remote sensing

MDPI

Article

Attention-Based Multi-Level Feature Fusion for Object Detection in Remote Sensing Images

Xiaohu Dong [1], Yao Qin [2], Yinghui Gao [3], Ruigang Fu [1,*], Songlin Liu [4] and Yuanxin Ye [5]

[1] College of Electronic Science, National University of Defense Technology, Changsha 410073, China
[2] Remote Sensing Laboratory, Northwest Institute of Nuclear Technology, Xi'an 710024, China
[3] Warfare Studies Institute, Academy of Military Sciences, Beijing 100091, China
[4] State Key Laboratory of Geo-Information Engineering, Xi'an 710024, China
[5] Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610031, China
* Correspondence: furuigang08@nudt.edu.cn

Abstract: We study the problem of object detection in remote sensing images. As a simple but effective feature extractor, Feature Pyramid Network (FPN) has been widely used in several generic vision tasks. However, it still faces some challenges when used for remote sensing object detection, as the objects in remote sensing images usually exhibit variable shapes, orientations, and sizes. To this end, we propose a dedicated object detector based on the FPN architecture to achieve accurate object detection in remote sensing images. Specifically, considering the variable shapes and orientations of remote sensing objects, we first replace the original lateral connections of FPN with Deformable Convolution Lateral Connection Modules (DCLCMs), each of which includes a 3×3 deformable convolution to generate feature maps with deformable receptive fields. Additionally, we further introduce several Attention-based Multi-Level Feature Fusion Modules (A-MLFFMs) to integrate the multi-level outputs of FPN adaptively, further enabling multi-scale object detection. Extensive experimental results on the DIOR dataset demonstrated the state-of-the-art performance achieved by the proposed method, with the highest mean Average Precision (mAP) of 73.6%.

Keywords: object detection; remote sensing; deformable convolution; multi-level feature fusion; attention module

Citation: Dong, X.; Qin, Y.; Gao, Y.; Fu, R.; Liu, S.; Ye, Y. Attention-Based Multi-Level Feature Fusion for Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3735. https://doi.org/10.3390/rs14153735

Academic Editors: Olga Sykioti, Gangyao Kuang and Xin Su

Received: 3 July 2022
Accepted: 1 August 2022
Published: 4 August 2022

1. Introduction

Thanks to the development of earth observation-related techniques, acquiring massive remote sensing images for automatic image interpretation has become increasingly available. As one of the fundamental but challenging tasks in automatic image interpretation, object detection in remote sensing images not only requires recognizing the categories of objects in remote sensing images, but also determining their exact locations and sizes. Albeit the great significance of this task in several real-world applications [1–5], multiclass remote sensing object detection remains a challenging task and deserved further exploration, due to the complex scenes and diverse objects of remote sensing images.

Recently, data-driven deep learning techniques have achieved great success in several fields [6–11]. In particular, the area of object detection in natural scene images has also been significantly revolutionized. Overall, existing detectors can be divided into two-stage and one-stage detectors. Two-stage detectors [9,12–21] first generate candidate regions by using a region proposal algorithm [15,22]. Then, these candidate regions are further refined and classified by a region-wise network. However, one-stage detectors [23–28] use convolutional neural networks (CNNs) to extract the convolutional features of the input image and directly predict the locations and categories of objects on these feature maps in a single pass, and hence do not require a proposal generation procedure. In general,

two-stage detectors have higher detection accuracy while one-stage detectors have faster detection speed.

Among all existing two-stage object detectors, the pioneering Region-based Convolutional Neural Network (R-CNN) [12] first generates a set of region proposals via selective search [22], and then utilizes Alexnet [29] to extract the convolutional features of each proposal. Finally, Support Vector Machines (SVM) [30] is applied to these features to recognize the object categories within each proposal. Albeit excellent detection performance, this method is time-consuming as it repeatedly applies Alexnet on a large number of overlapped proposals per image without shared computation. Motivated by this, Fast R-CNN [14] first feeds the whole image into CNNs only once to extract a feature map. Then, a Region of Interest (RoI) pooling layer is used to map each proposal on this feature map and extract a fixed-length feature vector for each proposal. Finally, each feature vector is passed through fully connected (FC) layers to perform object classification and bounding box regression. Fast R-CNN has greatly sped up the object detection efficiency but still relies on external methods, such as selective search to generate region proposals, which has become the bottleneck in improving detection efficiency. Therefore, Faster R-CNN [15] is proposed to use a region proposal network (RPN) to replace the selective search in Fast R-CNN for generating high-quality proposals, which simultaneously improves both efficiency and accuracy. However, since Faster R-CNN only utilizes the top layer feature map from CNNs for object detection and the top layer feature map has a fixed shape and single-scale receptive field, Faster R-CNN can hardly detect multi-scale objects well. To this end, Feature Pyramid Network (FPN) [9] is further introduced with a top-down path and lateral connections to construct a feature pyramid. The shallow and deep layers in the pyramid are responsible for detecting small and large objects, respectively. Mask R-CNN [16] introduces a segment branch based on FPN and simultaneously performs object detection and instance segmentation.

Compared with the two-stage object detectors, one-stage detectors consider object detection as a regression problem. For example, You Only Look Once (YOLO) [24] is proposed to apply CNNs on the whole image to extract its multi-scale feature maps and directly predict the bounding boxes and categories of objects in each position of the top layer feature map. Different from YOLO, which only utilizes the top layer feature map for object detection, single shot multibox detector (SSD) [23] sets default boxes on the multi-scale feature maps extracted by CNNs and employs the shallow and deep layers to detect small and large objects, respectively. However, only a few default boxes contain objects, which may cause the easy background samples to dominate the training process. Therefore, RetinaNet [27] is proposed with a focal loss to make the detector focus more attention on the hard samples that are difficult to classify during training.

On the other hand, thanks to the publicly available remote sensing datasets, such as DIOR [31], DOTA [32], FAIR1M [33], and RSOD [34], a number of detectors have been proposed for multi-scale remote sensing object detection [35–39]. In particular, most of these methods are two-stage object detectors [31] and employ FPN as the feature extractor. The details of FPN are shown in Figure 1. Specifically, given an input image, FPN first utilizes CNNs, such as ResNet-50, to extract its multi-scale convolutional features, then employs nearest neighbor interpolation and lateral connections (1×1 convolutional layers (Convs)) to fuse the adjacent feature maps extracted by CNNs and construct a feature pyramid. Finally, each feature map in the feature pyramid is responsible for detecting objects in a specific range of sizes. As a feature extractor, FPN has greatly improved the object detection accuracy and been widely used for multi-scale object detection.

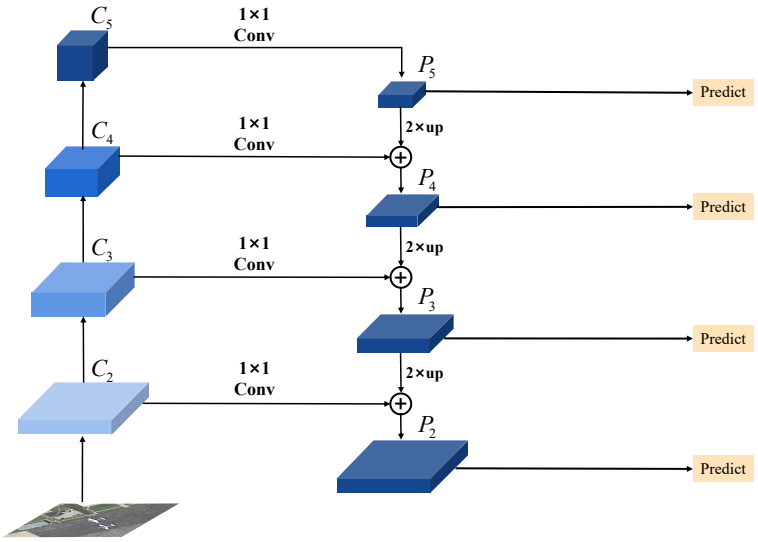

Figure 1. Flowchart of FPN.

Considering the objects in optical remote sensing images have variable shapes, orientations, and sizes, directly applying FPN to remote sensing object detection still faces the following two challenges. First, since the feature maps extracted by the FPN have fixed-shape receptive fields, as indicated in Figure 1, the original FPN, which only adopts 1×1 Convs as lateral connections to fuse these feature maps, can hardly generate precise feature maps with deformable fields to detect remote sensing objects with variable shapes and orientations. Second, remote sensing objects usually have large variations in size, and it is critically important to integrate the multi-level outputs of FPN to effectively detect these objects [18,39–41]. However, each proposal in the original FPN is predicted on a single feature level, whereas the useful information in other feature levels is ignored, which may harm its detection performance for multi-scale objects. To alleviate the above problem, several methods have been proposed to integrate the multi-level outputs of FPN. For instance, to enhance the current level features, the Cross-Scale Feature Fusion (CSFF) method [42] first concatenates the features of the remaining levels along the channel axis, and then adds the concatenated features with the current level features. Libra R-CNN [18] first integrates the multi-level outputs of FPN by using element-wise summation, then in turn adds the integrated features with the multi-level outputs of FPN to generate more discriminative pyramidal features for object detection. Although these two methods have effectively improved object detection accuracy, they still suffer from the following drawback. As different parts of the multi-level outputs of FPN show different significance to object detection, directly integrating them using concatenation or element-wise summation may not be able to generate optimal feature maps for detecting multi-scale remote sensing objects well.

To this end, we propose a new object detection framework based on FPN in this paper; the pipeline of our method can be seen in Figure 2. Compared with the original FPN (see Figure 1), the proposed method introduces two novel modules, i.e., Deformable Convolution Lateral Connection Module (DCLCM) and Attention-based Multi-Level Feature Fusion Module (A-MLFFM) to alleviate the above two problems, respectively. Specifically, we first replace the original lateral connections of FPN with the proposed DCLCMs. Each DCLCM includes a 3×3 deformable convolution to generate feature maps having deformable receptive fields for detecting remote sensing objects with various shapes and orientations better. Second, we propose several A-MLFFMs, each of which introduces a novel attention module to adaptively integrate the multi-level outputs of FPN. In this way, multi-level refined features $A_2 - A_5$ can be generated for multi-scale object detection. Specifically, A-MLFFM concatenates the features of the current layer with the features of the remaining layers to fully exploit the complementary information contained in different layers of FPN

and enhance the current layer features. Different from previous methods [42–44], which directly resize and concatenate multi-scale feature maps, in each A-MLFFM, a novel attention module is proposed and deployed after concatenation to emphasize the important features of the concatenated feature map.

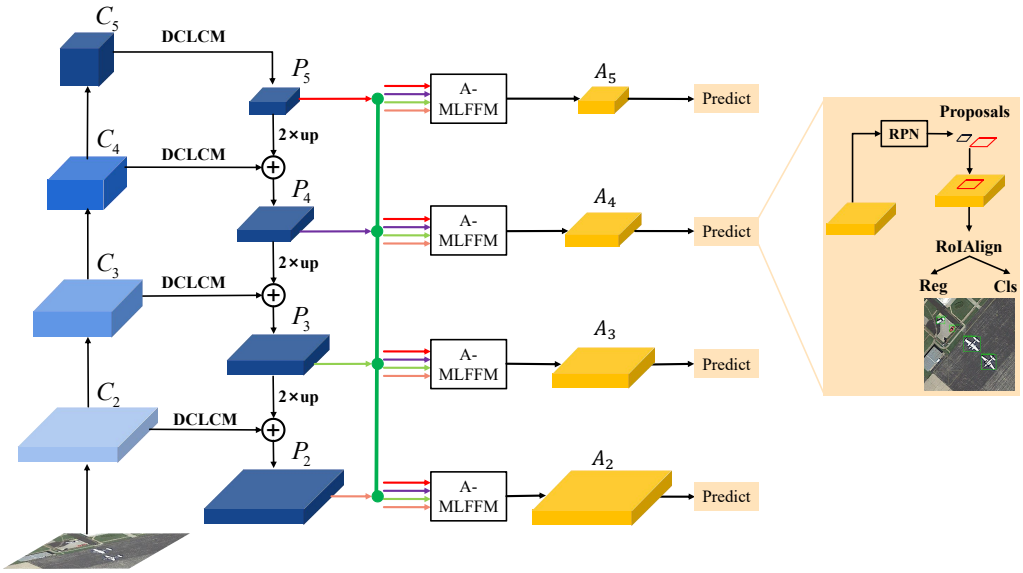

Figure 2. Flowchart of the proposed method.

The main contributions of this letter are summarized as follows.

(1) We replace the original lateral connections in FPN (1 × 1 Convs) with the proposed DCLCM, which aims to generate feature maps having deformable receptive fields, so as to effectively detect remote sensing objects with various shapes and orientations.
(2) Several A-MLFFMs, each of which contains a novel attention module, are proposed to adaptively fuse multi-level features and generate more powerful pyramidal features for object detection.
(3) Experimental results on the DIOR dataset validate the state-of-the-art performance of the proposed method.

The rest of this paper is structured as follows: in Section 2, we briefly introduce the literature related to our work, and the details of the proposed method are described in Section 3, including the DCLCM and A-MLFFM. In Section 4, we describe the experimental setup and implementation details, and the experimental results are reported in Section 5. Finally, conclusions are summarized in Section 5.

2. Related Work

In this section, we will briefly introduce the literature related to our work, including the studies on object detection in remote sensing images, multi-level feature fusion, and attention mechanism.

2.1. Object Detection in Remote Sensing Images

Object detection in remote sensing images has been studied for decades. Earlier works typically use handcrafted features to detect remote sensing objects, producing very limited detection performance. In addition, most of them are designed to detect a single class of objects, and often fail to deal with objects with cluttered background.

Inspired by the great success of CNNs-based object detection methods in natural scene images, extensive studies have been devoted to object detection in optical remote sensing images recently. For instance, SCRDet [45] uses the supervised pixel attention network and channel attention network to suppress noise and highlight target features, thus improving the detection performance of small and cluttered objects. Yang et al. [46] propose a Dense

Feature Pyramid Network (DFPN), which builds high-level semantic feature maps for all scales by means of dense connections to enhance the feature propagation and feature reuse. Yao et al. [47] design a unified EssNet backbone, which applies dilated convolution to maintain the resolution of deep level features, and then generate high-quality feature maps for detecting multi-scale objects. Dong et al. [35] develop a Receptive Field Expansion Block (RFEB) and add it on the top of the backbone of FPN to expand the receptive field of the whole network adaptively. In this way, the context information surrounding each object can be captured to help object detection. Qian et al. [48] propose a Multi-Level Feature Fusion (MLFF) module, which concatenates the feature grids pooled from the multi-level outputs of FPN to handle the problem of multi-scale object detection in aerial images. Lu et al. [49] design a multi-layer feature fusion structure to enhance the semantic information of the shallow features for detecting small objects better. Dong et al. [39] aggregate the multi-level outputs of FPN and the global context of the whole image by averaging them, and then in turn add the aggregated features with the multi-level outputs of FPN to generate more discriminative pyramidal features for remote sensing object detection.

In addition, some methods have been proposed to detect objects in wide area motion imagery in recent years. For example, taking a set of extremely large video frames as input, ClusterNet [50] combines the motion and appearance information within the convolutional architecture to locate multiple objects simultaneously. Zhou et al. [51] propose a new object detector, which first uses background subtraction with a low threshold to identify a large number of potential detections, then uses two CNNs considering both spatial-temporal information to remove false alarms and disentangle merged detections, respectively.

2.2. Multi-Level Feature Fusion

Remote sensing objects usually have large-scale variations, earlier object detectors, such as region-based CNN (R-CNN) [12], Fast R-CNN [14], Faster R-CNN [15], and Region-based Fully Convolutional Network (R-FCN) [52], for which only utilizing the top layer feature map of CNNs for object detection is suboptimal. To alleviate this issue, a lot of detectors have been proposed to exploit the multi-level outputs of CNNs for object detection. For instance, a single shot multibox detector (SSD) [23] directly utilizes the shallow and deep layer features of CNNs to detect small and large objects, respectively. However, the shallow layer features of CNNs lack semantic information, which may decrease its recognition performance for small objects. To this end, FPN [9] introduces a top-down path and lateral connections to pass the semantic information from deep layers to shallow layers, and construct a feature pyramid. The shallow and deep layers in the pyramid are responsible for detecting small and large objects, respectively. FPN has greatly increased the detection accuracy and become the mainstream way for multi-scale object detection. Based on FPN, a lot of detectors have been proposed. For example, an additional bottom-up path is created in PA-Net [40] to shorten the information path between the shallow layers and deep layers, and thus the accurate localization information existed in shallow layers can be passed to deep layers. NAS-FPN [53] adopts Neural Architecture Search (NAS) [54] to find the optimal feature pyramid architecture automatically. In addition, EfficientDet [19] finds that PANet suffers from too many parameters and computational load, and prunes PA-Net [40] to improve the model efficiency. In the field of remote sensing object detection, in order to fully exploit the multi-level outputs of FPN, and enhance the current level features, the CSFF method [42] first concatenates the features of the remaining levels along the channel axis, and then adds the concatenated features with the current level features.

In the aforementioned models, when fusing multi-scale features with different resolutions, most models first resize them to the same resolution, and then sum or concatenate them. In this paper, referring to the previous methods [42–44], we use concatenation to fuse the multi-level outputs of FPN, and make two improvements to fuse them better. First, since different parts of the multi-level outputs of FPN show different importance for object detection, a novel attention module is proposed and deployed after concatenation to emphasize the important features of the concatenated feature map. In this way, we can adaptively

exploit the complementary information contained in different outputs of FPN. Second, since the objects in remote sensing images have variable shapes and orientations, the original outputs of FPN, which have fixed-shape receptive fields, can hardly handle these objects well. To this end, we replace the original lateral connections of FPN with DCLCMs, each of which contains a 3×3 deformable convolution to generate feature maps with deformable receptive fields before concatenating them. More details are described in Section 3.

2.3. Attention Mechanism

In recent years, the attention mechanism has been widely used in various computer vision tasks, and achieved promising results [39,55–61]. Among all the existing attention modules, Squeeze-and-Excitation (SE) attention [62] is the most popular one. It computes channel attention by using 2D global average pooling. However, 2D global average pooling may cause SE attention loss of the positional information of objects, which is critically important for remote sensing object detection. To alleviate the positional information loss, the coordinate attention block [63] exploits two 1D global average pooling to pool the input feature map along the horizontal and vertical directions, generating two feature maps, respectively. Then, these two feature maps are encoded into two attention maps by using 1×1 convolution layers. Finally, both attention maps are multiplied with the input feature map to emphasize the features of targets. The positional information of targets can thus be preserved by the coordinate attention. However, as described above, coordinate attention computes the importance of each row or each column of the input feature map only using the information in the corresponding row or column, and the contextual information surrounding the corresponding row or column is not exploited, which is suboptimal for object detection. Based on coordinate attention, this paper proposes a novel attention module, which computes the importance of each row or each column of the input feature map using both the information in the corresponding row or column and the information surrounding the corresponding row or column; thus, it can produce more precise attention maps for object detection.

3. Materials and Methods

3.1. Overview of the Proposed Method

The framework of the proposed method is summarized in Figure 2. Specifically, given an input image, we first use ResNet-50 [6] to extract its multi-scale convolutional features C_2-C_5. Then, we replace the original lateral connections of the FPN with the proposed DCLCMs, and obtain new pyramidal features P_2-P_5 following the operation of FPN. To make full use of P_2-P_5 for object detection and detect multi-scale objects better, we propose several A-MLFFMs, each of which adaptively concatenates the features of the current layer with the features of the remaining layers to enhance the current layer features. In this way, the complementary information contained in different outputs of FPN is fully exploited, and the corresponding generated multi-scale feature maps A_2-A_5 are more conducive to multi-scale remote sensing object detection. In the following, we will elaborate the details of DCLCM and A-MLFFM.

3.2. Deformable Convolution Lateral Connection Module (DCLCM)

Given an input feature map x, the standard 2D convolutional process consists of two steps—first, sampling over the input feature map using a regular grid R (for a 3×3 standard convolution, $R = \{(-1, -1), (-1, 0), \ldots, (1, 0), (1, 1)\}$); second, summing the sampled values weighted by w.

Specifically, for each location p on the output feature map y, the output value $y(p)$ is computed as follows:

$$y(p) = \sum_{p_n \in R} (w(p_n) \cdot x(p + p_n)) \tag{1}$$

where p_n enumerates the locations in R. $x(p + p_n)$ represents the input of convolution when the sampling location is $p + p_n$.

In deformable convolution (DConv), as shown in Figure 3, the regular grid R is augmented with offsets, which are learned from the input feature map. This paper uses $\{\Delta p_n | n = 1, 2, ..., N\}$, $N = |R|$ to represent the learned offsets, and Equation (1) is thus modified as follows for DConv:

$$y(p) = \sum_{p_n \in R} (w(p_n) \cdot x(p + p_n + \Delta p_n)) \quad (2)$$

Due to the irregular offsets, DConv can adaptively adjust the sampling locations according to the input feature map and output a feature map with deformable receptive fields.

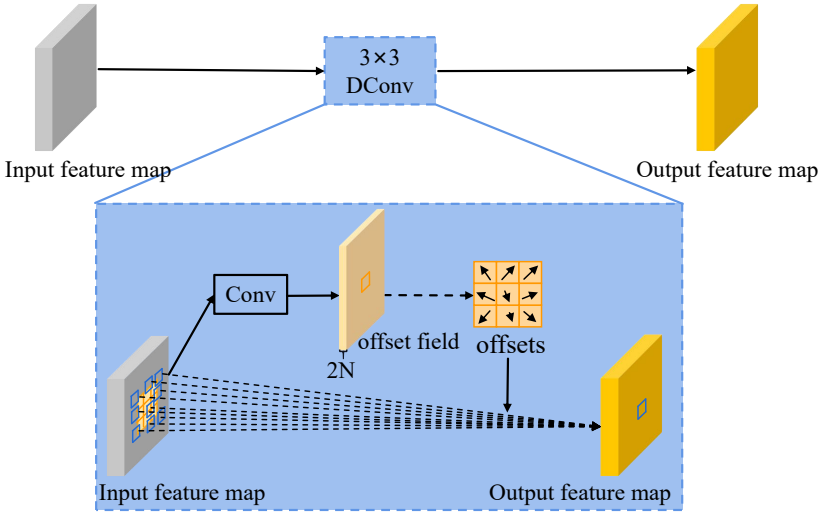

Figure 3. Flowchart of DCLCM.

Since the feature maps extracted by the backbone of FPN have fixed-shape receptive fields, the original FPN, which only adopts 1×1 convolution layers as the lateral connections, can hardly generate precise feature maps with deformable fields to detect remote sensing objects with variable shapes and orientations. To this end, as shown in Figure 2, we replace the original lateral connections of FPN with DCLCMs, each of which includes a 3×3 DConv to generate new pyramidal features for detecting remote sensing objects better.

3.3. Attention-Based Multi-Level Feature Fusion Module (A-MLFFM)

A-MLFFMs take $P_2 - P_5$ as inputs and generate more powerful pyramidal representations $A_2 - A_5$ for the subsequent object detection. Figure 4 illustrates the details of the proposed A-MLFFM for generating A_3. Likewise, A_2, A_4, and A_5 can be generated in a similar way. As shown in Figure 4, when generating A_3, A-MLFFM regards P_3 as the base feature map, while regarding the other three feature maps, i.e., P_2, P_4, and P_5 as the auxiliary feature maps. The base feature map is used to construct the primary features of A_3, while the auxiliary feature maps offer complementary features to remedy the inadequacy of P_3 for object detection. Specifically, the proposed A-MLFFM first introduces four 1×1 convolutional layers (Convs) and applies them after the multi-level outputs of FPN to generate four intermediate feature maps, namely $P_2^{'} - P_5^{'}$. During this process, for the other three auxiliary feature maps, the output channels of 1×1 Conv are set to 128. For the base feature map, the output channels of 1×1 Conv are set to 256. In addition, to save computational cost and memory usage, if the input level is higher than the output one (e.g., P_4 and P_5), the input feature map is processed by the operations of Conv followed by bilinear interpolation upsampling. Conversely, for lower input level (e.g., P_2), the input feature map is first downsampled by average pooling and then passed to 1×1 Conv.

Then, $P_2^{'} - P_5^{'}$ with the same spatial resolution are concatenated along the channel axis to generate a multi-scale concatenated feature map named F_{con}. Since different parts of F_{con} show different importance to object detection, a novel attention module is added after F_{con} to weigh it. The attention module will assign larger weights to the features that are important to object detection, making A-MLFFM able to adaptively exploit the complementary information contained in different parts of F_{con}.

Finally, a 3×3 Conv is used to adjust the channels of the refined feature map from 640 to 256 for the final object detection.

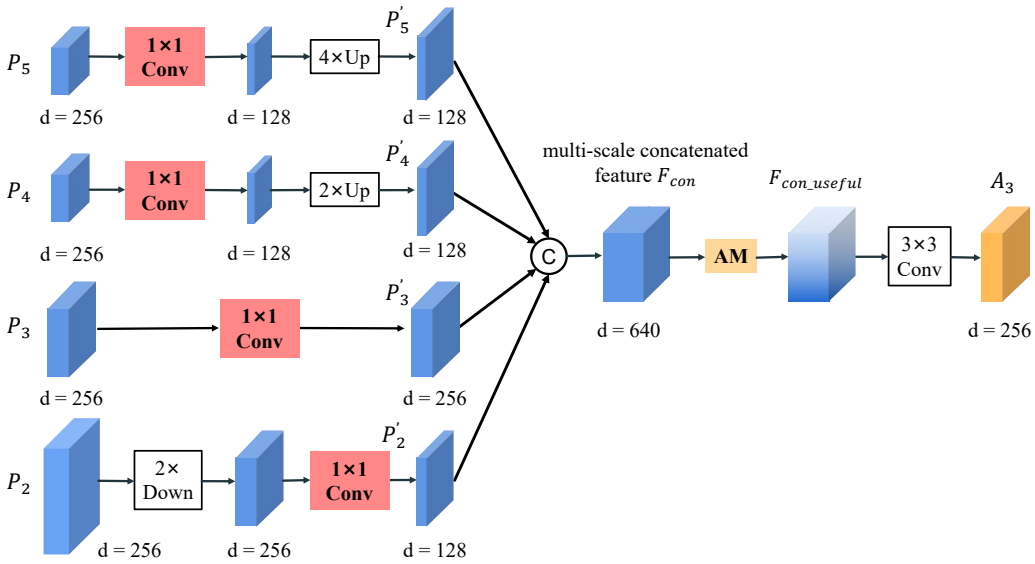

Figure 4. Structure of the proposed A-MLFFM, taking the generation of A_3 as an example.

In short, the overall process of A-MLFFM for generating A_3 is as follows:

$$P_2^{'} = Conv_{1\times1}(Down(P_2)) \tag{3}$$

$$P_3^{'} = Conv_{1\times1}(P_3) \tag{4}$$

$$P_4^{'} = Up(Conv_{1\times1}(P_4)) \tag{5}$$

$$P_5^{'} = Up(Conv_{1\times1}(P_5)) \tag{6}$$

$$F_{con} = Concat(P_2^{'}, P_3^{'}, P_4^{'}, P_5^{'}) \tag{7}$$

$$F_{con\_useful} = AM(F_{con}) \tag{8}$$

$$A_3 = Conv_{3\times3}(F_{con\_useful}) \tag{9}$$

where $Conv$ represents standard convolution. The subscripts 3×3 and 1×1 denote the kernel size. Up, $Down$, and $Concat$ denote the operations of bilinear interpolation, average pooling, and concatenation, respectively. AM is the proposed attention module.

In the following, we will describe the structure of the proposed attention module.

The Proposed Attention Module (AM)

The function of attention module here is to highlight the significant features of F_{con}. Among all the existing attention modules [55–63], Squeeze-and-Excitation (SE) attention [62] is the most popular one. Specifically, as shown in Figure 5a, given an input feature map, SE attention first uses a 2D global average pooling to aggregate the input feature map into a channel descriptor. Then, the channel descriptor is passed to two fully connected layers to compute the importance of each channel in the input feature map. However, 2D global average pooling may cause SE attention loss of the positional information of objects, which is critically important for remote sensing object detection. To alleviate the positional information

loss, the coordinate attention module [63] (see Figure 5b) uses two pooling layers with kernel size of (1, W) and (H, 1) to pool the input feature map along the horizontal and vertical directions, generating two feature maps, respectively. Then, these two feature maps are encoded into two attention maps by using 1×1 convolution layers. Finally, both attention maps are multiplied with the input feature map to emphasize the features of targets. The positional information of targets can thus be preserved by the coordinate attention. However, as described above, coordinate attention computes the importance of each row or each column of the input feature map only using the information in the corresponding row or column. The contextual information surrounding the corresponding row or column is not exploited, which is suboptimal for object detection. Based on coordinate attention, a novel attention module is proposed to produce more precise attention maps.

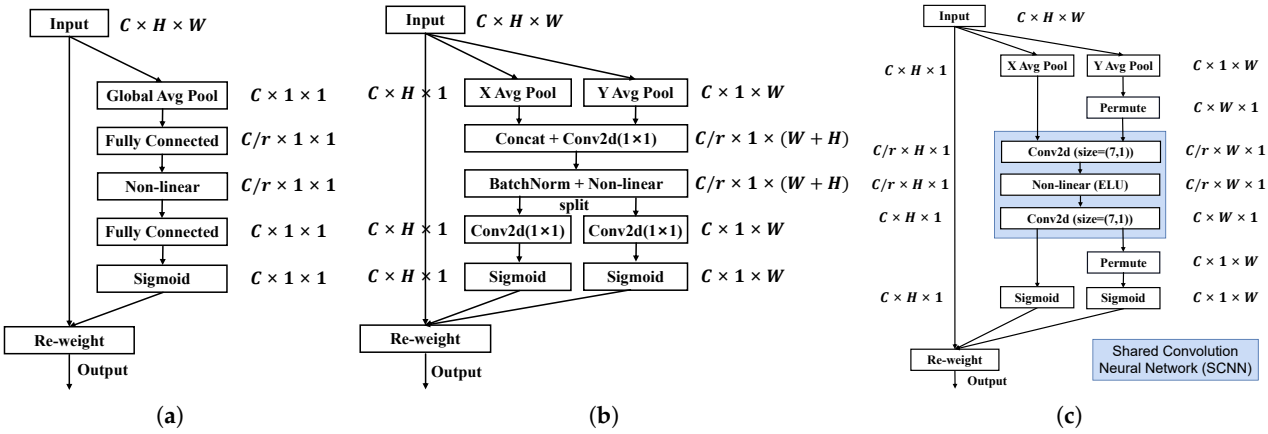

Figure 5. Schematic comparison of the classic SE attention (**a**), coordinate attention (**b**), and the proposed attention module (**c**). Here, 'X Avg Pool' and 'Y Avg Pool' represent two pooling layers with kernel size of (1, W) and (H, 1), respectively.

Specifically, as shown in Figure 5c, given the input feature map $F_{con} \in \mathbb{R}^{C*H*W}$, the proposed attention module first exploits two pooling layers with kernel size of (1, W) and (H, 1) to aggregate F_{con} into two feature maps. Then, unlike coordinate attention that sends these two feature maps to 1×1 convolutional layers to generate two corresponding attention maps, the proposed attention module sends them to a shared convolutional network, which contains two 7×1 convolutional layers to produce the above two attention maps. In this manner, the proposed attention module computes the importance of each row or each column of the input feature map using both the information in the corresponding row or column and the information surrounding the corresponding row or column and thus can produce more precise attention maps. In addition, different from the SE attention and coordinate attention, we use exponential linear unit (ELU) [64] instead of rectified linear unit (ReLU) between two fully connected layers to facilitate gradient propagation. Finally, following coordinate attention, both attention maps are multiplied with the input feature map to emphasize the features of targets.

4. Experiments and Results

4.1. Dataset

In this paper, the proposed method is evaluated on the DIOR dataset, which is one of the largest and most diverse datasets for object detection in remote sensing images. The DIOR dataset contained 23,463 images, and each image has the spatial size of 800×800 pixels. The training and testing datasets contain 11,725 and 11,738 images, respectively. In addition, 192,472 instances are annotated with horizontal rectangles in DIOR. The instances in DIOR have variable shapes, orientations, and sizes. Figure 6 shows the 20 object categories in DIOR.

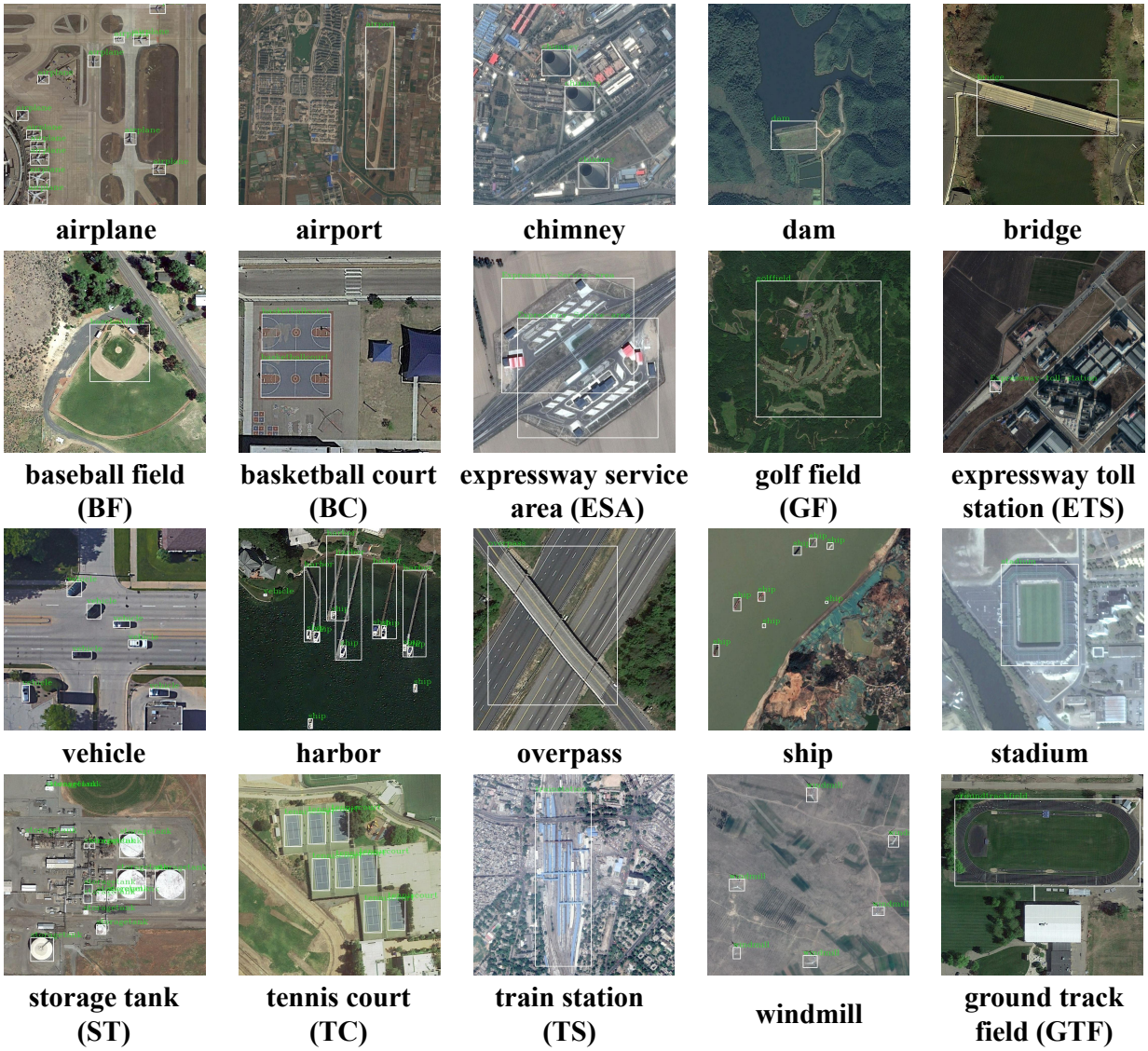

Figure 6. Object categories in the DIOR dataset. The white rectangles represent the ground truths in each image.

4.2. Evaluation Metrics

In object detection, each predicted bounding box can be divided into two types, i.e., true positive (TP) and false positive (FP). Specifically, if the Intersection over Union (IoU) between a predicted box and a ground truth exceeds a given threshold (set as 0.5 in this paper), the predicted box is recognized as TP. FP is just the opposite. In addition, the ground truth, which is not matched with any predicted bounding box is recognized as false negative (FN). Set N_{TP}, N_{FP} and N_{FN} as the number of TPs, FPs, and FNs, and the precision P and recall R can be obtained by the following equations:

$$P = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} = \frac{N_{\mathrm{TP}}}{\text{all detections}} \tag{10}$$

$$R = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} = \frac{N_{\mathrm{TP}}}{\text{all ground truths}} \tag{11}$$

Average Precision (AP) computes the mean value of P for recall value from $R = 0$ to $R = 1$ and higher AP indicates better detection performance. Thus, for multi-class object detection, mean AP (mAP), which is the mean value of AP over all classes, is adopted to

evaluate the performance of different detectors. In addition, we adopt the PASCAL VOC 07 metric [65] to calculate the AP of each category in this paper.

4.3. Implementation Details

In this paper, ResNet-50 [6] pretrained on the ImageNet [66] is chosen as the backbone for convolutional feature extraction. All detectors are trained on four NVIDIA TITAN X GPUs, and we adopt stochastic gradient descent algorithm to optimize the parameters of each detector. The number of total epochs and the initial learning rate are set to 12 and 0.02, respectively, and at the 8th and 11th epochs, the learning rate is divided by 10. In addition, in order to avoid the instability of convergence at the beginning of training, the warm-up learning rate is used (the learning rate is increased from 2×10^{-5} to 2×10^{-2} linearly during the first 800 iterations). The batch size is set to 12 (three images per GPU, and each image is resized to the size of 800×800 pixels).

4.4. Ablation Study and Analysis

The proposed method contains two modules, i.e., DCLCM and A-MLFFM. To demonstrate their individual effectiveness, an ablation study is conducted over the DIOR dataset. As shown in Table 1, we compare the detection mAP of FPN and FPN with different combinations of DCLCM and A-MLFFM on DIOR.

Table 1. Detection accuracy comparisons between FPN and FPN with different combinations of DCLCM and A-MLFFM on the DIOR Dataset.

Category	FPN [9]	FPN + DCLCM	FPN + A-MLFFM	FPN + DCLCM + A-MLFFM
airplane	70.2	70.6	70.3	70.7
airport	75.2	83.0	79.2	83.1
BF	71.5	71.6	71.8	71.9
BC	86.1	86.6	86.8	86.5
bridge	46.8	49.0	50.5	49.3
chimney	76.9	79.0	78.0	78.2
dam	64.4	68.8	69.2	70.3
ESA	76.4	78.8	82.0	83.7
ETS	69.7	73.4	74.3	76.7
GF	75.3	76.6	75.7	76.0
GTF	79.6	80.6	78.6	80.2
harbor	56.6	57.5	52.8	55.9
overpass	60.8	62.2	61.3	62.7
ship	88.5	88.7	88.9	89.0
stadium	60.7	69.8	69.5	71.3
ST	70.6	71.4	78.0	79.1
TC	81.4	81.3	81.2	81.4
TS	54.9	61.2	57.0	60.1
vehicle	54.4	54.8	55.5	55.6
windmill	88.5	88.2	88.6	89.4
mAP	70.4	72.7	72.5	73.6

- **DCLCM Only.** After we replace the original lateral connections of FPN with the proposed DCLCMs, 2.3% mAP improvement can be achieved on the DIOR dataset. The objects in remote sensing images have variable shapes and orientations, which may seriously affect the detection performance of FPN. By integrating the DCLCM into FPN, more precise feature maps with deformable receptive fields can be generated to detect these remote sensing objects well.
- **A-MLFFM Only.** When we insert the proposed A-MLFFMs into FPN, the detection mAP is improved by 72.1% on the DIOR dataset. Each A-MLFFM contains a novel attention module to fuse the multi-level outputs of FPN adaptively. In this way,

the complementary information contained in different outputs of FPN can be fully exploited to detect remote sensing objects with various sizes better.

- **Both DCLCM and A-MLFFM.** By introducing both of the proposed DCLCM and A-MLFFM into FPN simultaneously, the detection mAP can be further increased to 73.6% on the DIOR dataset, which demonstrates that DCLCM and A-MLFFM are two complementary modules, and the utilization of them allows the proposed method to achieve the best detection accuracy.

Qualitatively, as illustrated in Figure 7, compared with the baseline method FPN, the proposed method can reduce the number of false and missing detections, as well as locate remote sensing objects more precisely.

(**a**) (**b**)

Figure 7. Some detection results comparison between FPN and the proposed method. (**a**) prediction of FPN; (**b**) prediction of the proposed method. The green, blue, and red rectangles represent TPs, FPs, and FNs, respectively.

4.5. Effectiveness of the Attention Module in A-MLFFM

To clearly verify the effectiveness of the attention module in A-MLFFM, we conduct comparative experiments on the DIOR dataset. Specifically, we first remove the attention module from A-MLFFM, and obtain a new module named multi-level feature fusion module (MLFFM). Then, we add different attention modules in the proposed A-MLFFM. As shown in Table 2, compared to FPN with MLFFM, after we add different attention modules in A-MLFFM, the detection mAP can be increased, which demonstrate the effectiveness of the attention module in A-MLFFM. In addition, FPN with A-MLFFM, which contains the proposed attention module, can achieve the best detection mAP. This demonstrates the superiority of the proposed attention compared with SE attention and coordinate attention module.

Table 2. Detection accuracy comparison between FPN with MLFFM and FPN with A-MLFFM containing different attention modules on the DIOR Dataset.

Category	FPN + MLFFM	FPN + A-MLFFM (SE Attention [62])	FPN + A-MLFFM (CA Block [63])	FPN + A-MLFFM (the Proposed Attention)
airplane	70.4	70.2	70.3	70.3
airport	75.8	78.3	78.5	79.2
BF	71.4	71.7	71.8	71.8
BC	86.3	86.1	86.4	86.8
bridge	48.2	48.2	48.6	50.5
chimney	78.1	78.0	77.5	78.0
dam	65.0	66.0	66.5	69.2
ESA	79.4	80.6	78.8	82.0
ETS	69.8	73.9	73.3	74.3
GF	76.3	76.4	76.4	75.7
GTF	80.5	78.6	80.9	78.6
harbor	54.7	53.3	53.3	52.8
overpass	61.5	61.8	61.6	61.3
ship	88.9	88.5	88.7	88.9
stadium	71.3	69.7	71.6	69.5
ST	71.0	71.3	71.4	78.0
TC	81.4	81.4	81.5	81.2
TS	55.1	56.4	55.6	57.0
vehicle	55.0	55.2	55.3	55.5
windmill	88.6	89.0	88.6	88.6
mAP	71.4	71.7	71.8	72.5

4.6. Comparison with Other Multi-Level Feature Fusion Modules

In order to reveal the superiority of the proposed A-MLFFM, we compare it with three representative multi-level feature fusion modules, including the balanced feature pyramid (BFP) [18], cross-scale feature fusion (CSFF) [42], and adaptive feature pooling (AFP) [40] on the DIOR dataset. Specifically, we add them after the multi-level outputs of FPN. As can be seen from Table 3, compared with the other three methods, FPN with the proposed A-MLFFM can achieve the highest mAP of 72.5% on DIOR.

4.7. Comparison with Other State-of-the-Art Methods

To further evaluate the performance of the proposed method, we quantitatively compare it with seven considered state-of-the-art object detection methods, encompassing the original FPN [9], Libra R-CNN [18], Remote-sensing Spatial Adaptation DETector (RSADet) [67], deformable convolution networks (DCNs) [68], Double-head R-CNN [17], adaptive balanced network (ABNet) [69], and FPN with receptive field expansion block [35]. The detection accuracy of all these methods is illustrated in Table 4, from which we can see that the proposed method achieves the highest detection mAP of 73.6% on the DIOR dataset. Some visualization results of proposed method are shown in Figure 8. It can

be observed that, even remote sensing objects have different degrees of scale changes and deformation, the proposed method can detect them well.

Table 3. Detection accuracy comparisons of FPN with different multi-level feature fusion modules on the DIOR dataset.

Category	FPN + CSFF [42]	FPN + AFP [40]	FPN + BFP [18]	FPN + A-MLFFM
airplane	69.9	69.6	70.5	70.3
airport	74.4	75.4	78.1	79.2
BF	71.6	71.7	71.7	71.8
BC	86.3	86.4	86.3	86.8
bridge	47.0	47.1	47.4	50.5
chimney	78.1	77.8	77.9	78.0
dam	63.8	66.3	64.8	69.2
ESA	76.3	76.3	75.9	82.0
ETS	73.0	70.1	72.8	74.3
GF	75.4	74.3	75.3	75.7
GTF	80.8	80.5	80.7	78.6
harbor	53.9	52.8	55.9	52.8
overpass	61.0	61.1	60.3	61.3
ship	88.6	88.7	88.8	88.9
stadium	71.5	72.4	69.9	69.5
ST	71.1	71.0	70.8	78.0
TC	81.4	81.4	81.5	81.2
TS	55.1	56.6	56.3	57.0
vehicle	54.5	54.7	54.8	55.5
windmill	88.7	88.3	88.6	88.6
mAP	71.1	71.1	71.4	72.5

Table 4. Detection accuracy comparisons of different methods on the DIOR dataset. * indicates our implementation.

Category	FPN * [9]	Libra R-CNN * [18]	RSADet [67]	DCNs * [68]	Double-Head R-CNN * [17]	ABNet [69]	FPN with RFEB [35]	Ours
airplane	70.2	70.9	73.6	70.2	70.4	66.8	70.0	70.9
arport	75.2	76.5	86.0	81.7	80.2	84.0	83.9	83.1
BF	71.5	71.4	72.6	71.7	71.9	74.9	77.1	71.9
BC	86.1	86.0	89.6	86.2	86.9	87.7	86.9	86.5
bridge	46.8	47.4	43.6	48.0	49.3	50.3	47.8	49.3
chimney	76.9	77.4	75.3	75.7	78.5	78.2	77.6	78.2
dam	64.4	66.0	62.3	67.1	65.9	67.8	71.3	70.3
ESA	76.4	76.9	79.5	78.8	77.6	85.9	79.5	83.7
ETS	69.7	73.6	68.7	73.5	73.2	74.2	70.6	76.7
GF	75.3	76.1	78.6	75.5	76.2	79.7	77.3	76.0
GTF	79.6	80.4	79.1	79.6	81.5	81.2	78.7	80.2
harbor	56.6	55.9	57.9	57.2	56.2	55.4	57.7	55.9
overpass	60.8	61.2	59.2	61.2	61.8	61.6	60.4	62.7
ship	88.5	89.0	90.0	88.4	88.8	75.1	88.4	89.0
stadium	60.7	70.0	55.8	69.4	73.6	74.0	67.7	71.3
ST	70.6	71.2	77.0	77.1	71.3	66.7	76.2	79.1
TC	81.4	88.0	87.8	81.3	87.9	87.0	86.9	81.4
TS	54.9	55.3	65.3	60.4	58.7	62.2	63.7	60.1
vehicle	54.4	54.6	55.3	54.7	54.9	53.6	54.5	55.6
windmill	88.5	88.7	86.5	88.9	89.0	89.1	88.2	89.4
mAP	70.4	71.8	72.2	72.3	72.7	72.8	73.2	73.6

Figure 8. Some detection results of the proposed method on the DIOR dataset. The green rectangles represent TPs.

5. Conclusions

This paper proposes a novel object detection framework based on FPN to improve the object detection accuracy in remote sensing images. First, we replace the original lateral connections of FPN with DCLCMs to generate feature maps with deformable receptive fields for detecting remote sensing objects with variable shapes and orientations better. Second, taking these newly generated feature maps as input, A-MLFFMs generate more powerful pyramidal features by fusing them adaptively. In this way, the complementary information contained in different outputs of FPN can be fully exploited. The experiment results on the DIOR dataset indicate that, compared with the considered state-of-the-art methods, the proposed method can achieve highest detection mAP of 73.6%.

The proposed method has a limitation when detecting oriented objects, and we will expand its framework in the future to achieve the oriented object detection. In addition, we will consider the computational load of deformable convolution in DCLCM and adopt the structure like a residual module to optimize the detection speed of the proposed method.

Author Contributions: X.D. and Y.Q. proposed the method and wrote the original draft. R.F. and Y.G. reviewed the manuscript and supervised the study. R.F., S.L. and Y.Y. designed and carried out the experiments. All authors have read and improved the manuscript.

Funding: This work was funded by the Hunan Provincial Innovation Foundation for Postgraduate (No. CX20190020) and the National Natural Science Foundation of China (No. 42101344 and No. 62001482).

Data Availability Statement: Publicly available dataset was used in this paper. This data can be found here: http://www.escience.cn/people/gongcheng/DIOR.html (accessed on 6 May 2021).

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

References

1. Ahmad, K.; Pogorelov, K.; Riegler, M.; Conci, N.; Halvorsen, P. Social media and satellites. *Multimed Tools Appl.* **2019**, *78*, 2837–2875. [CrossRef]
2. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
3. 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]
4. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
5. Chen, X.; Xiang, S.; Liu, C.L.; Pan, C.H. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1797–1801. [CrossRef]
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
7. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
8. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2472–2481.
9. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
10. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni N.; Markham, A. Learning semantic segmentation of large-scale point clouds with random sampling. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021, *in press.* [CrossRef]
11. Hu, Q.; Yang, B.; Khalid, S.; Xiao, W.; Trigoni, N.; Markham, A. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *Int. J. Comput. Vis.* **2022**, *130*, 316–343. [CrossRef]
12. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–27 June 2014; pp. 580–587.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]

14. Girshick, R. Fast r-cnn. In Proceedings of the 2015 IEEE/CVF International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, 39, 1137–1149. [CrossRef] [PubMed]
16. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the 2017 IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
17. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10186–10195.
18. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 821–830.
19. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
20. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
21. Qiao, S.; Chen, L.C.; Yuille, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10213–10224.
22. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
23. 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 2016 European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
25. Redmon, J.; Redmon, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
26. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
27. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
28. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
29. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012.
30. Malisiewicz, T.; Gupta, A.; Efros, A.A. Ensemble of exemplar-svms for object detection and beyond. In Proceedings of the 2011 IEEE/CVF International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 89–96.
31. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
32. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
33. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
34. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
35. Dong, X.; Fu, R.; Gao, Y.; Qin, Y.; Ye, Y.; Li, B. Remote Sensing Object Detection Based on Receptive Field Expansion Block. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
36. Wu, Y.; Zhang, K.; Wang, J.; Wang, Y.; Wang, Q.; Li, Q. CDD-net: A context-driven detection network for multiclass object detection. *IEEE Geosci. Remote Sens. Lett.* 2020, *in press*. [CrossRef]
37. Ye, Y.; Ren, X.; Zhu, B.; Tang, T.; Tan, X.; Gui, Y.; Yao, Q. An Adaptive Attention Fusion Mechanism Convolutional Network for Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 516. [CrossRef]
38. Chen, L.; Liu, C.; Chang, F.; Li, S.; Nie, Z. Adaptive multi-level feature fusion and attention-based network for arbitrary-oriented object detection in remote sensing imagery. *Neurocomputing* **2021**, *451*, 67–80. [CrossRef]
39. Dong, X.; Qin, Y.; Fu, R.; Gao, Y.; Liu, S.; Ye, Y.; Li, B. Multi-Scale Deformable Attention and Multi-Level Features Aggregation for Remote Sensing Object Detection. *IEEE Geosci. Remote Sens. Lett.* 2022, *in press*.
40. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.

41. 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 33th AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 9259–9266.
42. Cheng, G.; Si, Y.; Hong, H.; Yao, X.; Guo, L. Cross-scale feature fusion for object detection in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 431–435. [CrossRef]
43. Huang, W.; Li, G.; Chen, Q.; Ju, M.; Qu, J. CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection. *Remote Sens.* **2021**, *13*, 847. [CrossRef]
44. Zhai, S.; Shang, D.; Wang, S.; Dong, S. DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion. *IEEE Access* **2020**, *8*, 24344–24357. [CrossRef]
45. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8232–8241.
46. 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]
47. Yao, Q.; Hu, X.; Lei, H. Multiscale convolutional neural networks for geospatial object detection in VHR satellite images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 23–27. [CrossRef]
48. Qian, X.; Lin, S.; Cheng, G.; Yao, X.; Ren, H.; Wang, W. Object detection in remote sensing images based on improved bounding box regression and multi-level features fusion. *Remote Sens.* **2020**, *12*, 143. [CrossRef]
49. Lu, X.; Ji, J.; Xing, Z.; Miao, Q. Attention and feature fusion SSD for remote sensing object detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–9. [CrossRef]
50. LaLonde, R.; Zhang, D.; Shah, M. Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4003–4012.
51. Zhou, Y.; Maskell, S. Detecting and tracking small moving objects in wide area motion imagery (wami) using convolutional neural networks (cnns). In Proceedings of the 22th International Conference on Information Fusion, Ottawa, ON, Canada, 2–5 July 2019; pp. 1–8.
52. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
53. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7036–7045.
54. Elsken, T.; Metzen, J.H.; Hutter, F. Neural architecture search: A survey. *J. Mach. Learn. Res.* **2019**, *20*, 1997–2017.
55. Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q.V. Attention augmented convolutional networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 3286–3295.
56. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. In Proceedings of the 32th International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018.
57. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the 2016 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
58. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. A simple and light-weight attention module for convolutional neural networks. *Int. J. Comput. Vis.* **2020**, *128*, 783–798. [CrossRef]
59. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
60. Hou, Q.; Zhang, L.; Cheng, M.M.; Feng, J. Strip pooling: Rethinking spatial pooling for scene parsing. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4003–4012.
61. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
62. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
63. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
64. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* **2015**, arXiv:1511.07289.
65. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
66. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]

67. Yu, D.; Ji, S. A new spatial-oriented object detection framework for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [CrossRef]
68. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the 2017 IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 764–773.
69. Liu, Y.; Li, Q.; Yuan, Y.; Du, Q.; Wang, Q. ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]