electronics

MDPI

Article

FCC-Net: A Full-Coverage Collaborative Network for Weakly Supervised Remote Sensing Object Detection

Suting Chen [1,*], Dongwei Shao [1], Xiao Shu [2], Chuang Zhang [1] and Jun Wang [3]

[1] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science & Technology, Nanjing 210044, China; dongwei.shao@nuist.edu.cn (D.S.); zhch_76@nuist.edu.cn (C.Z.)

[2] Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada; shux@mcmaster.ca

[3] School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing 210044, China; wangjun@nuist.edu.cn

* Correspondence: sutingchen@nuist.edu.cn; Tel.: +86-139-1386-4015

Received: 25 July 2020; Accepted: 15 August 2020; Published: 21 August 2020

Abstract: With an ever-increasing resolution of optical remote-sensing images, how to extract information from these images efficiently and effectively has gradually become a challenging problem. As it is prohibitively expensive to label every object in these high-resolution images manually, there is only a small number of high-resolution images with detailed object labels available, highly insufficient for common machine learning-based object detection algorithms. Another challenge is the huge range of object sizes: it is difficult to locate large objects, such as buildings and small objects, such as vehicles, simultaneously. To tackle these problems, we propose a novel neural network based remote sensing object detector called full-coverage collaborative network (FCC-Net). The detector employs various tailored designs, such as hybrid dilated convolutions and multi-level pooling, to enhance multiscale feature extraction and improve its robustness in dealing with objects of different sizes. Moreover, by utilizing asynchronous iterative training alternating between strongly supervised and weakly supervised detectors, the proposed method only requires image-level ground truth labels for training. To evaluate the approach, we compare it against a few state-of-the-art techniques on two large-scale remote-sensing image benchmark sets. The experimental results show that FCC-Net significantly outperforms other weakly supervised methods in detection accuracy. Through a comprehensive ablation study, we also demonstrate the efficacy of the proposed dilated convolutions and multi-level pooling in increasing the scale invariance of an object detector.

Keywords: object detection; collaborative learning; multiscale feature representation; weakly supervised learning; remote-sensing image

1. Introduction

Remote sensing is an interdisciplinary subject involving technologies [1] such as aviation, optical instrument, electronic sensor and computer science, etc. Over the past few decades, the development of these individual disciplines has armed remote sensing with unprecedented tools that enable extremely high image resolution and fidelity. The ever-more-capable remote-sensing imaging has become an essential part of many important applications including high-precision map surveying [2], weather forecasting [3], land planning [4] and disaster prevention [5], etc. With the fast-growing size of remote-sensing image data, a pressing problem is to efficiently extract useful information from these high-resolution images. In general, information extraction for remote-sensing images contains a few components: scene classification, object recognition, object detection and semantic segmentation.

The focus of this study is on object detection, which is the process of locating and classifying ground objects in a remote-sensing image.

Object detection technology has been widely used in transportation, industry and medical treatment [6–9]. Most existing object detection algorithms consist of a feature extraction stage and then an object recognition stage. Conventional object detection techniques generally employ predefined feature descriptors in feature extraction. After extracting the features in each sliding window, these techniques usually use classifiers, such as support vector machine (SVM) to determine the categories of objects contained in a region. One of the main drawbacks of these conventional techniques is their computational efficiency; they must examine a large number of sliding windows one-by-one. Another problem is that, as object boundaries are often not well-defined in remote-sensing images, the result of feature extraction algorithms, such as scale-invariant features (SIFT) [10], direction gradient histogram (HOG) [11], etc., are unreliable for object detection. Therefore, the bag-of-words (BoW) model [12] based on statistical histogram is no longer popular in image classification task. Fundamentally, the lack of high-level semantic information limits the effectiveness of feature descriptors in complex remote-sensing scenes. Modern object detection techniques use learned feature maps instead. Many convolutional neural networks (CNNs)-based models, such as AlexNet [13], VGG [14] and ResNet [15], utilize the high-level semantic feature extraction capabilities of neural networks and have demonstrated great strength on image classification tasks. In comparison with conventional techniques, CNNs are more robust against position shifting, scaling, stretching of objects, making them excellent choices for object detection. For most remote sensing object detection (RSOD) related tasks, the state of the art was set by CNNs [16,17].

Most CNN-based object detection methods employ end-to-end learning, mapping an image to the bounding boxes of objects in the image. To make the training of the networks more effective, these methods often use a region proposal network (RPN) [18] to pick out regions that may contain objects of interest. Then they extract the features of these candidate regions and feed them into an object classifier and a position estimator to get object types and positions, respectively. To increase their robustness against scale changes, some CNN methods refine the positions of sizes of object-bounding boxes using global-scale features and features of fully connected layers (FC). However, due to the substantial differences in object sizes in remote-sensing images, such a simple measure is often ineffective for achieving true scale invariance. To deal with the large variance of object size, Lin et al. [19] proposed a feature pyramid network (FPN) that utilizes multiscale features. The idea of FPN has shown great potential in improving scale robustness and has inspired several other methods. However, the weakness of FPN is its complexity. It can become overly complicated and unconducive to training if the multiscale feature maps are predicted separately. While FPN has difficulty balancing the contradiction between feature resolution and receptive field in high-resolution input images, there is also no real communication between multiscale features. In this study, we propose a cascade, low-complexity multiscale feature fusion module to strengthen the backbone against scale variations caused by large difference of objects sizes in the remote-sensing images, which is proved to be effective in challenging RSOD tasks.

No matter how effective machine learning techniques are on artificial samples, they all face the same problem in real-world applications: there are not enough labeled real remote-sensing images for training. As exemplified in Figure 1, although high-fidelity remote-sensing images containing a wide variety of objects are plentiful and readily available for researchers, most of these images do not have the corresponding information about the types and positions of objects in them. Manually labeling object-bounding boxes in these high-resolution images is a laborious work—prohibitively expensive for gathering any meaningful quantity of ground truth data for training. In contrast, with the help of learning-based automatic labeling methods, image-level labels are easy to obtain in mass. While image-level labels do not provide the same level of details as bounding-box labels, it can still be utilized to train an object detection network using weakly supervised learning.

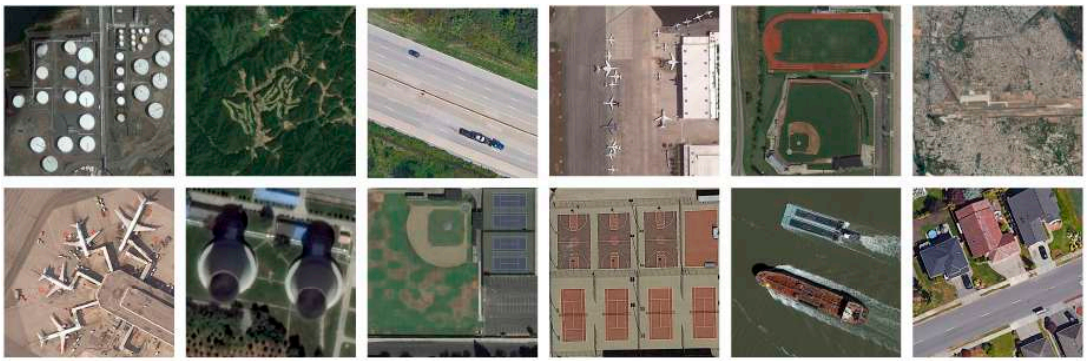


Figure 1. Complexity of multiobject and multiscale in remote-sensing images.

Based on the idea above, we propose a full-coverage collaborative network for weakly supervised RSOD. The evaluation results on two large-scale remote sensing datasets TGRS-HRRSD (High-resolution Remote Sensing Detection) [20] and DIOR [21] show that the proposed model achieves promising accuracy even without any bounding-box annotations. To make the proposed technique more effective against the previously discussed challenges, we also develop several novel techniques and make improvements upon existing approaches. Our main contributions are summarized as follows:

(1) We propose a novel end-to-end remote sensing object detection network (FCC-Net) combining a weakly supervised detector and a strongly supervised detector for addressing the challenge of insufficient labeled remote sensing data, which improves the performance of only using image-level labels training significant;
(2) We design a scale robust module on the top of the backbone using hybrid dilated convolutions and introduce a cascade multi-level pooling module for multiple feature fusion on the backend of the backbone, which promisingly suppresses the sensitivity of the network on scale changes to further enhance the ability of feature learning;
(3) We define a focal-based classification and distance-based regression multitask collaborative loss function that can jointly optimize the region classification and regression in the RPN phase;
(4) Our proposed method yields significant improvements compared with state-of-the-art methods on TGRS-HRRSD and DIOR datasets.

The remainder of this study is organized as follows: Section 2 discusses related works in the field of RSOD, including common problems and difficulties to be solved in multiobject detection, as well as existing solutions. In Section 3, we present the proposed method and its main components in detail. Section 4 introduces experimental datasets and evaluations, and Section 5 discusses the experimental results. Finally, we conclude and discuss future work in Section 6.

2. Related Work

CNNs have achieved great success in the research of natural scene image object detection, with the emergence of two-stage detection algorithm represented by R-CNN [22], fast R-CNN [23] and faster R-CNN [18] and the single-stage detection algorithm represented by YOLO [24], RetinaNet [25] and CornerNet [26]. Due to the particularity of its sensors and shooting angle, there are many differences between remote-sensing images and natural scene images, such as small objects, high resolution, complex background, imbalanced classes, insufficient training examples, etc. Therefore, how to improve the accuracy of RSOD has become a very promising subject. In this section, we review the related components of deep learning that are used in the proposed network for RSOD tasks.

2.1. Small Objects in Remote Sensing Images

Due to the large size of remote-sensing images, it is easy to miss or error-detect the small objects with only tens or hundreds of pixels. As the number of network layer increased, the feature information

of small objects will gradually weaken or even disappear. To address this problem, a large number of detection algorithms have been developed, and these methods tend to fall into two main categories: The first category of RSOD methods are feature learning-based detection algorithms, which enhances feature expression of small objects by extracting or fusing multiscale features. For instance, using unconventional convolutions to reduce the loss of feature extraction [27–30], constructing additional branches to fuse more contextual features [31–33], introducing attention mechanisms to enhance the relationship between global and local pixels [34,35], etc. Another category of RSOD methods are sample postprocessing-based detection algorithms, such as intersection over union (IoU) [36,37] and non-maximum suppression (NMS) [38,39].

2.2. Insufficient Training Examples of Remote Sensing Images

Since only image-level labels are required for training, weakly supervised learning has achieved great attention in the field of object detection. Most existing works adopt the idea of multiple-instance learning [40–45] to transform weakly supervised object detection into multilabel classification problems. Among them, the most representative model is the two-stream weakly supervised deep detection network (WSDDN) [40], which multiplies the score of classification and detection streams to select the high-confidence positive samples. Many subsequent CNN-based works [41–45] are built on WSDDN. Due to the lack of accurate bounding-box labels, the location regression performance of the weakly supervised detector is far worse than the strongly supervised detector. Therefore, some recent works [41,42,46–48] attempt to utilize the multiphase learning manner for weakly supervised object detection to improve the detection accuracy to a certain extent. However, these approaches cannot be directly applied to remote-sensing images. Thus, several efforts [49–54] are made to address the particularity of RSOD under weakly supervised learning. Although these methods achieve promising results, still have much room for improvement.

2.3. Foreground-Background Class Imbalance

In remote-sensing images, objects only occupy a small proportion of the large-scale image with complex backgrounds. Therefore, when generating proposals, a large number of proposals are the background, which dominates the gradient descent during training, resulting in a decrease in the performance of the detector. We can group the solutions for the foreground–background class imbalance into two: (1) hard sampling methods and (2) easy sampling methods. For example, the previous works [30,55,56] add online hard example mining (OHEM) to the detection networks for focusing on extracting features of hard samples so that the whole networks have a more robust classification capability. Focal loss can be introduced [57] or modified [58] to make the model focus on hard positive samples by reducing the weight of easy negative samples during training for improving the accuracy of vehicle detection.

3. Proposed Method

In this section, we present the proposed network and some related components that affect network performance in detail. This study primarily focuses on making full use of the image-level supervision information through a collaborative training network for realizing multiple-object remote sensing detection and the overall architecture of the FCC-Net is illustrated in Figure 2. As can be seen, it consists of three modules: (1) full-coverage residual network (FCRN), (2) cascade multi-level pooling module (CMPM) and (3) collaborative detection subnetwork. Moreover, we introduce focal loss and distance intersection over union (DIoU) simultaneously to define a focal-based classification and distance-based regression multitask collaborative loss function for generating proposals in the RPN phase. The strongly and weakly supervised detectors alternately optimize their own loss functions and cooperate with each other to further improve the detection performance under weakly supervised learning procedure. Specifically, training the FCC-Net contains the following three stages:

(1) Fine-tuning FCRN to extract more image edges and details information and utilizing CMPM to fuse multiscale features for better correlation between local–global information;
(2) Training a weakly supervised detector (WSD) using image-level labels and adopting the fine-tuning branch to refine the proposal results of the WSD for obtaining the final pseudo-ground-truths;
(3) Training a strongly supervised detector (SSD) with the pseudo-ground-truths generated by previous steps and minimizing the overall loss function in a stage-wise fashion to optimize the training process.

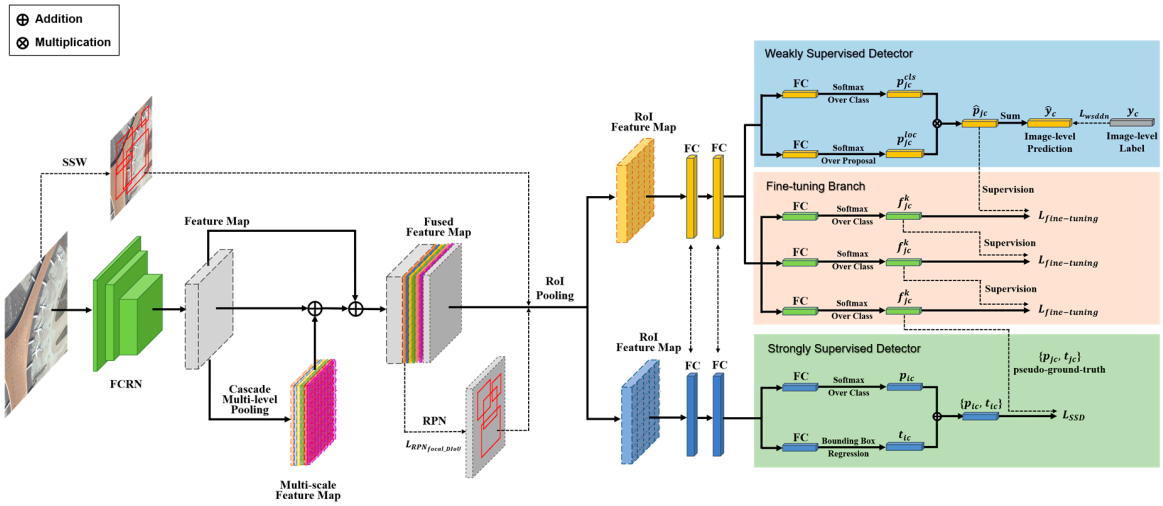


Figure 2. Overall framework of full-coverage collaborative network. (1) SSW—selective search windows; (2) FCRN—full-coverage residual network; (3) RPN—region proposal network; (4) RoI—region of interest; (5) FC—full connected layer.

In the following sections, we will describe the implementation details of the three stages.

3.1. Scale Robust Backbone for High-Resolution Remote Sensing Images

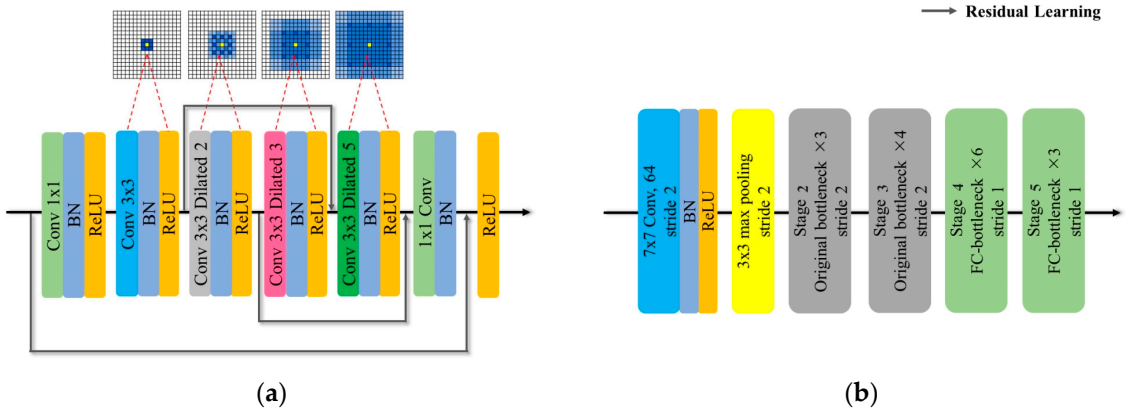
3.1.1. Full-Coverage Residual Network

Objects in high-resolution remote-sensing images usually appear in large resolution spans. Many researchers try to introduce dilated convolution to expand the receptive field—and thus, to extract features of the higher receptive field [59,60]. However, there is an inherent problem of dilated convolution called gridding artifacts, that is, the lack of interdependence between neighboring pixels that leads to the loss of local information. Given the above problem, this paper designs a novel backbone for RSOD by modifying ResNet-50.

We insert three 3×3 dilated convolutions with dilated rates of 2, 3 and 5 after the normal 3×3 convolution in the original bottleneck to form a contiguous dilated convolution combination with dilated rates of 1, 2, 3 and 5, thus constructing a new bottleneck block, namely full-coverage bottleneck (FC bottleneck), as shown in Figure 3a. Moreover, we add two shortcuts in the two contiguous dilated convolutions, respectively, which reuse the low-level features that contribute to object localization a lot. Compared with the dilated bottleneck in dilated residual network [55], the proposed FC bottleneck can acquire different levels of context information from a broader range of pixels without causing gridding artifacts. It also reduces the probability of dispersion or explosion that often occurs in gradient propagation due to the depth change.

We remain the first third stages of ResNet-50, then stack six and three FC bottlenecks in stages 4 and 5, respectively, replacing the original stages 4 and 5 to effectively enhance the correlation between long-ranged information. The whole structure of FCRN is shown in Figure 3b. The downsampling operation is removed in stages 4 and 5 so that the resolutions of output features are maintained at 1/8

of the original image. Moreover, stage 4 and stage 5 keep the same input channels as stage 3, i.e., 256, since contiguous dilated convolutions are time-consuming. Our experiments demonstrate that the proposed FCRN can increase the final receptive field to cover the entire area and avoid voids or loss of edge information, which directly improves the robustness of the detector to multiscale objects in remote-sensing images. Table 1 shows the architectures of ResNet-50 and our modified network.



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

Figure 3. Structure of components and architecture of full-coverage residual network. (**a**) FC bottleneck; (**b**) overall architecture of the FCRN.

Table 1. Architectures of ResNet-50 and our modified network.

	ResNet-50		**Modified**	
Stage	**Layer**	**Output_Size**	**Layer**	**Output_Size**
Stage 1	7×7, 64, stride 2	1/2	7×7, 64, stride 2	1/2
Stage 2	3×3, max-pooling, stride 2	1/4	3×3, max-pooling, stride 2	1/4
	$\begin{bmatrix} 1 \times 1.64 \\ 3 \times 3.64 \\ 1 \times 1.256 \end{bmatrix} \times 3$		$\begin{bmatrix} 1 \times 1.64 \\ 3 \times 3.64 \\ 1 \times 1.256 \end{bmatrix} \times 3$	
Stage 3	$\begin{bmatrix} 1 \times 1.128 \\ 3 \times 3.128 \\ 1 \times 1.512 \end{bmatrix} \times 4$	1/8	$\begin{bmatrix} 1 \times 1.128 \\ 3 \times 3.128 \\ 1 \times 1.512 \end{bmatrix} \times 4$	1/8
Stage 4	$\begin{bmatrix} 1 \times 1.256 \\ 3 \times 3.256 \\ 1 \times 1.1024 \end{bmatrix} \times 6$	1/16	$\begin{bmatrix} 1 \times 1.256 \\ 3 \times 3.256 \\ 3 \times 3, \textit{dilate}\ 2.256 \\ 3 \times 3, \textit{dilate}\ 3.256 \\ 3 \times 3, \textit{dilate}\ 5.25 \\ 1 \times 1.1024 \end{bmatrix} \times 6$	1/8
Stage 5	$\begin{bmatrix} 1 \times 1.512 \\ 3 \times 3.512 \\ 1 \times 1.2048 \end{bmatrix} \times 3$	1/32	$\begin{bmatrix} 1 \times 1.256 \\ 3 \times 3.256 \\ 3 \times 3, \textit{dilate}\ 2.256 \\ 3 \times 3, \textit{dilate}\ 3.256 \\ 3 \times 3, \textit{dilate}\ 5.256 \\ 1 \times 1.1024 \end{bmatrix} \times 3$	1/8
	Avg-Pooling, FC Layer, Softmax		Removed	

3.1.2. Cascade Multi-Level Pooling Module

To further strengthen the performance of the backbone in the feature learning stage, we introduce a cascade multi-level pooling module (CMPM) after the backend of FCRN. This module can freely fuse multiscale feature maps for adapting to the detection of various-sized objects in remote-sensing images, especially for strip objects in remote-sensing images. Figure 4 shows the structure of the proposed CMPM.

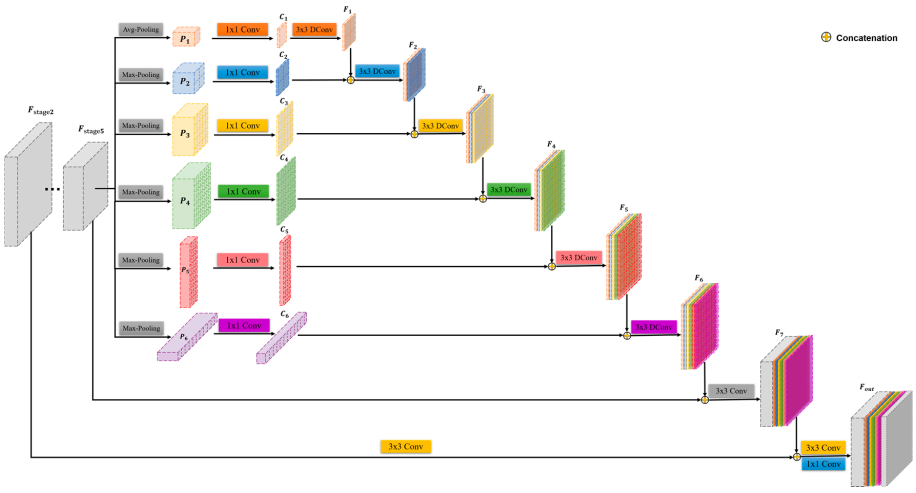


Figure 4. Structure of cascade multi-level pooling module.

First, this module adopts the multi-level pooling layer with six different-size adaptive pooling kernels to reduce the dimension of the output feature maps of FCRN: $F_{stage5} \in \mathbb{R}^{H_1 * W_1 * C_1}$, H, W and C represent the length, width and channel of the feature map, respectively.Then we can obtain contextual features at six different fixed spatial scales: $P_i = \{P_1 = 1 \times 1, P_2 = 2 \times 2, P_3 = 4 \times 4, P_4 = 8 \times 8, P_5 = 10 \times 2, P_6 = 2 \times 20\}$. Among them, the fifth and sixth pooling layers are specifically designed for difficult-to-detect objects in remote-sensing images, such as bridges and ships. Two small rectangular pooled features in vertical or horizontal directions P_5 and P_6 are added to further increase the feature representation of strip objects. In addition to the first level of pooling being average-pooling, the other five levels of pooling all utilize max-pooling.

Second, we use a 1×1 convolution to compress the dimension of P_i to 1/8 of input feature channels for limiting the weight of global features in the subsequent feature fusion stage and acquire the intermediate features: $C_i = \{C_1, C_2, C_3, C_4, C_5, C_6\}$. C_i are gradually upsampled through a 3×3 deconvolution from top to bottom and further concatenated on the channel dimension layer-by-layer to obtain the fused features: $F_i = \{F_1, F_2, F_3, F_4, F_5, F_6, F_7\}$, which can prevent the information loss in the process of directly upsampling the minimum size to the maximum size.

Finally, we downsample the output feature maps of stage 2 in FCRN: $F_{stage2} \in \mathbb{R}^{H_2 * W_2 * C_2}$, to the scale of F_7 obtained from the previous operation and perform three convolution operations after concatenation to get the final output of the backbone: $F_{Out} \in \mathbb{R}^{H_3 * W_3 * C_3}$. The convolution kernel sizes are 3×3, 3×3 and 1×1, respectively.

The module is defined as follows:

$$\begin{aligned} P_i &= \begin{cases} p_{avg}(F_{stage5}), & i = 1 \\ p_{\max}(F_{stage5}), & i \in \{2, 3, \ldots, 6\} \end{cases} \\ C_i &= f_{conv}(P_i), \; i \in \{1, 2, \ldots, 6\} \\ F_i &= \begin{cases} f_{dconv}(C_i), & i = 1 \\ f_{dconv}(C_i \oplus F_{i-1}), & i \in \{2, 3, \ldots, 6\} \\ f_{conv}(F_{stage5} \oplus F_i), & i = 7 \end{cases} \\ F_{out} &= f_{conv}(f_{conv}(F_{stage2}) \oplus F_7) \end{aligned} \tag{1}$$

where $p_{\text{avg}}(*)$ and $p_{\max}(*)$ represent the operation of avg-pooling and max-pooling, $f_{conv}(*)$ represents the operation of convolution and activation function, $f_{dconv}(*)$ represents the operation of deconvolution and activation function, $\oplus$ represents the concatenation operation of feature maps on the channel dimension. The module we propose is similar to the FPN, but we only utilize the fused feature from stage 2 and 5 for detection, instead of making predictions on features of each level. The experimental results show that CMPM can enable abundant different coarse-fine-grained features to be shared and

reused. Moreover, compared with the multiscale feature prediction and fusion operation in FPN, CMPM can maintain accurate feature expressions, especially for strip objects in remote-sensing images.

3.2. Collaborative Detection SubNetwork for Weakly Supervised RSOD

As mentioned above, since there are no bounding-box labels, it is a great challenge for weakly supervised methods to accurately predict the positions of the objects in remote-sensing images. Therefore, we attempt to balance this contradiction by using a two-phase training procedure, i.e., a multiple instance learning detector followed by a strongly supervised detector with bounding-box regression. In this work, we design a two-phase collaborative detection subnetwork with both multiple instance learning and bounding-box regression branches that share the same backbone, and we introduce it to the RSOD task. Specifically, we choose WSDDN as the baseline to generate pseudo-ground-truths and integrate faster R-CNN for more accurate bounding-box regression.

3.2.1. Weakly Supervised Detector (WSD)

As shown in Figure 3, selective search windows (SSW) [61] is first utilized to propose J_w region proposals per remote-sensing image in WSDDN. Then these images are fed into our proposed backbone in the previous article and an ROI pooling layer. This can deal with the features of different scales and improve the robustness of the network to scale changes of input images. After two FC layers, the outputs of WSDDN are split into two branches: one branch predicts the probability p_{jc}^{cls} that the object in the proposal j belonging to class c; another predicts the probability p_{jc}^{loc} of the object contained in the proposal j according to its position. Next, the predictions of the two branches are multiplied by element-wise product to get the class label $\hat{p}_{jc}$ of the proposal j. Finally, we sum up the class labels from all proposals in a remote-sensing image as its corresponding prediction $\hat{y}_c$ of the image-level multiclass label. The binary cross-entropy (BCE) loss is adopted to train the initial instance classifier with predict labels $\hat{y}_c$ and ground-truths y_c, as shown in Equation (2):

$$L_{wsddn} = L_{BCE}(y_c, \hat{y}_c) = -\sum_{c=1}^{C} (y_c \log(\sum_{j=1}^{J_w} \hat{p}_{jc}) + (1 - y_c) \log(1 - \sum_{j=1}^{J_w} \hat{p}_{jc})) \tag{2}$$

Moreover, we introduce the online instance classifier refinement (OICR) algorithm [41] into the WSD for further optimizing the local optimal problem of WSDDN in the regression phase. Specifically, we add three same fine-tuning branches ($k = 3$) in the subnetwork, which are parallel to the two branches in WSDDN. Each branch applies a max-out strategy to select the highest-confidence proposal and the proposals that overlaps it highly as reference so that filters out most redundancy predictions, as given in Equation (3). The loss function for each fine-tuning branch is Equation (4):

$$j_c = \arg\max \hat{p}_{jc} \tag{3}$$

$$L_{fine-tuning} = -\frac{1}{J_w} \sum_{j=1}^{J_w} \sum_{c=1}^{C+1} w_j^k \hat{y}_{jc}^k \log f_{jc}^k \tag{4}$$

where f_{jc}^k represents the output of the k^{th} fine-tuning branch, $\hat{y}_{jc}^k \in \{0, 1\}$ is image-level pseudo-ground-truth for each proposal. w_j^k represents the weight for each proposal. Unlike the original OICR, the weight w_j^k in our WSD module is modified by imposing a spatial restriction on negative labeling for alleviating the problem that multiple objects of the same class are easy to be mislabeled in remote-sensing images in Equation (5) as follows:

$$w_j^k = \begin{cases} f_{jc}^{k-1}, & \text{if IoU}(j, j_c) < i_t \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where i_t is a threshold, and we set $i_t = 0.1$. For more details, please refer to [62].

After that, the final loss of the WSD is defined as Equation (6):

$$L_{WSD} = L_{wsddn} + \sum_{k=1}^{3} L_{fine-tuning} \tag{6}$$

3.2.2. Strongly Supervised Detector (SSD)

The network initializes from the training of WSD and when the loss drops below a threshold, we assume that the region proposals with high scores proposed from the network are reliable enough. Therefore, these proposals are provided as pseudo-ground-truths to train a SSD, i.e., faster R-CNN. The purpose of this branch is to take advantage of faster R-CNN in bounding box regression since the pseudo-ground-truths generated in the previous two branches are too rough. To reduce the number of parameters and accelerate the network's convergence, SSD and WSD share the same backbone and the weights of part of FC layers.

Faster R-CNN can be regarded as composed of RPN and fast R-CNN, in which RPN is the core part. However, directly using RPN for RSOD tasks has two limitations: (1) RPN overcomes the impact of class-imbalance on network performance by setting the ratio of positive and negative samples, but it also results in losing the diversity of proposals during training; (2) RPN suffers from the problems of slow convergence and inaccurate regression in the bounding box regression process, especially for multiscale objects in remote-sensing images. Inspired by the observations, we attempt to optimize the loss function of RPN by introducing focal loss [25] and DIoU [63] simultaneously. In this study, we design a focal-based classification and distance-based regression multitask collaborative loss function to replace the original loss function in RPN for accurate bounding boxes and positive proposals in remote-sensing images. The flow chart of this loss function is shown in Figure 5. The formulation of $L_{RPN_{focal\_DIoU}}$ is given as follows:

$$L_{RPN_{focal\_DIoU}} = \frac{1}{N_{cls}} \sum\nolimits_i L_{focal}(p_{ic}, p^*_{ic}) + \frac{\lambda}{N_{reg}} \sum\nolimits_i p^*_{ic} L_{DIoU}(t_{ic}, t^*_{ic}) \tag{7}$$

where p_{ic} represents the predicted probability of the i-th proposal of class c, p^*_{ic} is the i-th proposal of class c. t_{ic} and t^*_{ic} are the coordinate vectors of the predicted proposal and ground truth, respectively. λ is the weight of balancing the classification loss and bounding box regression loss of RPN. Focal loss L_{focal} and DIoU loss L_{DIoU} are expressed as follows, respectively:

$$L_{focal} = -(\alpha p^*_{ic}(1-p_{ic})^{\gamma} \log(p_{ic}) + (1-\alpha)(1-p^*_{ic}) p^{\gamma}_{ic} \log(1-p_{ic})) \tag{8}$$

$$L_{DIoU} = 1 - IoU + (d^2_{\rho}/d^2_c) \tag{9}$$

where α and γ are the hyperparameters. d_c represents the diagonal length of the smallest enclosing box covering predicted anchor box and the ground-truth, d_{ρ} represents the distance of central points of predicted anchor boxes and ground-truths.

Then, the region proposals selected by NMS are output to train the subsequent fast R-CNN. Generally, training a fast R-CNN involves a classification loss and a bounding box regression loss. Due to the lack of refined bounding-box labels, the actual supervision information in our SSD branch is the pseudo-ground-truths $\left\{(p_{jc}, t_{jc})\right\}$ generated by the first two weakly supervised branches. Considering that both the WSD and SSD are used to predict the object-bounding boxes, thus we adopt a consistency prediction loss to constrain the training of the two detectors. The prediction consistency loss consists of the loss between the SSD and the WSD and the internal loss of the SSD. The loss function for the prediction consistency loss is defined as:

$$L_{SSD} = \sum_{j=1}^{J_w} \sum_{i=1}^{J_s} \sum_{c=1}^{C} I_{ij}(\beta L_{cls\_inter}(p_{ic}, p_{jc}) + (1-\beta) L_{cls}(p_{ic}) + p_{jc} L_{reg}(t_{ic}, t_{jc})) \tag{10}$$

where L_{cls_inter} represents the consistency of class predictions between two detectors by using multiclass cross-entropy. L_{reg} represents the smooth L_1 loss function and I_{ij} indicates the overlap of the IoU of the

regions proposed by the two detectors. I_{ij} will set to be 1 if the IoU is greater than 0.5 and otherwise 0. β is a hyperparameter between 0 and 1 to balance the consistency of the predictions of the two detectors. The larger the β indicates that the SSD trusts the object-bounding boxes predicted by WSD more.

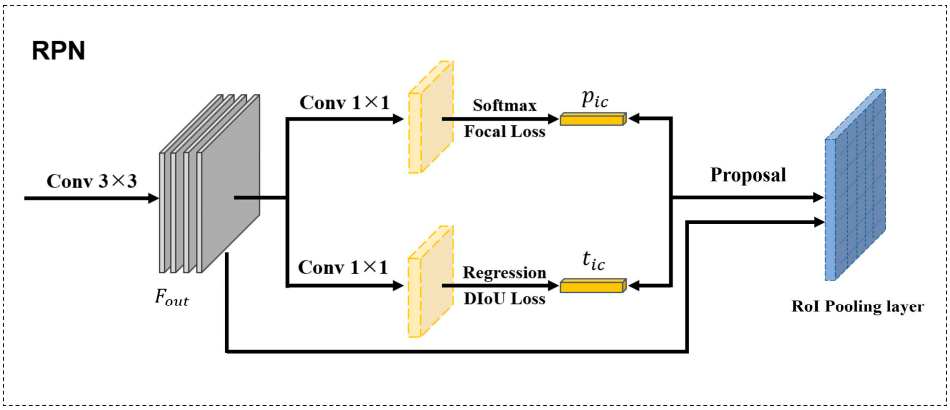


Figure 5. Flow chart of the multitask collaborative loss function.

3.3. Overall, Loss Function

After introducing our collaborative detection subnetwork, we can formulate the total loss function for FCC-Net. Our FCC-Net is trained by optimizing the following composite loss functions from the four components using stochastic gradient descent:

$$L_{total} = L_{wsddn} + L_{fine-tuning} + L_{RPN_{focal\_DIoU}} + L_{SSD} \tag{11}$$

Empirically, we set the hyperparameters $\alpha = 0.25$, $\beta = 0.8$, $\gamma = 2$ and $\lambda = 1$ of the individual loss functions in the following experiments.

To make the overall training strategy of FCC-Net clearer, we summarize the process in Algorithm 1.

Algorithm 1 FCC-Net Algorithm

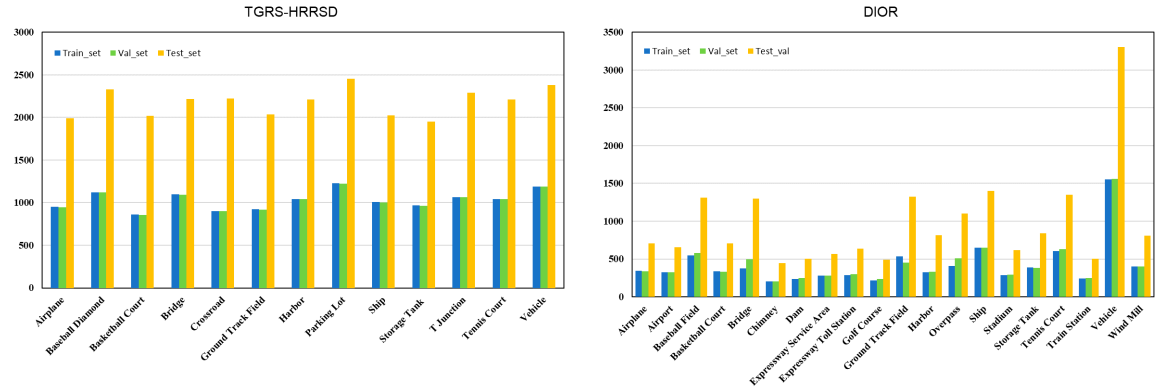
Inputs: remote-sensing image I and image-level labels y_c
Output: Detection results

Step 1: *Feature Extraction and Fusion*
(1) Extract feature of I by the backbone FCRN
(2) Conduct cascade multi-level pooling to realize multiscale feature fusion by Equation (1)
Step 2: *Collaborative Training*
while iterations less than max iteration or *Loss* larger than threshold **do**
(1) Propose region proposals of I by SSW
(2) For the region proposals:
Compute classification probability p_{jc}^{cls}, position probability p_{jc}^{loc}
Compute detection probability $\hat{p}_{jc} = S_{jc}^{cls} * S_{jc}^{loc}$
(3) Generate image-level prediction $\hat{y}_c = \sum_{j=1}^{N} \hat{p}_{jc}$
(4) Max-out on the predictions generated from the previous branch by Equation (3)
(5) Fine-tune and generate pseudo-ground-truth labels $\{(\boldsymbol{p_{jc}}, \boldsymbol{t_{jc}})\}$ by Equation (4)
(6) Generate region proposals using RPN by Equation (7)
(7) Generate bounding-box predictions $\{(\boldsymbol{p_{ic}}, \boldsymbol{t_{ic}})\}$
(8) Collaborative training to optimize Equation (10)
if the total loss Equation (11) converges **then**
Stop
end if
end while

4. Experimental Settings

4.1. Datasets

We evaluate the superiority and generalization of our method on two large public multiclass remote-sensing image datasets as follows: Figure 6 illustrates the number of objects of each class on two datasets.



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

Figure 6. Objects number of each class on two benchmarks. (**a**) TGRS-HRRSD dataset; (**b**) DIOR dataset.

TGRS-HRRSD dataset contains a total of 21,761 high-altitude images from Google Earth and Baidu Map. The minimum resolution of these images ranges from 0.15 m to 1.2 m, with a tremendous difference in resolution, which is difficult for the object detection task. The whole dataset contains a total of 55,740 target object instances in 13 categories. This dataset is divided into three subsets, in which the training set contains 5401 images, the validation set contains 5417 images and the test set contains 10,943 images. The trainval set and test set each account for 50% of the total dataset.

DIOR dataset contains 23,463 high-altitude remote-sensing images and 192,472 object instances covered by 20 categories. The size of images in the dataset is 800 × 800 pixels and the spatial resolutions range from 0.5 m to 30 m. These images are carefully selected from Google Earth by researchers and this dataset has the largest scale on both the number of images and the number of object categories. Researchers fully consider the different weather, seasons, imaging conditions and image quality when collecting these remote-sensing images, which makes the background variations of the dataset rich and diverse. Moreover, since there are many categories of target objects, this dataset has a high interclass similarity and intraclass diversity, thus making it much challenging for the object detection task. This dataset is divided into three subsets, in which the training set contains 5862 images, the validation set contains 5863 images and the test set contains 11,738 images. The trainval set and test set each account for 50% of the total dataset.

4.2. Evaluation Metrics

To quantitatively evaluate the performance of the proposed method in this work, we adopt the widely used precision–recall curve (PRC), average precision (AP), mean average precision (mAP) and correct location (CorLoc).

4.2.1. Precision–Recall Curve

The PRC is characterized by precision as the *Y*-axis and recall as the *X*-axis, such that before generating the PRC, we need to calculate the precision and recall first. Precision refers to the proportion of correctly detected objects in all detected objects and recall refers to the proportion of correctly detected objects in all positive examples detected.

4.2.2. Average Precision and Mean Average Precision

AP is a more general metric in the field of object detection and information retrieval. As normally defined, the average precision refers to the average precision value within the interval from 0 to 1 for the recall rate, which is also the area under the precision–recall curve. Normally, higher average precision results in better model performance. Moreover, mAP refers to the mean of the AP for each class.

4.2.3. Correct Location

CorLoc is another common evaluation metric for weakly supervised object detection. CorLoc measures the accuracy of localization by calculating the proportion of the detected correct images to all true positive in the dataset. CorLoc is evaluated on the union of the training set and validation set, and AP is measured on the test set.

4.3. Implementation Details

The network proposed in this study can be trained end to end. The whole framework is implemented on Ubuntu 16.04 and Python 2.7 with Pytorch 0.2. In the training progress, we adopt the stochastic gradient descent (SGD) with a batch size of 4 for optimization. The momentum and weight decay are set as 0.9 and 0.0001, respectively. The proposed FCC-Net is trained by 20 K iterations, and the initial learning rate is set as 0.001 for the first 10 K iterations—then it decades to 0.0001 for the next 10 K. All backbones adopt ImageNet pre-trained weights when possible and then fine-tune on the two benchmarks used in this work. We conducted all experiments on a computer with a Intel Xeon E5-2650 v2 CPU and a single NVIDIA GTX TITAN-V GPU for acceleration. The whole training converges in 47 h on TGRS-HRRSD dataset and 44 h on DIOR dataset. Our method achieves ~0.8 fps with an input of ~1000 × 1000 pixels (i.e., TGRS-HRRSD dataset) and ~1.1 fps with an input of 800 × 800 pixels images (i.e., DIOR dataset).

5. Experimental Results and Discussion

5.1. Evaluation on TGRS-HRRSD Dataset

Table 2 presents the quantitative comparison results of the eleven different methods in terms of AP and mAP on the TGRS-HRRSD dataset. To save space, we represent the class names with C1 to C13 according to the order in Figure 6a. The double underscore in table is used to distinguish between strongly supervised and weakly supervised methods for a clear illustration. It can be seen that our proposed FCC-Net significantly outperforms the other three traditional approaches and two weakly supervised methods in terms of mAP by a large margin of about at least 10.3% improvements. For classes of airplane, ship, ground track field and storage tank, our approach achieves better AP values than the WSDDN and OICR, which is also competitive compared with the other four strongly supervised methods. After our observation, we believe that this is because the objects of these categories have moderate size, high shape recognition and a relatively low probability of co-occurrence with other classes of objects, so they are easy to distinguish. However, for classes of crossroad, T Junction and parking lot, our method is far less accurate than the strongly supervised methods. Through the experimental results of faster R-CNN, we can find some commonalities with BoW method. For example, the AP value of basketball court and parking lot are low on both models. Moreover, compared with the two-stage models, we perform experiments on the single-stage model YOLOv2. The mAP value also reaches 65.8 and the AP value of each category is similar to faster R-CNN. Whether single-stage or two-stage models, the methods based on deep learning all show a similar performance, which we think is caused by the similarity of the backbone in feature extraction.

Table 2. Average precision (AP) (%) of different detection networks on the TGRS-HRRSD dataset.

Methods	The Class AP													mAP
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	
BoW [12]	36.8	7.0	4.4	7.2	12.2	42.1	24.1	5.3	39.3	51.5	0.7	12.5	11.4	18.9
SSCBoW [64]	59.3	35.0	7.9	9.3	11.6	7.6	53.5	6.8	37.6	26.3	0.8	12.2	30.5	23.0
FDDL [65]	30.5	20.4	2.5	10.1	18.2	15.0	19.3	7.4	33.2	72.7	1.9	16.0	4.0	19.3
COPD [66]	62.5	52.2	9.6	7.1	8.6	72.4	47.8	17.4	45.3	50.1	1.6	58.1	32.7	35.8
Transformed CNN [13]	77.5	57.6	18.0	20.0	25.9	76.3	54.1	16.6	49.7	79.1	2.4	70.8	41.3	45.3
RICNN [67]	78.1	59.6	23.0	27.4	26.6	78.0	47.8	20.5	56.5	81.0	9.3	66.4	52.0	48.2
YOLOv2 [68]	84.6	62.2	41.3	79.0	43.4	94.4	74.4	45.8	78.5	72.4	46.8	67.6	65.1	65.8
Fast R-CNN [23]	83.3	83.6	36.7	75.1	67.1	90.0	76.0	37.5	75.0	79.8	39.2	75.0	46.1	66.5
Faster R-CNN [18]	90.8	86.9	47.9	85.5	88.6	90.6	89.4	63.3	88.5	88.7	75.1	80.7	84.0	81.5
WSDDN [40]	47.9	51.4	13.6	3.0	18.0	89.6	22.6	13.4	31.8	51.5	5.1	32.8	13.7	31.1
OICR [41]	34.2	33.5	23.1	2.9	13.1	88.9	9.0	17.6	50.9	73.3	13.2	36.1	14.6	32.3
FCC-Net (FCRN + CMPM + MCL)	64.6	52.3	21.1	22.4	28.6	90.3	18.2	25.3	60.5	72.6	18.2	43.6	35.8	42.6

Meanwhile, we can see the performance measured in terms of CorLoc in Table 3. Compared with WSDDN and OICR, our method obtains 19.1% and 13.3% gains, respectively. The possible explanation is that the introduction of the spatial restriction in fine-tuning branches leads to more makes objects of the same class remain and improves the ability of multi-instance mining. In general, although there is a certain gap between our method and the strongly supervised methods such as YOLO and faster R-CNN, we have achieved remarkable success in narrowing the gap between weakly and strongly supervised methods.

Table 3. CorLoc (%) of different detection networks on the TGRS-HRRSD dataset.

Methods	mAP	CorLoc
WSDDN [40]	31.1	31.6
OICR [41]	32.3	37.4
FCC-Net (FCRN + CMPM + MCL)	42.6	50.7

The above results show that a variety of potential factors can affect the performance of RSOD. First of all, the complexity of the backbone network and the robustness of multiscale objects are the primary problems faced by remote-sensing images feature extraction. Different object sizes and spatial image resolution result in the limitation of the general image classification network in RSOD task, so we need to design a new backbone. Second, the size of objects in remote-sensing images is generally small. When the backbone is used to extract the high-level features of small objects, the image resolution is low so that some feature information may be lost, while the low-level features have high resolution and contain more position and detail information. Therefore, fusing multiscale features can enhance the relevance of context information and improve the performance of small object detection. Lastly, the optimization strategy in RPN can also enhance the robustness of the detection subnet to objects of different shapes and angles in remote-sensing images to a certain extent.

5.2. Evaluation on DIOR Dataset

To further test the effect of our method on remote sensing datasets with more categories, we also conduct some comparative experiments on the DIOR dataset. We test the dataset on faster R-CNN, and the mAP value reached 54.1%, with a performance degradation of more than 25% compared to the TGRS-HRRSD dataset, which is very consistent with the description of the dataset in Section 4.1. Due to the complexity of this dataset, it is a greater difficulty for the backbone to extract features of high robustness, and the effect of the object detection is reduced consequently. In addition, the different spatial resolutions of the data make it hard to learn multiscale features in a relatively shallow backbone. To this end, we need to further modify the backbone, that is, expand both the depth and width. To verify this idea, we also carried out evaluations on RetinaNet, a single-stage detector. Resnet-50 and

Resnet-101 are used as the backbones, respectively. In Table 4, the experimental results show that a deeper and more complex backbone can bring about performance improvements, increasing the mAP from 65.7% to 66.1%. The weak improvement of 0.4% may be due to the reason that RetinaNet itself has reached a saturation state on this benchmark. Furthermore, although more complex backbones can indeed extract features of high robustness, it may be difficult to improve further due to the limitations of the network itself. In the evaluation of our method, we set Res50, FRCN and FRCN-CMPM as the backbones, respectively, and carry out more comparisons. The results are consistent with the distribution of the TGRS-HRRSD dataset in Table 2, and mAPs are slightly improved. However, the optimal mAP on this dataset is nearly 30% lower than that on the TGRS-HRRSD dataset. The gap between these models is not large, which confirms our hypothesis of the dataset and the effectiveness of this work.

Table 4. Mean average precision (mAP) (%) of different detection networks on the DIOR dataset.

Methods	Faster R-CNN [18]	RetinaNet [25]		FCC-Net		
	VGG16	ResNet-50	ResNet-101	ResNet-50	FRCN	FRCN + CMPM
mAP	54.1	65.7	66.1	17.2	17.7	18.1

In Table 5, we plot the AP values of each class in the DIOR dataset. To save space, we represent the class names with C1 to C20 according to the order in Figure 6b. The double underscore in table is used to distinguish between strongly supervised and weakly supervised methods for a clear illustration. It is can be seen clearly that the DIOR dataset is much more challenging than the TGRS-HRRSD dataset, the detection performance of two strongly supervised methods is much better than all these three weakly supervised methods. However, compared with the other two state-of-the-art weakly supervised methods, our method outperforms WSDDN and OICR with an improvement of 5% and 1.8% in terms of mAP, respectively. For the two classes of ship and bridge that are difficult to detect, our method achieves the best performance among the three weakly supervised methods. Moreover, for classes of baseball field, ground track field and stadium, our method is still very competitive compared with the two strongly supervised methods without any bounding-box supervision information. At the same time, we can see in Table 6 that our method improved by 9.3% and 6.9% in terms of CorLoc compared with WSDDN and OICR, respectively.

Combining the two evaluation metrics, we can find that our proposed network does not decrease in performance on more difficult and large-scale benchmarks. There are three main reasons as follows: (1) The combination of FCRN and CMPM fully improves the utilization and sharing of abundant multiscale features so that increases the network's ability to express the features of these hard instances; (2) The addition of three fine-tuning branches enables WSD to generate more accurate pseudo-ground-truths, providing to SSD for further bounding-box regression; (3) The multitask collaborative loss function mitigates the adverse effects of complex background and prevents the network from overfitting negative samples in the classification stage.

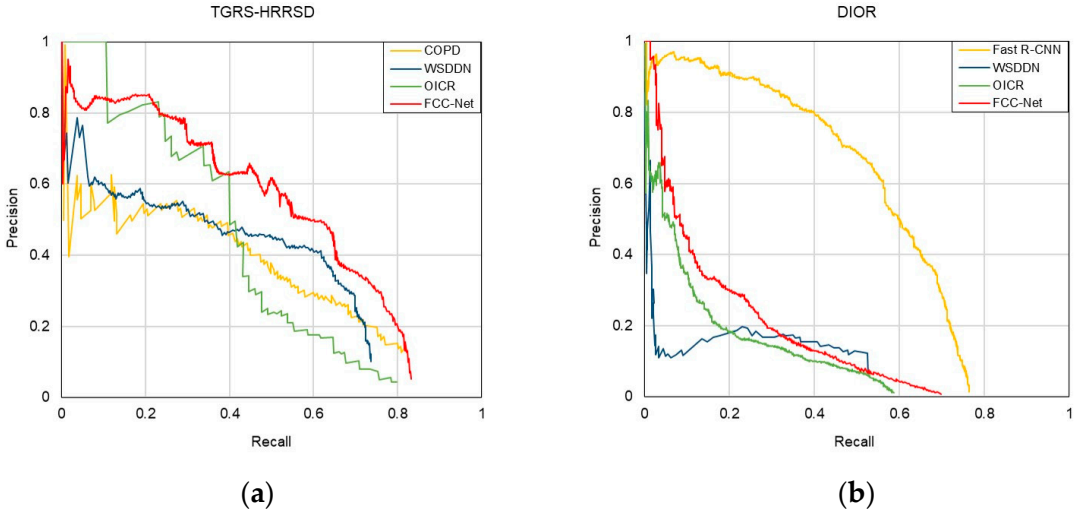
For a more intuitive understanding, we draw the PRC on the two benchmarks of the proposed network in this work. As shown in Figure 7, with the same Precision, the Recall of FCC-Net is higher than any of the comparison methods, which means that our method can detect more actual objects; with the same Recall, the precision of FCC-Net is higher than any of the comparison methods, which means that the false alarm rate of our method is lower. These observations adequately demonstrate that our method is highly competitive compared to other state-of-the-art weakly supervised object detection methods.

Table 5. AP (%) of different detection networks on the DIOR dataset.

Methods	The Class AP																			mAP
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	
Fast R-CNN [23]	44.2	66.8	67.0	60.5	15.6	72.3	52.0	65.9	44.8	72.1	62.9	46.2	38.0	32.1	71.0	35.0	58.3	37.9	19.2	38.1	50.0
Faster R-CNN [18]	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	54.1
WSDDN [40]	9.1	39.7	37.8	20.2	0.3	12.2	0.6	0.7	11.9	4.9	42.4	4.7	1.1	0.7	63.0	4.0	6.1	0.5	4.6	1.1	13.3
OICR [41]	8.7	28.3	44.1	18.2	1.3	20.2	0.1	0.7	29.9	13.8	57.4	10.7	11.1	9.1	59.3	7.1	0.7	0.1	9.1	0.4	16.5
FCC-Net (FRCN + CMPM + MCL)	20.1	38.8	52.0	23.4	1.8	22.3	0.2	0.6	28.7	14.1	56.0	11.1	10.9	10.0	57.5	9.1	3.6	0.1	5.9	0.7	18.3

Table 6. CorLoc (%) of different detection networks on the DIOR dataset.

Methods	mAP	CorLoc
WSDDN [40]	13.3	32.4
OICR [41]	16.5	34.8
FCC-Net (FCRN + CMPM + MCL)	18.3	41.7



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

Figure 7. Precision–recall (PR) curve of our model on two benchmarks: (**a**) TGRS-HRRSD dataset: (**b**) DIOR dataset.

5.3. Ablation Experiments

To evaluate the effectiveness of our proposed FCC-Net, we constructed some ablation experiments on the TGRS-HRRSD dataset to future analyze the contributions of the three key components in this work, such as FCRN, CMPM and multitask collaborative loss. In the ablation experiments, we gradually added the components used in this work and carry out comparative experiments. In Table 7, FCC–FCRN represents the collaborative subnetwork based on FRCN, FCC–FCRN–CMPM represents the addition of CMPM on this basis, FCC–FCRN–CMPM–MCL represents the further addition of multitask collaborative loss.

(1) Impact of full-coverage residual network. Under the collaborative detection subnetwork of VGG16 as the backbone, the mAP value reaches 46.6%, which proves that collaborative training has more advantages than the mere weakly supervised model. Then, we tried two deeper and more complex backbones, i.e., ResNet-50 and ResNet-101 and the results show that complex networks have more advantages in feature representation and can bring a certain degree of improvement. Our proposed FCRN outperforms ResNet-101 with an improvement of 0.4% in terms of mAP, which confirms that the FC bottleneck greatly reduces the loss of information during the feature extraction process and improves the detection performance of the entire network.

(2) Impact of cascade multi-level pooling module. Using FCRN and CMPM simultaneously, we find that mAP value is further improved by 0.6%. For classes of bridge and ship, FCC–FCRN–CMPM achieves better AP results than the previous weakly supervised methods. CMPM successfully deals with the problem of low feature recognition of strip objects in remote-sensing images, which future confirms that the multifeature fusion strategy is effective to mine objects under weakly supervised settings, especially for the objects of small size and special shape. The combination of FRCN and CMPM leads to that the scale robustness of the network is significantly improved.

(3) Impact of Multitask Collaborative Loss. We combined all the components proposed in this study to reach a mAP value of 48.3%, the accuracy of the model fluctuated slightly by 0.1%. The main reason is that the predicted bounding boxes of the objects extracted by PRN used in our framework are

very dense, the calculation results of the IoU between the predicted bounding boxes and ground-truth boxes are always in an ideal horizontal interval. Thus, the improvement of our proposed network accuracy by multitask collaborative loss is relatively limited, but still much higher than WSDDN.

Table 7. AP (%) of the components of FCC-Net on the TGRS-HRRSD dataset.

Methods	The Class AP													mAP
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	
FCC–VGG16	62.3	49.0	17.8	21.5	26.9	82.3	10.4	21.5	55.8	69.3	15.0	42.1	31.9	38.9
FCC–Res50	62.5	49.1	18.0	20.9	27.6	84.7	10.1	23.9	56.0	69.4	14.9	42.7	32.4	39.4
FCC–Res101	63.0	49.1	18.2	21.7	27.7	83.9	10.4	24.0	55.9	68.6	15.7	42.6	33.3	39.5
FCC–FCRN	63.5	49.7	19.2	22.6	29.9	83.3	11.3	25.1	57.8	72.6	18.2	43.9	32.1	40.7
FCC–FCRN–CMPM	64.7	51.6	20.6	24.2	28.1	88.3	19.2	26.2	59.2	72.4	18.1	44.5	35.0	42.5
FCC–FCRN–CMPM–MCL	64.6	52.3	21.1	22.4	28.6	90.3	18.2	25.3	60.5	72.6	18.2	43.6	35.8	42.6

5.4. Qualitative Results

We present part of the detection results on the TGRS-HRRSD and DIOR datasets in Figures 8 and 9, respectively. It can be seen that our method can successfully detect most classes of objects and give precise and tight bounding boxes on both two datasets. When faced with multiple objects to be detected in a remote-sensing image, our method also shows better performance. However, we also observe that our method is easy to treat densely arranged small objects as a whole and cannot detect every single object, such as the class of ship and vehicle. This is also the direction we will focus on in our next work.

Figure 8. Partial detection results of FCC-Net on the TGRS-HRRSD dataset.

Figure 9. Partial detection results of FCC-Net on the DIOR dataset.

6. Conclusions

The traditional strongly supervised object detection methods require a large amount of finely labeled image data, which has a high cost and poor generalization performance. Considering the gap between remote-sensing images and general images, we propose a novel collaborative network (FCC-Net) in this study. We jointly train WSD and SSD in an end-to-end manner and feed the prediction results of the WSD into the SSD as the pseudo-ground-truths. This makes full use of their respective advantages to realize the collaborative optimization of region classification and regression. Moreover, a scale robust backbone is designed to enhance the feature learning of the multiscale objects in remote-sensing images. The quantitative evaluations on the TGRS-HRRSD and DIOR datasets demonstrate the effectiveness of the proposed method. Specifically, compared with two previous state-of-the-art weakly supervised methods and four traditional methods, FCC-Net archives the substantial improvements in terms of mAP. Moreover, the detection performance of FCC-Net in several classes is still highly competitive compared with some state-of-the-art strongly supervised methods. Through experiments, we also reveal an important mechanism, that is, the deeper and more complex backbone has more advantages than the shallow backbone, which is more suitable for diversified RSOD tasks.

Author Contributions: Conceptualization, S.C. and D.S.; methodology, S.C. and D.S.; software, D.S.; validation, D.S. and C.Z.; formal analysis, D.S.; investigation, C.Z.; writing—original draft preparation, D.S.; writing—review and editing, S.C., X.S. and J.W.; visualization, D.S.; supervision, S.C., X.S.; project administration, S.C.; funding acquisition, S.C. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Nos. 61906097, 41875184) and a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Acknowledgments: The authors would like to thank all reviewers and editors for their constructive comments for this study.

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

References

1. Campbell, J.B.; Wynne, R.H. *Introduction to Remote Sensing*, 5th ed.; Guilford Press: New York, NY, USA, 2011; ISBN 978-1-60918-176-5.
2. Iliffe, J.; Lott, R. *Datums and Map Projections for Remote Sensing, GIS and Surveying*, 2nd ed.; Whittles Publishing: Scotland, UK, 2008.
3. Gherboudj, I.; Ghedira, H. Assessment of solar energy potential over the United Arab Emirates using remote sensing and weather forecast data. *Renew. Sustain. Energy Rev.* **2016**, *55*, 1210–1224. [CrossRef]
4. Lindgren, D. *Land Use Planning and Remote Sensing*; Taylor & Francis: Milton Park, UK, 1984; Volume 2.
5. Liu, Y.; Wu, L. Geological Disaster Recognition on Optical Remote Sensing Images Using Deep Learning. *Procedia Comput. Sci.* **2016**, *91*, 566–575. [CrossRef]
6. Chintalacheruvu, N. Video Based Vehicle Detection and its Application in Intelligent Transportation Systems. *J. Transp. Technol.* **2012**, *2*, 305–314. [CrossRef]
7. Massad-Ivanir, N.; Shtenberg, G.; Raz, N.; Gazenbeek, C.; Budding, D.; Bos, M.P.; Segal, E. Porous Silicon-Based Biosensors: Towards Real-Time Optical Detection of Target Bacteria in the Food Industry. *Sci. Rep.* **2016**, *6*, 38099. [CrossRef] [PubMed]
8. Jalilian, E.; Xu, Q.; Horton, L.; Fotouhi, A.; Reddy, S.; Manwar, R.; Daveluy, S.; Mehregan, D.; Gelovani, J.; Avanaki, K. Contrast-enhanced optical coherence tomography for melanoma detection: An in vitro study. *J. Biophotonics* **2020**, *13*, e201960097. [CrossRef]
9. Turani, Z.; Fatemizadeh, E.; Blumetti, T.; Daveluy, S.; Moraes, A.F.; Chen, W.; Mehregan, D.; Andersen, P.E.; Nasiriavanaki, M. Optical radiomic signatures derived from optical coherence tomography images improve identification of melanoma. *Cancer Res.* **2019**, *79*, 2021–2030. [CrossRef] [PubMed]
10. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
11. Mizuno, K.; Terachi, Y.; Takagi, K.; Izumi, S.; Kawaguchi, H.; Yoshimoto, M. Architectural study of HOG feature extraction processor for real-time object detection. In Proceedings of the 2012 IEEE Workshop on Signal Processing Systems (SiPS), Quebec City, QC, Canada, 17–19 October 2012; pp. 197–202.
12. Xu, S.; Fang, T.; Li, D.; Wang, S. Object classification of aerial images with bag-of-visual words. *IEEE Geosci. Remote Sens. Lett.* **2009**, *7*, 366–370.
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
15. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
16. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
17. Lu, X.; Zheng, X.; Yuan, Y. Remote sensing scene classification by unsupervised representation learning. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5148–5157. [CrossRef]
18. 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]
19. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
20. Zhang, Y.; Yuan, Y.; Feng, Y.; Lu, X. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5535–5548. [CrossRef]
21. 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]

22. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
23. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
25. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
26. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 765–781.
27. Zhang, W.; Wang, S.; Thachan, S.; Chen, J.; Qian, Y. Deconv R-CNN for small object detection on remote sensing images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 22–27 July 2018; pp. 2483–2486.
28. Deng, Z.; Sun, H.; Lei, L.; Zou, S.; Zou, H. Object Detection in Remote Sensing Imagery with Multi-scale Deformable Convolutional Networks. *Acta Geod. Et Cartogr. Sin.* **2018**, *47*, 1216–1227.
29. Liu, W.; Ma, L.; Wang, J.; Chen, H. Detection of multiclass objects in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 791–795. [CrossRef]
30. Ding, P.; Zhang, Y.; Deng, W.; Jia, P.; Kuijper, A. A light and faster regional convolutional neural network for object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 208–218. [CrossRef]
31. 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]
32. Ji, H.; Gao, Z.; Mei, T.; Li, Y. Improved faster R-CNN with multiscale feature fusion and homography augmentation for vehicle detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1761–1765. [CrossRef]
33. 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**. [CrossRef]
34. Ying, X.; Wang, Q.; Li, X.; Yu, M.; Jiang, H.; Gao, J.; Liu, Z.; Yu, R. Multi-attention object detection model in remote sensing images based on multi-scale. *IEEE Access* **2019**, *7*, 94508–94519. [CrossRef]
35. Chen, J.; Wan, L.; Zhu, J.; Xu, G.; Deng, M. Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 681–685. [CrossRef]
36. Yan, J.; Wang, H.; Yan, M.; Diao, W.; Sun, X.; Li, H. IoU-adaptive deformable R-CNN: Make full use of IoU for multi-class object detection in remote sensing imagery. *Remote Sens.* **2019**, *11*, 286. [CrossRef]
37. 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]
38. Qiu, S.; Wen, G.; Deng, Z.; Liu, J.; Fan, Y. Accurate non-maximum suppression for object detection in high-resolution remote sensing images. *Remote Sens. Lett.* **2018**, *9*, 237–246. [CrossRef]
39. Zhu, M.; Xu, Y.; Ma, S.; Li, S.; Ma, H.; Han, Y. Effective Airplane Detection in Remote Sensing Images Based on Multilayer Feature Fusion and Improved Nonmaximal Suppression Algorithm. *Remote Sens.* **2019**, *11*, 1062. [CrossRef]
40. Bilen, H.; Vedaldi, A. Weakly supervised deep detection networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2846–2854.
41. Tang, P.; Wang, X.; Bai, X.; Liu, W. Multiple instance detection network with online instance classifier refinement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2843–2851.
42. Tang, P.; Wang, X.; Bai, S.; Liu, W.; Yuille, A. PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 176–191. [CrossRef]
43. Wan, F.; Wei, P.; Han, Z.; Jiao, J.; Ye, Q. Min-Entropy Latent Model for Weakly Supervised Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2395–2409. [CrossRef]

44. Inoue, N.; Furuta, R.; Yamasaki, T.; Aizawa, K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5001–5009.
45. Wan, F.; Liu, C.; Ke, W.; Ji, X.; Jiao, J.; Ye, Q. C-MIL: Continuation multiple instance learning for weakly supervised object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2199–2208.
46. Zhang, Y.; Bai, Y.; Ding, M.; Li, Y.; Ghanem, B. W2f: A weakly-supervised to fully-supervised framework for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 928–936.
47. Wang, J.; Yao, J.; Zhang, Y.; Zhang, R. Collaborative learning for weakly supervised object detection. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 971–977.
48. Wang, X.; Wang, J.; Tang, P.; Liu, W. Weakly-and Semi-Supervised Fast Region-Based CNN for Object Detection. *J. Comput. Sci. Technol.* **2019**, *34*, 1269–1278. [CrossRef]
49. Han, J.; Zhang, D.; Cheng, G.; Guo, L.; Ren, J. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Trans. Geosci. Remote Sens.* **2014**, *53*, 3325–3337. [CrossRef]
50. Zhou, P.; Cheng, G.; Liu, Z.; Bu, S.; Hu, X. Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping. *Multidim. Syst. Sign. Process.* **2015**, *27*, 925–944. [CrossRef]
51. Zhang, F.; Du, B.; Zhang, L.; Xu, M. Weakly supervised learning based on coupled convolutional neural networks for aircraft detection. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 5553–5563. [CrossRef]
52. Ji, J.; Zhang, T.; Yang, Z.; Jiang, L.; Zhong, W.; Xiong, H. Aircraft Detection from Remote Sensing Image Based on A Weakly Supervised Attention Model. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 322–325.
53. Xu, J.; Wan, S.; Jin, P.; Tian, Q. An active region corrected method for weakly supervised aircraft detection in remote sensing images. In Proceedings of the Eleventh International Conference on Digital Image Processing, Guangzhou, China, 11–13 May 2019; Volume 11179, p. 111792H.
54. Wu, X.; Hong, D.; Tian, J.; Kieft, R.; Tao, R. A weakly-supervised deep network for DSM-aided vehicle detection. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1318–1321.
55. Cai, B.; Jiang, Z.; Zhang, H.; Zhao, D.; Yao, Y. Airport Detection Using End-to-End Convolutional Neural Network with Hard Example Mining. *Remote Sens.* **2017**, *9*, 1198. [CrossRef]
56. Koga, Y.; Miyazaki, H.; Shibasaki, R. A CNN-based method of vehicle detection from aerial images using hard example mining. *Remote Sens.* **2018**, *10*, 124. [CrossRef]
57. Wu, Z.; Sang, J.; Zhang, Q.; Xiang, H.; Cai, B.; Xia, X. Multi-Scale Vehicle Detection for Foreground-Background Class Imbalance with Improved YOLOv2. *Sensors* **2019**, *19*, 3336. [CrossRef] [PubMed]
58. Sergievskiy, N.; Ponamarev, A. Reduced focal loss: 1st place solution to xview object detection in satellite imagery. *arXiv* **2019**, arXiv:1903.01347.
59. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460.
60. Yu, F.; Koltun, V.; Funkhouser, T. Dilated residual networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 472–480.
61. 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]
62. Kosugi, S.; Yamasaki, T.; Aizawa, K. Object-aware instance labeling for weakly supervised object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6064–6072.
63. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000. [CrossRef]

64. Sun, H.; Sun, X.; Wang, H.; Li, Y.; Li, X. Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Trans. Geosci. Remote Sens.* **2012**, *9*, 109–113. [CrossRef]
65. Han, J.; Zhou, P.; Zhang, D.; Cheng, G.; Guo, L.; Liu, Z.; Bu, S.; Wu, J. Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. *ISPRS J. Photogramm. Remote Sens.* **2014**, *89*, 37–48. [CrossRef]
66. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
67. Peicheng, Z.; Cheng, G.; Junwei, H. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7416.
68. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.