	c18	c19	c20
R-CNN [26]	VGG16	37.7	35.6	43.0	53.8	62.3	15.6	53.7	33.7	50.2	33.5	50.1	49.3	39.5	30.9	9.1	60.8	18.0	54.0	36.1	9.1	16.4
Faster R-CNN [26]	VGG16	54.1	53.6	49.3	78.8	66.2	28.0	70.9	62.3	69.0	55.2	68.0	56.9	50.2	50.1	27.7	73.0	39.8	75.2	38.6	23.6	45.4
Faster R-CNN * [26]	ResNet-50	63.1	54.1	71.4	63.3	81.0	42.6	72.5	57.5	68.7	62.1	73.1	76.5	42.8	56.0	71.8	57.0	53.5	81.2	53.0	43.1	80.9
Faster R-CNN * [26]	ResNet-101	65.1	54.0	74.5	63.3	80.7	44.8	72.5	60.0	75.6	62.3	76.0	76.8	46.4	57.2	71.8	68.3	53.8	81.1	59.5	43.1	81.2
Mask R-CNN * [26]	ResNet-50	63.5	53.8	72.3	63.2	81.0	38.7	72.6	55.9	71.6	67.0	73.0	75.8	44.2	56.5	71.9	58.6	53.6	81.1	54.0	43.1	81.1
Mask R-CNN * [26]	ResNet-101	65.2	53.9	76.6	63.2	80.9	40.2	72.5	60.4	76.3	62.5	76.0	75.9	46.5	57.4	71.8	68.3	53.7	81.0	62.3	43.0	81.0
PANet [26]	ResNet-50	63.8	61.9	70.4	71.0	80.4	38.9	72.5	56.6	68.4	60.0	69.0	74.6	41.6	55.8	71.7	72.9	62.3	81.2	54.6	48.2	86.7
PANet [26]	ResNet-101	66.1	60.2	72.0	70.6	80.5	43.6	72.3	61.4	72.1	66.7	72.0	73.4	45.3	56.9	71.7	70.4	62.0	80.9	57.0	47.2	84.5
CBD-E [50]	ResNet-101	67.8	54.2	77.0	71.5	87.1	44.6	75.4	63.5	76.2	65.3	79.3	79.5	47.5	59.3	69.1	69.7	64.3	84.5	59.4	44.7	83.1
SSD300 [26]	VGG16	58.6	59.5	72.7	72.4	75.7	29.7	65.8	56.6	63.5	53.1	65.3	68.6	49.4	48.1	59.2	61.0	46.6	76.3	55.1	27.4	65.7
YOLO-v3 [26]	Darknet-53	57.1	72.2	29.2	74.0	78.6	31.2	69.7	26.9	48.6	54.4	31.1	61.1	44.9	49.7	87.4	70.6	68.7	87.3	29.4	48.3	78.7
RetinaNet500 [26]	ResNet-50	65.7	53.7	77.3	69.0	81.3	44.1	72.3	62.5	76.2	66.0	77.7	74.2	50.7	59.6	71.2	69.3	44.8	81.3	54.2	45.1	83.4
RetinaNet500 [26]	ResNet-101	66.1	53.3	77.0	69.3	85.0	44.1	73.2	62.4	78.6	62.8	78.6	76.6	49.9	59.6	71.1	68.4	45.8	81.3	55.2	44.4	85.5
RefineDet320 [4]	VGG16	67.1	69.5	80.4	74.4	81.1	40.0	72.7	68.8	80.2	58.9	77.7	74.2	61.3	57.8	63.3	75.3	47.3	81.3	65.7	34.7	78.2
CornerNet [26]	Hourglass-104	64.9	58.8	84.2	72.0	80.8	46.4	75.3	64.3	81.6	76.3	79.5	79.5	26.1	60.6	37.6	70.7	45.2	84.0	57.1	43.0	75.9
M2Det320 [42]	VGG16	44.0	54.7	61.4	67.1	54.6	16.7	61.6	33.2	60.1	51.7	58.5	60.2	19.6	32.7	31.3	63.0	12.4	71.4	21.5	9.0	38.2
CenterNet [16]	Hourglass-104	52.4	50.2	51.2	62.2	62.3	31.7	61.0	38.5	63.1	57.0	57.3	56.6	26.2	41.1	58.3	54.1	49.7	73.6	41.7	40.5	66.8
RICNN [26]	VGG16	44.2	39.1	61.0	60.1	66.3	25.3	63.3	41.1	51.7	36.6	55.9	58.9	43.5	39.0	9.1	61.1	19.1	63.5	46.1	11.4	31.5
RICAOD [26]	VGG16	50.9	42.2	69.7	62.0	79.0	27.7	68.9	50.1	60.5	49.3	64.4	65.3	42.3	46.8	11.7	53.5	24.5	70.3	53.3	20.4	56.2
RIFD-CNN [26]	VGG16	56.1	56.6	53.2	79.9	69.0	29.0	71.5	63.1	69.0	56.0	68.9	62.4	51.2	51.1	31.7	73.6	41.5	79.5	40.1	28.5	46.9
MFPNet320	ResNet50	69.1	73.6	80.6	80.2	80.9	41.1	74.0	69.3	84.2	63.3	76.3	74.6	62.8	58.2	69.9	71.5	55.5	82.3	66.4	38.5	79.1
MFPNet320	VGG16	71.2	76.6	83.4	80.6	82.1	44.3	75.6	68.5	85.9	63.9	77.3	77.2	62.1	58.8	77.2	76.8	60.3	86.4	64.5	41.5	80.2

Comparing with existing approaches, MFPNet can better handle the main problems in ORSIs including complex background interference, targets scale-differences, and numerous scattered instances in one image. Figures 12–14 demonstrate the superior performance of MFPNet. In each case, the first column stands for Faster R-CNN and the second column shows the results of RetinaNet. The results of MFPNet are shown in the third column.

As shown in Figure 12, there are train stations or playgrounds in the complex background. Faster R-CNN and RetinaNet miss several targets while MFPNet detects all of them. In Figure 13, it is hard for Faster R-CNN or RetinaNet to detect huge harbors and small ships simultaneously. However, MFPNet can achieve better results. As for Figure 14, MFPNet detects almost all scattered targets such as oil tanks and ships, surpassing other methods.

Though the results of MFPNet are satisfying in most cases, there are actually some limitations. False cases are shown in Figure 15. MFPNet generates horizontal anchors. When facing oriented targets, these anchors may extract features from the background, misleading the detector and resulting in false detection. In response to this limitation, we will introduce oriented anchors to MFPNet in the future investigation.

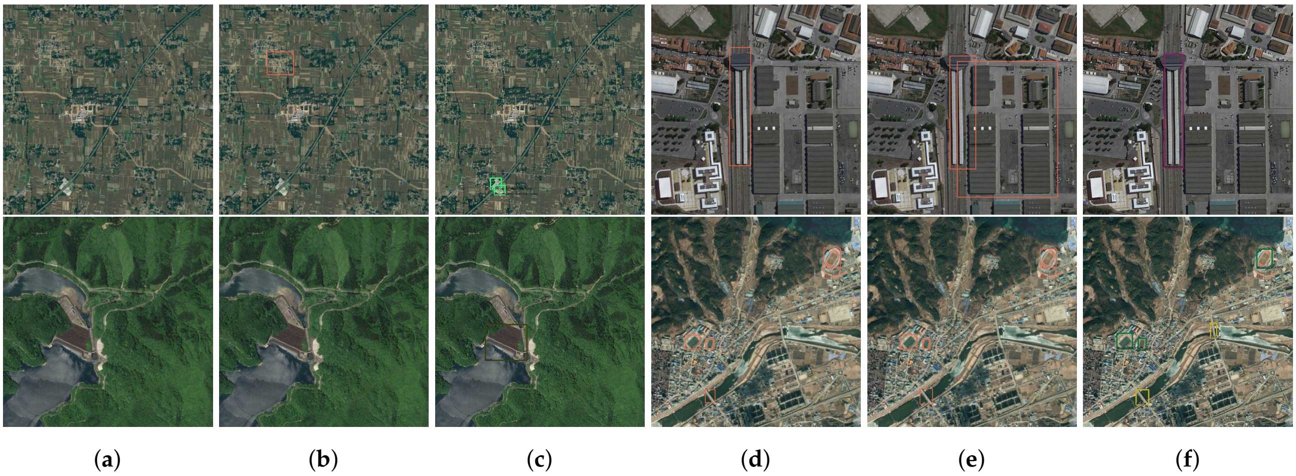

Figure 12. The results of images with complex background. Faster R-CNN (**a**,**d**) and RetinaNet (**b**,**e**) have several false or missed detection. MFPNet (**c**,**f**) gives better results.

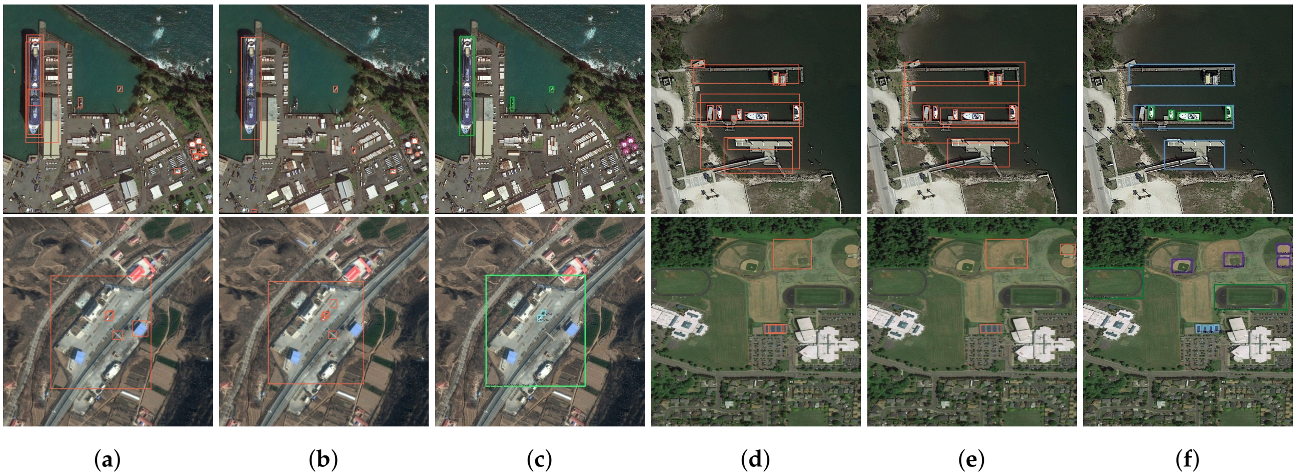

Figure 13. The results of images having targets with scale differences. Comparing with Faster R-CNN (**a**,**d**) and RetinaNet (**b**,**e**), MFPNet (**c**,**f**) makes better detection.

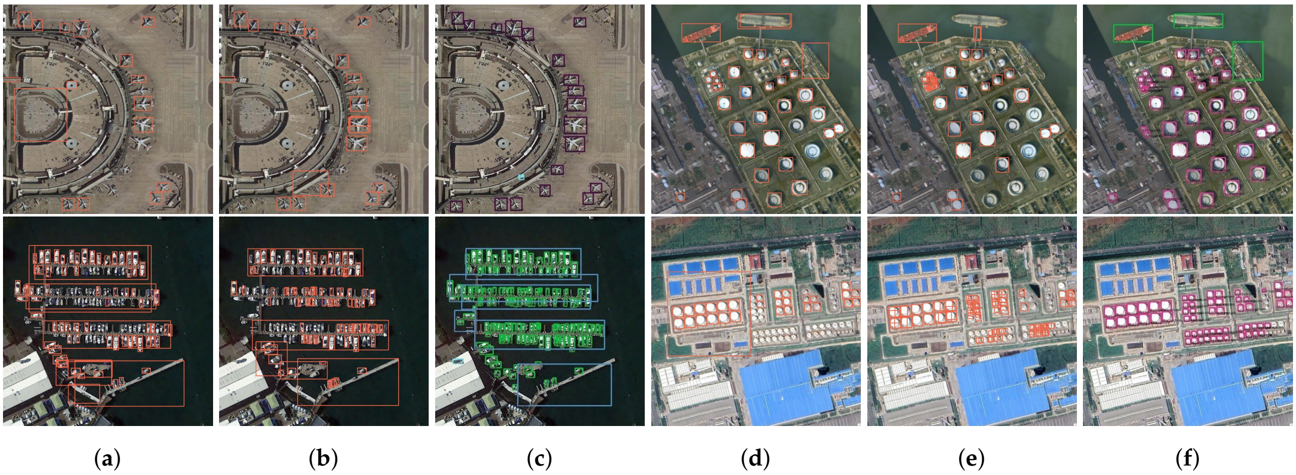

Figure 14. The results of images with numerous scattered targets. Faster R-CNN (**a**,**d**) and RetinaNet (**b**,**e**) have many missed detection, while MFPNet (**c**,**f**) detects almost all objects.

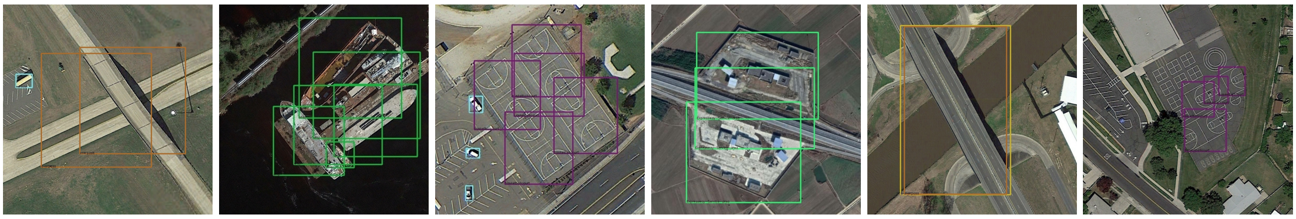

Figure 15. False cases in DIOR dataset. False detection may occur when facing dense arrays of oriented objects.

4.4. Ablation Study

In order to verify the contribution of all modules, we test each module through the ablation studies. We construct five different networks. They adopt the same hyper-parameters and use the same VGG16 as the backbone network. In order to prove the validity of our method, all experiments were trained and tested on the DIOR dataset. The results are shown in Table 6.

Multi-Feature Pyramid Module: In order to prove the validity of the Multi-Feature Pyramid Module, we remove the first cascaded feature pyramid and send Conv3_3, Conv4_3, Conv5_3 and Conv_fc7 of VGG16 directly to the detection network. The RFB are retained for context information extraction of Conv3_3. As shown in the forth line of Table 6, the final mAP is reduced by 6.3% when testing without M-FPM. The decline indicates that the Multi-Feature Pyramid Module enhances the ability to integrate features. The object features with different scales are also enhanced. The result indicates that with the help of the M-FPM, the network itself has a comprehensive improvement in detecting optical remote sensing objects. The detection results of most classes are higher than those without the M-FPM. The Floating Point Operations (FLOPs) and Multiply-Adds (MAdd) drops about a quarter, suggesting that M-FPM is a main part of the network. The favorable improvement made by this module demonstrates the importance of multi-scale and global-local feature fusion for optical remote sensing objects detection.

Receptive Field Block: In order to verify the validity of the Receptive Field Block, we retain the Multi-Feature Pyramid Module and remove the Receptive Field Block that extracts context information from the cascaded feature map. As shown in the third line of Table 6, the final test result is reduced by 0.6%. It is because that the RFB enhances the context information of the network, and it has a positive effect on both small and large objects, which improves the performance obviously. As shown in Table 6, when working without cascaded convolutional map, the feature extraction effect on small objects will be more obvious. It works better for texture feature extraction when acting on the cascaded convolutional map, as evidenced by the outstanding result of detecting expressway service areas, expressway toll stations and ground track fields. The MAdd and FLOPs only drops about a tenth. It means that the RFP is an efficient module to improve the performance of the network.

Table 6. Results of ablation experiment. All models are trained on the DIOR trainval set and tested on the DIOR test set.

Cascade Layers	M-FPM	RFB	MAdd	FLOPs	mAP	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16	c17	c18	c19	c20
5	✓	✓	200.6 G	99.8 G	**71.2**	76.6	83.4	80.6	82.1	44.3	75.6	68.5	85.9	63.9	77.3	77.2	62.1	58.8	77.2	76.8	60.3	86.4	64.5	41.5	80.2
4	✓	✓	140.5 G	71.7 G	67.3	70.6	84.1	74.9	81.3	40.9	74.6	70.1	82.2	59.0	77.7	75.2	62.2	58.3	55.8	76.8	45.0	81.3	66.5	32.6	76.3
5	✓	-	181.4 G	90.2 G	70.6	77.1	81.4	80.6	81.1	43.7	74.3	69.8	84.5	63.0	77.3	76.4	62.2	58.9	76.2	76.8	60.0	86.7	62.2	40.8	79.9
5	-	✓	153.9 G	76.5 G	64.9	66.7	78.7	71.8	81.0	35.0	74.0	65.2	79.5	53.7	78.5	73.9	61.7	55.6	57.8	79.5	41.2	81.0	59.8	30.8	72.5
5	-	-	134.7 G	66.9 G	63.2	60.8	78.1	71.0	80.7	32.7	72.7	62.4	77.8	52.8	75.9	70.1	60.1	53.2	59.5	76.7	40.5	80.5	58.8	29.4	70.2

Feature Map Scale: Experiments are also conducted for verifying the impact of the number of layers in the M-FPM. The cascade layers in Table 6 refers to the number of layers in the M-FPM. For the model with five cascade layers, Conv3_3, Conv4_3, Conv5_3, Conv_fc7 and an extra Conv layer are included in the cascaded feature pyramid. When the number of cascade layers is four, only Conv4_3, Conv5_3, Conv_fc7 and the extra layer are sent to the M-FPM. That is, the model with five layers in the pyramid has larger size feature maps. As shown in the second line in Table 6, when using four layers in M-FPM, the final result is reduced by 3.9%. The drop in mAP indicates that local features of the images contained in Conv3_3 have a positive effect on the detection of ORSIs, because the distribution of remote sensing objects is dispersed, and the local features have a greater influence on the detection accuracy of dispersed objects. The increase of local features is more conducive to detect scattered distribution objects. As shown in Table 6, the detection result of scattered objects such as airplanes, storage tanks and vehicles has been greatly improved.

5. Conclusions

Aiming at the problems such as scale-differences between classes, numerous instances in one image and complex background texture in ORSIs object detection, we propose an end-to-end network called MFPNet. We construct a Multi-Feature Pyramid Module to combine the global semantic features and local detailed features. A Receptive Field Block is designed to increase the receptive field and obtain local context information. Our method can better extract features of scattered or multi-scale objects with complicated background in ORSIs, and performs better than current state-of-the-art methods on the Levir benchmark, achieving 89.8% mAP . On DIOR dataset, the result is 71.2% mAP, which is higher than other approaches by more than 4.1%. MFPNet has a limitation when detecting oriented objects. In the future, we will explore the potential of anchor-free or oriented anchor-based methods for small object detection and hope to more accurately detect oriented objects and further improve the performance of the network.

Author Contributions: Conceptualization, D.Z.; methodology, Z.L.; software, Z.L.; validation, Z.Y., Z.L. and C.Z.; formal analysis, Z.Y.; resources, J.Q.; writing—original draft preparation, Z.Y. and Z.L.; writing—review and editing, C.Z. and D.Z.; visualization, C.Z.; supervision, J.Q.; funding acquisition, D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under Grant No. 2019YFC1510905, and in part by the Air Force Equipment Pre-research Project under Grant 303020401 (Corresponding author: Danpei Zhao).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

Abbreviations

The following abbreviations are used in this manuscript:

CNNs	convolutional neural networks
ORSIs	optical remote sensing images
MFPNet	Multi-Feature Pyramid Network
RFB	Receptive Field Block
ROIs	Regions of Interest
RPN	Region Proposal Network
IoU	Intersection over Union
M-FPM	Multi-Feature Pyramid Module
mAP	mean average precision
Conv	Convolution Layer
TCB	Transfer Connection Block
SGD	Stochastic Gradient Descent
FLOPs	Floating Point Operations
Madd	Multiply-Adds

Notations

The following notations are used in this manuscript:

i	the identifier of an anchor box
$\mathcal{L}_d$	loss of the first step detection network
N_d	the number of true anchors in the first step detection
$\mathcal{L}_b$	the classifying loss in the first step detection, which is a cross entropy loss
p_i	the confidence level of being an object
l_i^*	the category label
$\mathcal{L}_r$	the Smooth-L1 regression loss
x_i	the position of the anchor box obtained by the first step detection
g_i^*	the ground truth's location
$\mathcal{L}_u$	loss of the second step detection network
N_u	the number of true anchors in the second step detection
$\mathcal{L}_m$	the classifying loss in the second step detection, which is a softmax loss
c_i	the predicted category of the anchor box
t_i	the final prediction of location
$\mathcal{L}$	total loss of MFPNet
K	the total number of classes
n	the identifier of a category
R_n	the recall for a given class n
P_n	the precision for a given class n
TP	the number of true positives
FN	the number of false negatives
FP	the number of false positives

References

1. Lin, T.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
2. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision-ECCV 2016-14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I; pp. 21–37.
3. Redmon, J.; Divvala, S.K.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
4. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-Shot Refinement Neural Network for Object Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4203–4212.
5. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
6. Girshick, R.B. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
7. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
8. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
9. Lin, T.; Dollár, P.; Girshick, R.B.; He, K.; Hariharan, B.; Belongie, S.J. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
10. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
11. Everingham, M.; Eslami, S.M.A.; Gool, L.V.; Williams, C.K.I.; Winn, J.M.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]

12. Everingham, M.; Gool, L.V.; Williams, C.K.I.; Winn, J.M.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
13. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision-ECCV 2014-13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V; pp. 740–755.
14. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. *arXiv* **2019**, arXiv:1904.08189.
15. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. *arXiv* **2019**, arXiv:1904.01355.
16. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
17. Bhagavatula, C.; Zhu, C.; Luu, K.; Savvides, M. Faster than Real-Time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; pp. 4000–4009.
18. Zheng, Y.; Pal, D.K.; Savvides, M. Ring Loss: Convex Feature Normalization for Face Recognition. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5089–5097.
19. He, T.; Shen, C.; Tian, Z.; Gong, D.; Sun, C.; Yan, Y. Knowledge Adaptation for Efficient Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 578–587.
20. Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; Wang, J. Structured Knowledge Distillation for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 2604–2613.
21. Tian, Z.; He, T.; Shen, C.; Yan, Y. Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 3126–3135.
22. Liang, X.; Wang, T.; Yang, L.; Xing, E.P. CIRL: Controllable Imitative Reinforcement Learning for Vision-Based Self-driving. In Proceedings of the Computer Vision-ECCV 2018-15th European Conference, Munich, Germany, 8–14 September 2018; Proceedings, Part VII; pp. 604–620.
23. Wang, D.; Devin, C.; Cai, Q.; Yu, F.; Darrell, T. Deep Object-Centric Policies for Autonomous Driving. In Proceedings of the International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, 20–24 May 2019; pp. 8853–8859.
24. Zou, Z.; Shi, Z. Random Access Memories: A New Paradigm for Target Detection in High Resolution Aerial Remote Sensing Images. *IEEE Trans. Image Process.* **2018**, *27*, 1100–1111. [CrossRef] [PubMed]
25. Xia, G.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.J.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
26. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *arXiv* **2019**, arXiv:1909.00133.
27. Cheng, G.; Han, J.; Zhou, P.; Xu, D. Learning Rotation-Invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection. *IEEE Trans. Image Process.* **2019**, *28*, 265–278. [CrossRef]
28. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
29. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2337–2348. [CrossRef]
30. Liu, L.; Pan, Z.; Lei, B. Learning a Rotation Invariant Detector with Rotatable Bounding Box. *arXiv* **2017**, arXiv:1711.09405.
31. Liu, W.; Ma, L.; Chen, H. Arbitrary-Oriented Ship Detection Framework in Optical Remote-Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 937–941. [CrossRef]
32. Tang, T.; Zhou, S.; Deng, Z.; Lei, L.; Zou, H. Arbitrary-Oriented Vehicle Detection in Aerial Imagery with Single Convolutional Neural Networks. *Remote Sens.* **2017**, *9*, 1170. [CrossRef]
33. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
34. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
35. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the Computer Vision-ECCV 2018-15th European Conference, Munich, Germany, 8–14 September 2018; Proceedings, Part XIV; pp. 765–781.
36. Liu, W.; Ma, L.; Wang, J.; Chen, H. Detection of Multiclass Objects in Optical Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 791–795. [CrossRef]
37. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
38. Han, X.; Zhong, Y.; Zhang, L. An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]

39. Zhong, J.; Lei, T.; Yao, G. Robust Vehicle Detection in Aerial Images Based on Cascaded Convolutional Neural Networks. *Sensors* **2017**, *17*, 2720. [CrossRef] [PubMed]
40. Tang, T.; Zhou, S.; Deng, Z.; Zou, H.; Lei, L. Vehicle Detection in Aerial Images Based on Region Convolutional Neural Networks and Hard Negative Example Mining. *Sensors* **2017**, *17*, 336. [CrossRef]
41. Fu, C.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD : Deconvolutional Single Shot Detector. *arXiv* **2017**, arXiv:1701.06659.
42. Zhao, Q.; Sheng, T.; Wang, Y.; Tang, Z.; Chen, Y.; Cai, L.; Ling, H. M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network. In Proceedings of the The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019; pp. 9259–9266.
43. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Singh, S.P., Markovitch, S., Eds.; AAAI Press: San Francisco, CA, USA, 2017; pp. 4278–4284.
44. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, 22–29 October 2017; IEEE Computer Society; pp. 764–773.
45. Chen, L.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
46. Liu, S.; Huang, D.; Wang, Y. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the Computer Vision-ECCV 2018-15th European Conference, Munich, Germany, 8–14 September2018; Proceedings, Part XI; pp. 404–419.
47. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
50. Zhang, J.; Xie, C.; Xu, X.; Shi, Z.; Pan, B. A Contextual Bidirectional Enhancement Method for Remote Sensing Image Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2020**, *13*, 4518–4531. [CrossRef]