remote sensing

MDPI

Article

Context Information Refinement for Few-Shot Object Detection in Remote Sensing Images

Yan Wang, Chaofei Xu, Cuiwei Liu and Zhaokui Li *

School of Computer Science, Shenyang Aerospace University, Shenyang 110136, China; wy4615@sau.edu.cn (Y.W.); 192106074097@email.sau.edu.cn (C.X.); liucuiwei@sau.edu.cn (C.L.)
* Correspondence: lzk@sau.edu.cn

Abstract: Recently, few-shot object detection based on fine-tuning has attracted much attention in the field of computer vision. However, due to the scarcity of samples in novel categories, obtaining positive anchors for novel categories is difficult, which implicitly introduces the foreground–background imbalance problem. It is difficult to identify foreground objects from complex backgrounds due to various object sizes and cluttered backgrounds. In this article, we propose a novel context information refinement few-shot detector (CIR-FSD) for remote sensing images. In particular, we design a context information refinement (CIR) module to extract discriminant context features. This module uses dilated convolutions and dense connections to capture rich context information from different receptive fields and then uses a binary map as the supervision label to refine the context information. In addition, we improve the region proposal network (RPN). Concretely, the RPN is fine-tuned on novel categories, and the constraint of non-maximum suppression (NMS) is relaxed, which can obtain more positive anchors for novel categories. Experiments on two remote sensing public datasets show the effectiveness of our detector.

Keywords: convolutional neural network (CNN); few-shot object detection; remote sensing images; context information

Citation: Wang, Y.; Xu, C.; Liu, C; Li, Z. Context Information Refinement for Few-Shot Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3255. https://doi.org/10.3390/rs14143255

Academic Editors: Pedram Ghamisi, Danfeng Hong, Xin Wu and Sicong Liu

Received: 24 June 2022
Accepted: 30 June 2022
Published: 6 July 2022

1. Introduction

The code is available at https://github.com/Li-ZK/CIR-FSD-2022 (accessed on 29 June 2022). Object detection has always been a research hotspot in remote sensing and computer vision fields. In the past few years, object detection has made significant progress due to the rapid development of deep convolutional neural networks (CNNs). A series of excellent object detection algorithms based on CNN have emerged in natural scene images [1,2]. The object detection frameworks are generally divided into two main types according to whether they contain region proposals, i.e., one-stage and two-stage detectors. One-stage detectors, represented by the You Only Look Once (YOLO) series [3–7], directly generate class-related bounding boxes and their probabilities at each spatial location. In contrast, two-stage detectors, represented by the region-based CNN (R-CNN) series [8], including Fast R-CNN [9] and Faster R-CNN [10], adopt a region proposal algorithm to improve the performance of object detection.

Compared with natural scene images, remote sensing images (RSIs) have the characteristics of arbitrary directions, different object sizes and complex backgrounds. To deal with the problems mentioned above, researchers have proposed many excellent solutions based on CNN [11]. Wu et al. [12] proposed an Optical Remote Sensing Imagery detector (ORSIm detector) that incorporates feature extraction, feature learning, fast image pyramid matching, and boosting strategies. Qian et al. [13] incorporated a multi-level feature fusion module into the existing hierarchical deep network, which can fully use the multi-level features. Cheng et al. [14] proposed a two-stage oriented detector for detecting arbitrary-oriented objects in RSIs. They generated high-quality oriented proposals

through an oriented RPN, and refined these proposals through an oriented R-CNN head. Yang et al. [15] proposed a sampling fusion network to improve sensitivity to small objects. They use a supervised multi-dimensional attention network to attenuate the noise in remote sensing images and highlight object information. Zheng et al. [16] proposed a multitask learning network that treats each small-scale training dataset as a task. They utilize shared branches to extract shared features across tasks to better adapt to remote sensing images.

Although the object detection method based on CNN has achieved great success, training a deep detector usually requires sufficient annotation data. Collecting annotated data in the real world is time-consuming and expensive, which makes it difficult to obtain enough annotated data. This raises considerable attention about learning efficient detectors with limited training samples. Chen et al. [17] proposed a low-shot transfer detector for object detection in few-shot cases, which transfers rich source domain knowledge to the target domain. Xu et al. [18] designed a cross-domain ship detection method, which can transfer labeled optical ship images to unlabeled SAR images. Wu et al. [19] designed a multi-source active fine-tuning network to achieve vehicle detection without the requirement for well-annotated samples. Few-shot object detection (FSOD) has gradually become an effective mechanism to address this issue, which can learn new concepts from limited training samples.

Currently, FSOD is mainly divided into three categories: meta-learning-based, metric-learning-based, and fine-tuning-based approaches [20,21]. Meta-learning [22–25] usually utilizes many episodes to learn task-agnostic notions (e.g., meta-parameters), which might be meaningful for quickly adapting to a new session. Kang et al. [22] used a meta feature learner to extract meta-features from the query images and designed a reweighting module to acquire the global features of the support images. Li et al. [23] designed a meta-learning model with multiscale architecture to solve the inherent scale fluctuations in remote sensing images by introducing feature pyramid network (FPN) [26]. Cheng et al. [25] designed a prototype learning network (PLN) to obtain the prototypes of each class, and used the prototypes to guide a region proposal network to generate higher-quality proposals, which can more effectively choose foreground objects from complicated backgrounds in RSIs. Meta-learning approaches divide so many small tasks and design a complex episodic training scheme, which can cause a lot of training time and more memory with an increasing number of categories in the support set.

Metric-learning [27–29] focuses on learning a robust encoding function and a rating function that measure the similarity of a query image's embedding vectors to each category prototype. Karlinsky et al. [28] proposed a novel metric-learning-based method for representing each category that uses a mixed model with multiple modes, and they take the centers of these modes as the category's representative vectors. During the training process, the method concurrently learns the embedding space, the model weights and the representative vectors for each category. Yang et al. [29] found that negative proposals, especially hard negative ones, are essential for learning an embedding space. Therefore, they introduced a new metric learning framework inference scheme based on negative and positive representations. The embedding space representing the positive and negative vectors in both methods is learned by utilizing a triplet loss [30]. Metric-learning approaches require the use of training data to learn a robust collection of class prototypes as task-specific parameters, which makes building robust class prototypes problematic when the dataset contains numerous outliers (such as occlusion).

Recently, Two-stage Fine-tune Approach (TFA) has shown great potential in the field of FSOD [31]. Compared to meta-learning and metric-learning approaches, TFA can yield competitive performance through a simple fine-tuning strategy. TFA utilizes a simple two-stage treatment on Faster R-CNN, which freezes the pre-trained weights in the first stage and fine-tunes the last layers in the second stage. Wu et al. [32] proposed a scale-aware network based on TFA to distinguish positive–negative exemplars by combining the FPN with object pyramids. Zhang et al. [33] proposed a novel data hallucination-based approach to address the problem of lack of variety in training data,

which efficiently transfers common patterns of within-category variation from base categories to novel categories. Zhao et al. [34] designed a path-aggregation multiscale few-shot detector for remote sensing images (PAMS-Det), which can mitigate the scale scarcity in novel categories by adding a path-aggregation module. In addition, PAMS-Det designed an involution backbone to improve the classification ability of the object detector in remote sensing images. Huang et al. [35] designed a shared attention module and a balanced fine-tuning strategy to cope with large size variations and improve the classification accuracy. Li et al. [36] designed a few-shot correction network to eliminate false positives caused by class confusion. This can improve the limitation of TFA for classifier rebalancing. Sun et al. [37] introduced contrastive learning into TFA to learn contrastive-aware object proposal embeddings, which is helpful to classify the detected objects. Sun et al. observed that the positive anchors for novel objects received relatively low scores from region proposal network, resulting in fewer positive anchors passing non-maximum suppression (NMS) and becoming proposals. The low-score positive anchors for novel objects are mostly regarded as background noise, which introduces the problem of foreground–background imbalance.

Due to various object sizes and cluttered backgrounds, it is still a challenging problem to identify foreground objects from complex backgrounds in remote sensing images, even with the help of the aforementioned FSOD methods. Firstly, the receptive field of FPN is insufficient to capture rich context information for objects of different sizes since the effective receptive fields of CNN are substantially smaller than the expected receptive fields [38,39], which may lead to the failure of FPN to detect objects correctly. Yang et al. [40] proposed Densely connected Atrous Spatial Pyramid Pooling (DenseASPP) for semantic segmentation of street scenes and achieved remarkable success. They used the dilated convolution to obtain different receptive fields and used the dense connections to aggregate multiple atrous-convolved features as the final feature representation. In addition, due to the complexity of RSIs, excessive background noise might override the target information, and the boundary between the targets will become blurred, which will lead to missed detection. Wang et al. [41] designed a multiscale refinement FPN and nonlocal-aware pyramid attention to suppress background noise and focus more on the valuable object features. Finally, because of the scarcity of samples in novel categories, it is difficult to obtain the positive anchors for novel categories, which implicitly introduces the foreground-background imbalance problem. In this article, to tackle the above challenges, we introduce a fine-tuning-based method for few-shot object detection, which designs a novel context information refinement few-shot detector (CIR-FSD) for remote sensing images. In order to better extract discriminative context features, we devise a context information refinement (CIR) module. In CIR, firstly, the dilated convolutions and dense connections are used to capture rich context information from different receptive fields. Then, a binary map is used as supervision labels to refine context information, which can suppress the noise and enhance the object information. In addition, baseline TFA usually needs to freeze all parameters trained on base categories and fine-tune only the box classification layer and regression layer on novel categories, which prevents RPN from learning the features related to novel categories. In our method, in addition to the box classifier and regressor, RPN is also fine-tuned on base and novel categories, which can increase the confidence of positive anchors for novel categories. Further, we relax the constraint of NMS on the confidence of anchors. Fine-tuning RPN and relaxing NMS can obtain more positive anchors for novel categories, which can alleviate the imbalance between the foreground and background. Our main contributions are highlighted as follows:

(1) We design a novel context information refinement few-shot detector for remote sensing images, which can effectively detect objects of different scales and cluttered objects in complex backgrounds with a few annotated samples.
(2) A CIR module is designed to obtain rich context information from different receptive fields and refine it at the same time, which can learn discriminative context features.

(3) Our proposed method increases the confidence of positive anchors for novel categories by fine-tuning RPN, and relaxes the constraint of NMS on the confidence of anchors, which can obtain more positive anchors for novel categories to alleviate the imbalance between the foreground and background.

2. The Proposed RSI Few-Shot Object Detection Framework

2.1. Overall Architecture

Depending on the high accuracy and recall, a two-stage detector is widely employed for object detection in the remote sensing of images, such as various improved algorithms based on the popular Faster R-CNN [10]. Shivappriya et al. [42] applied the Additive Activation Function (AAF) to Faster R-CNN to improve the efficiency of object detection. In the mainstream detection frameworks (i.e., Detectron2 [43] and MMDetection [44]), two-stage object detector generally includes multiple modules, such as Backbone, FPN [26], RPN and Roi Feature Extractor. These detectors usually generate region of interests (RoIs) using independent RPN, and then further classify these RoIs and accurately regress them as the final results. As shown in Figure 1, our proposed network architecture is built upon the popular detector Faster R-CNN. First, the images in the training set are fed into the Backbone for basic feature extraction. Then a CIR module is implemented for extracting discriminative context features. Next, these features are processed by a top-down module and lateral connections in FPN to generate multi-scale feature maps for objects of various sizes. These multi-scale feature maps are fed into the RPN for RoIs generation. Finally, these RoIs are pooled to a uniform scale and perform the final classification and regression tasks. The total loss function of the network is as follows:

$$Loss = L_{RPN} + L_{cls} + L_{reg} + \lambda L_{CIR}, \quad (1)$$

where the first term denotes RPN loss, the second term represents classifier loss, the third term expresses regression loss, and the fourth term denotes CIR loss. The equilibrium parameter lambda represents the coefficient of CIR loss, which is taken as 0.35 in experiments and more details are described in Section 3.2.

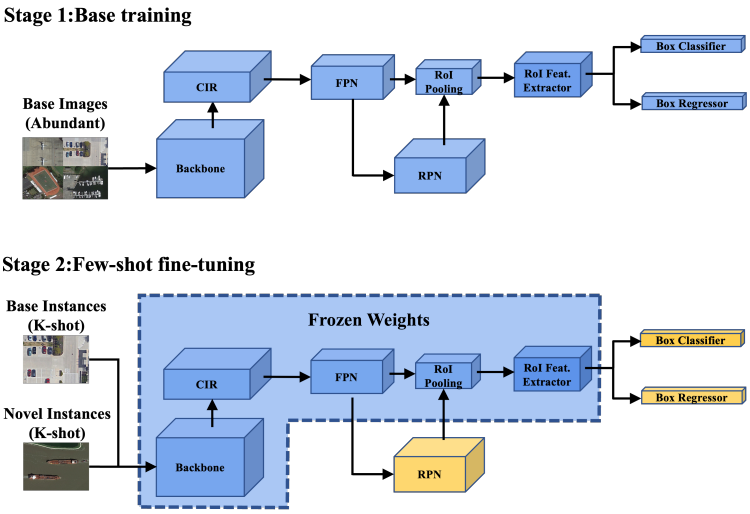

Figure 1. The overview architecture of the proposed few-shot RSI object detector.

2.2. Two-Stage Fine-Tuning Strategy

Usually, the dataset of the experiment is divided into a training set D_{train} and a test set D_{test}. Under the few-shot detection scenario, the categories in the D_{train} are divided into base categories D_{base} and novel categories D_{novel}, $D_{\text{base}} \cap D_{\text{novel}} = \emptyset$. To train a robust model, the basic categories require as many samples as feasible, whereas the novel categories usually contain several annotated samples. To make full use of the rich prior

knowledge embedded in large-scale samples and transfer this knowledge to a few novel samples, the training of our CIR-FSD is divided into two stages: base training stage and few-shot fine-tuning stage.

In the first stage, the network is trained on the base categories with abundant labels to learn prior knowledge. For each sample (X_{train}, Y_{train}) in D_{base}, X_{train} is an image ($X_{train} \in \mathbb{R}^{H \times W \times 3}$) and Y_{train} is the label of the image X_{train}. It is generally considered that such prior knowledge is stored in the feature extraction modules, such as Backbone, CIR, FPN, RoI feature extractor, etc., so that the parameters of these models are usually frozen.

In the second stage, both prior knowledge learned from base categories and new knowledge related to novel categories are utilized to detect the targets of novel classes. For several instances (I_{train}, Y_{train}) in $D_{base} \cup D_{novel}$, I_{train} is the instances of the image ($I_{train} \in X_{train}$), Y_{train} is the label of the instances I_{train}, and the maximum number of instances per category was set to k (k is generally no more than 20).

In our CIR-FSD, RPN acts as a binary classification network responsible for filtering out possible foreground objects from the background. If the RPN parameters are kept frozen, novel categories can easily be taken as background due to their low confidence. Different from the feature extraction module that is only responsible for extracting category-independent features, RPN needs to extract category-related features. To have the discriminative ability to identify novel categories from complex backgrounds, RPN needs to learn the knowledge of novel categories. Therefore, we improve RPN to enhance the confidence of novel categories in the fine-tuning stage. More novel categories are separated from the background and treated as foreground. In addition, the box classifier and regressor in the first stage are designed for base categories, and we also fine-tune them in the second stage to adapt to novel categories.

In the process of fine-tuning, the base categories often introduce catastrophic forgetting problems as it receives less attention, and there is a great deal of research examining this difficulty. Wang et al. [45] proposed a new online continual learning dataset and evaluation metrics, which can sufficiently evaluate catastrophic forgetting. Fan et al. [46] proposed a novel fine-tuning method, called Retentive R-CNN, which avoids catastrophic forgetting by combining pretrained RPN and fine-tuned RPN. Guirguis et al. [47] propose a constraint-based fine-tuning approach to mitigate catastrophic forgetting. To prevent catastrophic forgetting, we used a few images from the base categories for fine-tuning, which preventing the model from overfitting the novel categories. For model testing, we test our CIR-FSD method on the test set D_{test}. Algorithm 1 depicts the whole training and testing process.

Algorithm 1 Process of Training and Testing for the CIR-FSD

1: Create a large-scale training set D_{base} out of base classes, and a small-scale training set D_{novel} out of novel classes, $D_{\text{base}} \cap D_{\text{novel}} = \emptyset$.
2: Construct a testing set D_{test} for evaluation.
3: Initialize the parameters ψ, θ, ϕ, χ in the Backbone module, CIR module, FPN module, and RoI Feature extractor module.
4: **for** each sample $(X_{train}, Y_{train}) \in D_{base}$ **do**
5: Base training.
6: **end for**
7: Keep the network parameters ψ, θ, ϕ, χ fixed.
8: **for** k instances per class $(I_{train}, Y_{train}) \in D_{novel} \cup D_{base}$ **do**
9: Few-shot fine-tuning.
10: **end for**
11: **for** each sample $(X_{test}, Y_{test}) \in D_{test}$ **do**
12: Generate bounding boxes and category scores on each image.
13: Calculate the accuracy and recall of all correctly identified objects.
14: **end for**

2.3. Context Information Refinement

It is generally believed that a large receptive field can capture richer contextual information. However, the receptive field of the feature pyramid network is insufficient to capture the contextual information for objects of different sizes. In particular, excessive background noise in complex remote sensing images can result in an overabundance of object data and a blurring of object boundaries. To address these problems, we designed a new CIR module to learn the discriminative context features, which can classify objects correctly and localize objects precisely. Specifically, as shown in Figure 2, we adopt the backbone of Faster R-CNN (i.e., Resnet-101 [48]) to implement the feature extractor . Then, the output feature maps of the backbone network are fed into our CIR, which is composed of multi-path dilated convolutional layers with rates of 3, 6, 12, 18, and 24 to obtain multiple receptive fields.

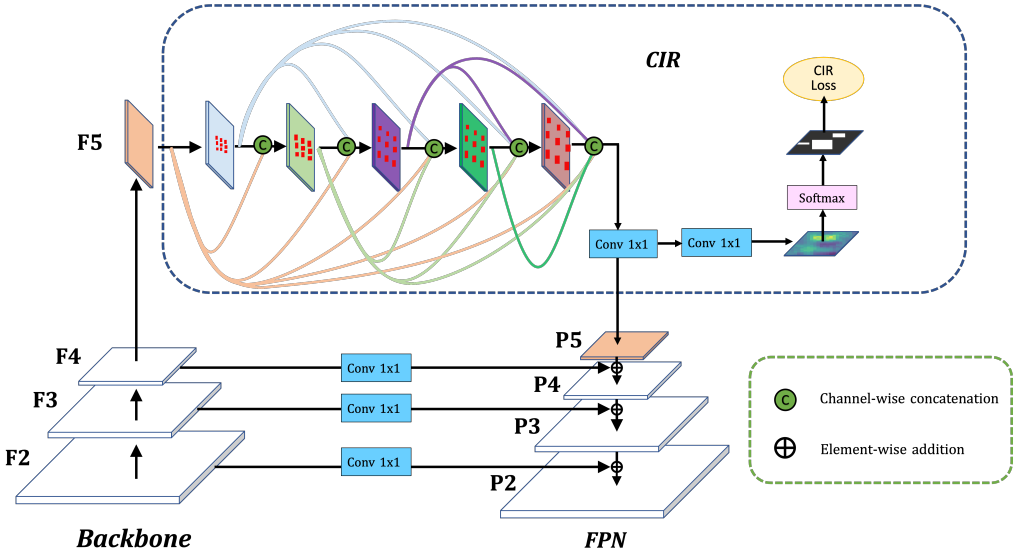

Figure 2. The diagram of the context information refinement module.

After applying multi-path dilated convolution to ResNet-101, the multiple feature maps in various receptive fields can be derived. In particular, deformable convolutions are introduced into each path, which can adapt to different scales and shapes of RSI objects. In addition, in our CIR, dense connections are adopted between each dilated convolutional layer, which can fuse better multi-scale context information. Finally, the output of the last dilated layer is sent into a 1×1 convolutional layer to fuse the multi-scale features. The structural details of the CIR implementation are described in Section 3.2.

To better refine context information, the output of the last dilated layer is also sent into two 1×1 convolutional layers to learn a two-channel saliency map, which indicates the foreground and background scores, respectively. Then, the value of the saliency map is normalized to between [0,1] by executing the Softmax function. We take a binary map obtained from ground truth as the supervision label, and then calculated the CIR loss between the binary map and the saliency map to suppress noise and enhance object information. In our method, CIR loss is essentially cross-entropy loss, and it is calculated as follows:

$$L_{CIR} = \frac{\lambda}{h \times w} \sum_{i}^{h} \sum_{j}^{w} L_{att}\left(u'_{ij}, u_{ij}\right), \tag{2}$$

where h and w denote the feature map's width and height, u'_{ij} and u_{ij} denote the prediction of mask's pixel and label respectively, and Latt is pixelwise Softmax cross-entropy.

2.4. Improved RPN

RPN is considered to be a category-independent network, which uses foreground–background classifier to select RoIs without considering their exact category, as shown in Figure 3. To prevent over-fitting, most prior studies believed that a pre-trained RPN could produce high-quality suggestions for a new assignment. As a result, they tended to freeze all parameters of such an RPN. We found that this strategy prevents RPN from learning features related to the novel categories, resulting in low confidence of the positive anchors for novel classes. Meanwhile, NMS in RPN tends to treat the novel categories as background, resulting in foreground–background imbalance. To alleviate the imbalance between the foreground and background, the RPN is fine-tuned in the fine-tuning stage and the constraint of NMS is relaxed on the confidence of positive anchors.

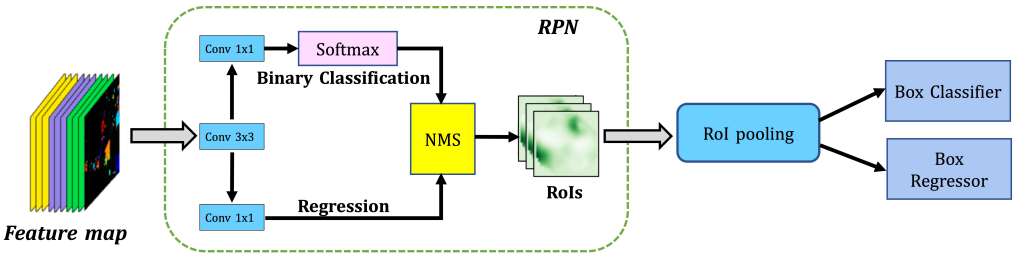

Figure 3. The diagram of the relaxed region proposal network module.

As shown in Figure 4, for the convenience of presentation, we utilize the original image to replace the input feature map. The yellow cross in the picture depicts the feature map's stride relative to the original image, and the green box represents the anchor generated by the RPN. Let the size of the ith feature map generated by FPN be $H_i \times W_i \times C_i$, where W and H denote the feature map's width and height, respectively, and C denotes the number of the feature channels.

In each feature map, $H_i \times W_i \times k$ anchor boxes are first generated by anchor generator, corresponding to the anchors around the yellow cross in the figure. For visualization, we do not draw all the anchors around the yellow cross. Usually, there are nine different anchors around each yellow cross, consisting of three sizes and three ratios. These anchors perform classification and regression tasks via 3×3 convolution and 1×1 convolution, respectively. Then, RPN randomly selects the top m positive anchors that may contain objects in each level, and P_m denotes their probability of containing objects. Usually, an intersection over union (IoU) threshold t is set to distinguish foreground and background. If an anchor satisfies $P_m > t$, it is treated as foreground, otherwise as background.

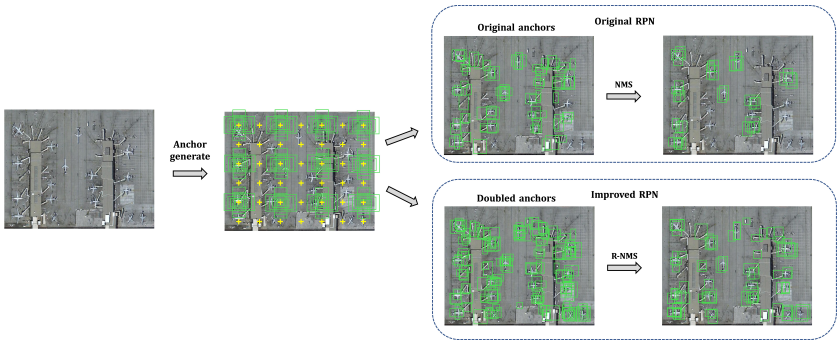

Figure 4. The proposals generated by RPN and improved RPN.

To retain more positive anchors for novel categories, we improve the RPN. Specifically, we double the preset m and propose a relaxed NMS (R-NMS). The IoU threshold t is slightly reduced to allow more anchors to be selected, and more potential targets will be relieved of inhibitions. Finally, these anchors are fine-tuned into RoIs, and RPN selects the top n RoIs to the subsequent networks for a more refined bounding box regression and

multi-classification. To compensate for positive anchors, we slightly increase the value of n. In RPN, the loss is a sum of the classification and bounding box regression losses, where the classification loss is calculated by cross-entropy and the regression loss is as follows:

$$L_{loc} = \sum_{u} \sum_{n \in x,y,w,h} smooth_{L1}(u_n(P_t) - u_n(G_t)) \tag{3}$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases} \tag{4}$$

where u denotes the foreground anchors, P_t denotes the prediction, G_t denotes the ground thuth, and the positions of the boxes are denoted by x, y, w, h. Fine-tuning RPN and relaxing NMS can obtain more positive anchors for novel categories, thus improving the ability of RPN to adapt to novel classes with fewer samples.

3. Experiments and Results

3.1. Datasets and Evaluation Metrics

DIOR [49]: A dataset consists of 23,463 images and 192,472 instances, with the training set having 5862 images, the evaluation set having 5863 images, and the test set having 11,738 images. The spatial resolution in this dataset varies from 0.5 to 30 m, and the picture size is 800 × 800 pixels. DIOR contains 20 common object categories, including airplane, tennis court, baseball field, bridge, windmill, airport, harbor, chimney, ground track field, expressway service area, dam, golf course, expressway toll station, overpass, stadium, ship, storage tank, vehicle, train station, and basketball court.

NWPU VHR-10 [50]: An open Level 10 geographic remote sensing dataset with a resolution of 0.5–2.0 m for multicategory object detection, with 10 categories of objects, including airplane, tennis court, harbor, storage tank, baseball diamond, basketball court, bridge, ground track field, vehicle, and ship. The images are rectangles of approximately 500 to 1200 pixels on the long side and are separated into two sets: negative image set and positive image set. The negative image set has 150 photos without any item categories provided, while the positive image set has 650 photos, each containing at least one detectable target.

Our experimental settings are exactly the same as that in [23], i.e., three novel categories in NWPU VHR-10 dataset, and five novel categories in DIOR dataset, with the rest of the categories in the dataset being regarded as base categories. Specifically, the three novel categories in the NWPU VHR-10 dataset are airplane, tennis court, and baseball diamond, and the five novel categories in the DIOR dataset are airplane, tennis court, baseball field, windmill, and train station. K-shot novel instances from the novel categories are randomly picked for few-shot training, where K is set to 3, 5, 10, 20, 30. To obtain relatively stable results, we set 10 random seeds and calculate average over the 10 random seeds. We take the mean average precision (mAP) as evaluating metric, and evaluate the performance of the detector through PASCAL VOC2007 development kit. The fraction of true positives (TP) is used to calculate the precision, and the detection area overlap between ground truth and detection is generally considered to be greater than the defined IoU threshold, such as 0.5. The recall formula is used to measure the fraction of correctly identified positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$AP_c = \int_0^1 P_c(R) dR \tag{7}$$

Usually, the AP of the c-th category AP_c is calculated by Formula (7), which is also considered as the area under the P-R curve. To more accurately evaluate the few-shot

object detection approaches, the mAPs for novel categories with different K values is calculated as:

$$mAP_k = \frac{1}{k}\sum_{c=1}^{k} AP_c, \tag{8}$$

where k denotes the amount of shots selected for the training set and mAP_k denotes the average accuracy of the K-shot model.

3.2. Experiment Settings

We use two RTX 2080Ti GPUs for model training, each with a batch size of two. The stochastic gradient descent (SGD) optimization algorithm is employed in the base training stage, and the initial learning rate is set to 0.005. We train 8000 iterations on the NWPU VHR-10 dataset, where the learning rate is divided by 10 at 4000 and 6000 iterations, respectively. On the DIOR dataset, we train 7000 iterations, in which the learning rate is divided by 10 at 3000 and 5000 iterations, respectively. In the fine-tuning stage, the same initial learning rate and optimization algorithm are adopted. Training the model up to 3000 iterations on two datasets is sufficient to achieve good performance. The proposed method is implemented by the Detecron2 [43], which is a free and open-source object detection framework developed and maintained by Facebook AI Research.

For our CIR, we use dilated convolutions and dense connections to obtain global context information, and use 1×1 convolutions and binary maps to refine the context information. Specifically, in CIR, we first reduce the F5 from 2048 channels to 512 channels, and then utilize several 3×3 deformable convolution layers with different dilated rates to obtain various receptive fields. Finally, 1×1 convolutions are used to reduce the channel to 256 for feeding into the top-down structure of the FPN, and further reduce the channel to 2 to calculate the loss using binary mapping. The network architecture of our proposed CIR module is shown in Table 1.

Table 1. Network architecture of the CIR module.

Module	Module Details	Input Shape	Output Shape
CIR_3_1×1	Conv	(2048, w, h)	(512, w, h)
CIR_3_3×3	DeformConv(dilate=3)	(512, w, h)	(256, w, h)
CIR_concat_1	Concatenation(C5, CIR_3_3×3)	(2048, w, h)⊕(256, w, h)	(2304, w, h)
CIR_6_1×1	Conv	(2304, w, h)	(512, w, h)
CIR_6_3×3	DeformConv(dilate=6)	(512, w, h)	(256, w, h)
CIR_concat_2	Concatenation(CIR_concat_1, CIR_6_3×3)	(2304, w, h)⊕(256, w, h)	(2560, w, h)
CIR_12_1×1	Conv	(2560, w, h)	(512, w, h)
CIR_12_3×3	DeformConv(dilate=12)	(512, w, h)	(256, w, h)
CIR_concat_3	Concatenation(CIR_concat_2, CIR_12_3×3)	(2560, w, h)⊕(256, w, h)	(2816, w, h)
CIR_18_1×1	Conv	(2816, w, h)	(512, w, h)
CIR_18_3×3	DeformConv(dilate=18)	(512, w, h)	(256, w, h)
CIR_concat_4	Concatenation(CIR_concat_3, CIR_18_3×3)	(2816, w, h)⊕(256, w, h)	(3072, w, h)
CIR_24_1×1	Conv	(3072, w, h)	(512, w, h)
CIR_24_3×3	DeformConv(dilate=24)	(512, w, h)	(256, w, h)
CIR_concat_5	Concatenation(CIR_3_3×3, CIR_6_3×3, CIR_12_3×3, CIR_18_3×3, CIR_24_3×3)	(256, w, h)⊕(256, w, h)⊕(256, w, h)⊕ (256, w, h)⊕(256, w, h)	(1280, w, h)
CIR_golobal_context_reduce_1×1	Conv	(1280, w, h)	(256, w, h)
CIR_context_refine_reduce_1×1	Conv	(256, w, h)	(2, w, h)

In addition, to explore the optimal parameters of CIR in our detector, we conducted a comparison experiment on the selection of the parameter λ in L_{CIR} on the DIOR [49] dataset. When the value of the parameter λ is 0, it means that the L_{CIR} is not used, as demonstrated in Table 2. When the parameter λ is 0.35, the 5-shot, 10-shot, and 20-shot all achieve the best performance. As a result, we choose λ = 0.35 as the CIR's equilibrium parameter. During the fine-tuning stage, we conducted a comparison experiment with hyperparameters in the model. Specifically, the fine-tuning stage mainly involves three hyperparameters, namely m, n, and t, where m represents the number of positive anchors selected, n represents the number of ROIs selected, and t represents the IoU threshold. We take 5 shot results on the DIOR dataset as the evaluation metric, and set m and n between 1500 and 3500, and the threshold between 0.2 and 0.7. To choose appropriate hyperparameters, we choose two strategies, the first is to fix the IoU threshold, compare m and n. The second is to fix m and n, compare the IoU threshold. As is shown in Figures 5 and 6, when the IoU threshold is fixed, the model has the best value when both m and n are 3000. When m and n are set to 3000 and fixed, the model achieves the best value at the IoU threshold of 0.4. Therefore, we take m, n, and the IoU threshold as 3000, 3000, and 0.4, respectively, as the optimal hyperparameters.

Table 2. Equilibrium parameter λ of L_{CIR}.

λ	5 Shot	10 Shot	20 Shot
0	0.301	0.360	0.400
0.10	0.302	0.362	0.405
0.20	0.312	0.366	0.413
0.25	0.311	0.366	0.414
0.30	0.313	0.370	0.417
0.35	**0.325**	**0.375**	**0.427**
0.40	0.315	0.372	0.420

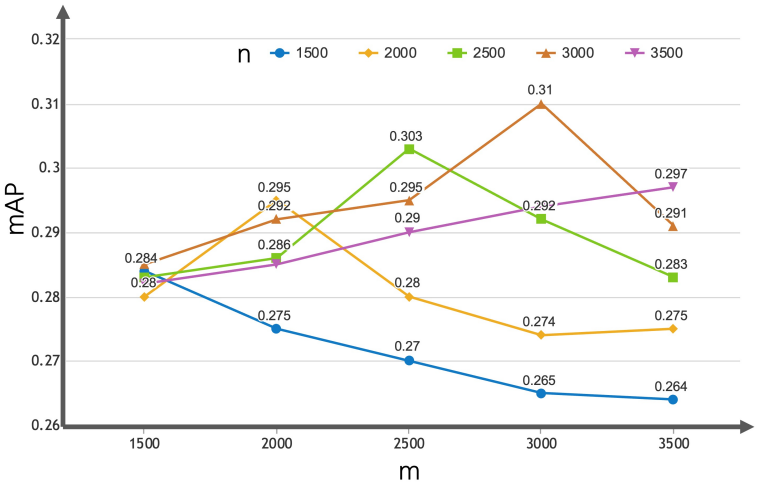

Figure 5. Equilibrium parameter m and n.

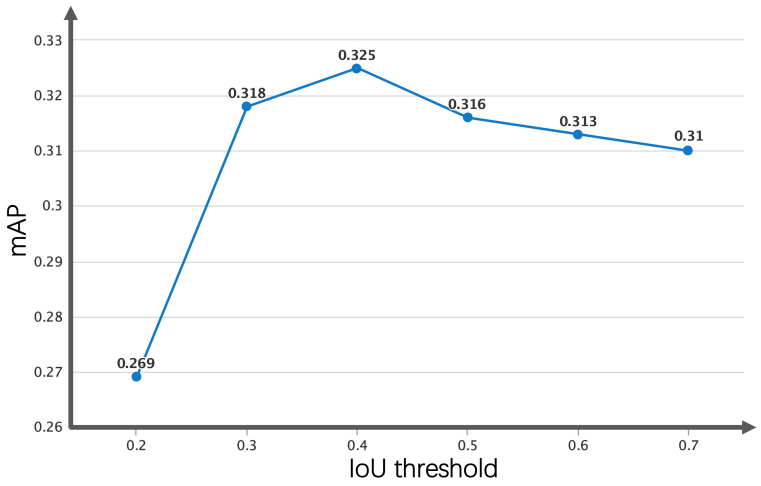

Figure 6. Equilibrium parameter IoU threshold.

3.3. Comparing Methods

We compare our proposed few-shot detector with the state-of-the-art (SOTA) FSOD methods such as meta-learning-based FSRW [22] and FSODM [23], metric-learning-based RepMet [28], fine-tuning-based TFA [31] and PAMS-Det [34]. In addition, we also choose the two-stage detector Faster R-CNN [10] algorithm and one-stage detector YOLO v5 [7] algorithm for comparison, which does not belong to FSOD method. For a fair comparison, most of the methods adopt the Faster R-CNN [10] as the base architecture, which uses the same pre-trained feature extraction network, ResNet-101 [48], on the ImageNet dataset. Exceptionally, FSRW [22] and FSODM [23]are based on the YOLO v3 [5] architecture with DarkNet53 as the feature extractor.

For the DIOR dataset, we validate the mAPs of the five novel categories at 5, 10, and 20 shots, respectively. Similarly, for the NWPU dataset, we validate the mAPs of the three novel categories at 3, 5, and 10 shots, respectively. In addition, we also validate the results of these methods for the base classes, except for the RepMet [28] method, which is based on metric-learning and cannot be validated for the base categories alone.

3.4. Results on DIOR

The detection performance of our and the comparison methods on different shots of the DIOR dataset is reported in Table 3. The proposed method significantly outperforms all previous works in any shots, as shown in Table 3. Due to training directly on few novel samples, the detection performance of YOLO v5 [7] and Faster R-CNN [10] are obviously inferior to that of FSOD methods. Compared with baseline TFA [31], PAMS-Det [34] increases mAP by 3%, 3% and 1% in 5-shot, 10-shot, and 20-shot settings, respectively. However, our proposed method is still the best to obtain detection performance. Specifically, compared with PAMS-Det, our method increases mAP by 5%, 5%, and 5% in 5-shot, 10-shot, and 20-shot settings, respectively, demonstrating the method's effectiveness.

As shown in Table 3, TFA freezes the pre-trained network parameters to ensure that the prior knowledge is preserved, resulting in higher accuracy for the base categories than metric-learning- and meta-learning-based methods. However, the improvement of TFA for the novel categories is reduced, the results under the settings of 5-shot, 10-shot, and 20-shot settings are close to the meta-learning-based FSODM. This is due to the scarcity of samples in novel categories, introducing a foreground–background imbalance. Our proposed method fine-tunes RPN and relaxes NMS, which obtains more positive anchors for the novel categories and makes better use of contextual information in fewer samples, so the mAPs for novel categories are significantly improved.

We select two current mainstream few-shot algorithms for visual comparison: one is the FSODM algorithm based on meta-learning, and the other is the baseline TFA algorithm

based on fine-tuning. The two comparison methods and our proposed CIR-FSD are visualized on DIOR and NWPU VHR-10 datasets. The results of the novel and base classes are visualized in the two datasets, respectively.

Figures 7 and 8 show the qualitative inference results of the novel and base classes on the DIOR dataset. In Figure 7, we visualize the detection results for each novel category, and in Figure 8, we select three representative categories from the base categories for analysis. As shown in Figure 7, our proposed method can obtain outstanding detection results compared to FSODM [23] and TFA [31]. For airplane, both FSODM and TFA have a lot of missed objects, especially in such complex remote sensing scenarios, and our method hardly misses any.

As shown in Figure 8, when facing dense and small objects, such as vehicle and storage tank, CIR-FSD uses contextual information to obtain the correlation between objects and uses the fusion of different scale features in FPN to detect these challenging objects. FSODM and TFA are difficult to identify all foreground objects in the images. Our CIR-FSD uses a binary map for the refinement of these objects at the instance level, which can better distinguish object boundaries and achieve accurate localization.

Table 3. Comparison of the base and novel classes results on the DIOR dataset.

Class	Shot	YOLO v5	Faster R-CNN	RepMet ‡	FSRW †	FSODM †	TFA ⋆	PAMS-Det ⋆	Ours ⋆
Base Classes Results									
airport		0.59	0.73	-	0.59	0.63	0.76	0.78	0.87
basketball court		0.71	0.69	-	0.74	0.80	0.78	0.79	0.88
bridge		0.26	0.26	-	0.29	0.32	0.52	0.52	0.55
chimney		0.68	0.72	-	0.70	0.72	0.66	0.69	0.79
dam		0.40	0.57	-	0.52	0.45	0.54	0.55	0.72
expressway service area		0.55	0.59	-	0.63	0.63	0.66	0.67	0.86
expressway toll station		0.45	0.45	-	0.48	0.60	0.60	0.62	0.78
golf course		0.60	0.68	-	0.61	0.61	0.79	0.81	0.84
ground track field		0.65	0.65	-	0.54	0.61	0.77	0.78	0.83
harbor		0.31	0.31	-	0.52	0.43	0.50	0.50	0.57
overpass		0.46	0.45	-	0.49	0.46	0.50	0.51	0.64
ship		0.10	0.10	-	0.33	0.50	0.66	0.67	0.72
stadium		0.65	0.67	-	0.52	0.45	0.75	0.76	0.77
storage tank		0.21	0.21	-	0.26	0.43	0.55	0.57	0.70
vehicle		0.17	0.19	-	0.29	0.39	0.52	0.54	0.56
mean		0.45	0.48	-	0.50	0.54	0.63	0.65	**0.74**
Novel Classes Results									
airplane	5	0.02	0.03	0.09	0.09	0.09	0.13	0.14	0.20
	10	0.08	0.09	0.13	0.15	0.16	0.17	0.17	0.20
	20	0.09	0.09	0.14	0.19	0.22	0.24	0.25	0.27
baseball field	5	0.09	0.09	0.19	0.33	0.27	0.51	0.54	0.50
	10	0.27	0.31	0.33	0.45	0.46	0.53	0.55	0.55
	20	0.30	0.35	0.34	0.52	0.50	0.56	0.58	0.62
tennis court	5	0.10	0.12	0.11	0.47	0.57	0.24	0.24	0.50
	10	0.12	0.13	0.24	0.54	0.60	0.41	0.41	0.50
	20	0.20	0.21	0.29	0.55	0.66	0.50	0.50	0.55
train station	5	0.00	0.00	0.01	0.09	0.11	0.13	0.17	0.24
	10	0.00	0.02	0.01	0.07	0.14	0.15	0.17	0.23
	20	0.02	0.04	0.03	0.18	0.16	0.21	0.23	0.28
wind mill	5	0.01	0.01	0.01	0.13	0.19	0.25	0.31	0.20
	10	0.10	0.12	0.01	0.18	0.24	0.30	0.34	0.36
	20	0.12	0.21	0.03	0.26	0.29	0.33	0.36	0.37
mean	5	0.04	0.05	0.08	0.22	0.25	0.25	0.28	**0.33**
	10	0.11	0.13	0.14	0.28	0.32	0.31	0.33	**0.38**
	20	0.15	0.18	0.16	0.34	0.36	0.37	0.38	**0.43**

† marks meta-learning based methods. ‡ marks metric-learning based methods. ⋆ marks the average of the experiments on 10 random seeds.

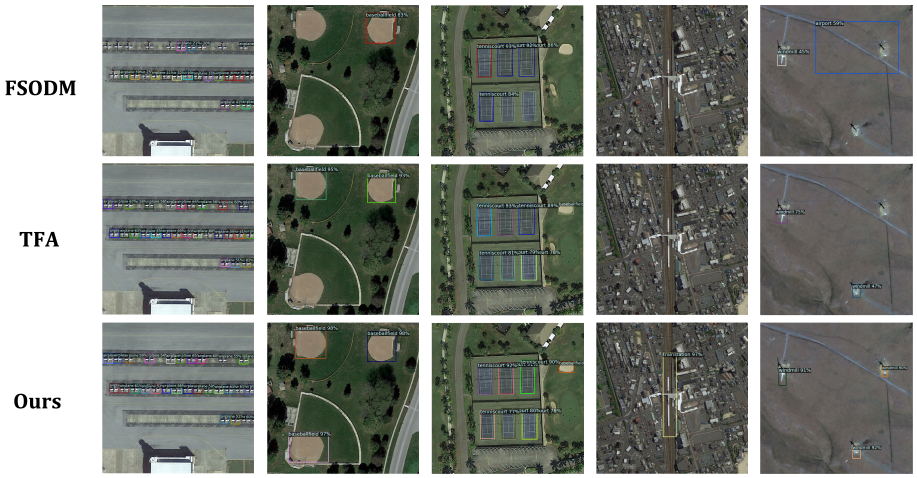

Figure 7. Qualitative inference results of novel categories on the DIOR dataset. (**a**) airplane, (**b**) baseball field, (**c**) tennis court, (**d**) train station, (**e**) wind mill.

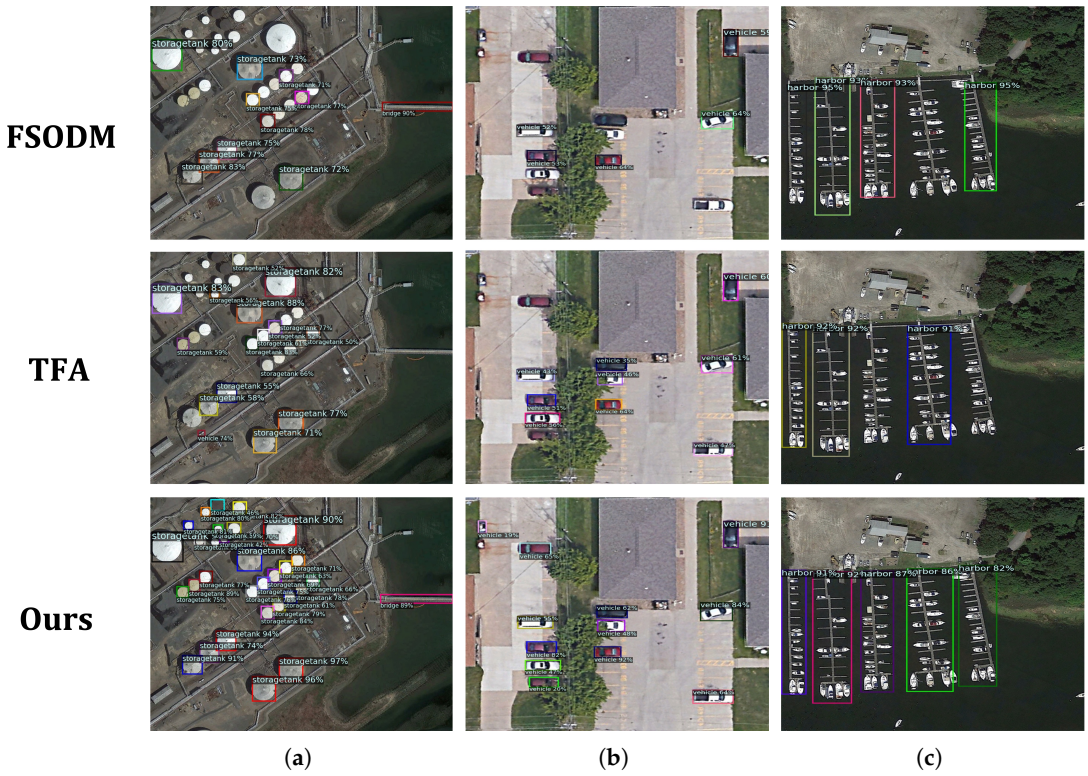

Figure 8. Qualitative inference results of base categories on the DIORdataset. (**a**) storage tank, (**b**) vehicle, (**c**) harbor.

3.5. Results on NWPU VHR-10

To further validate the advantages of our CIR-FSD, we conducted experiments on the NWPU VHR-10 dataset. Table 4 lists the detection accuracy of the proposed CIR-FSD and comparison methods on three novel categories of the NWPU dataset. Our method shows obvious advantages over the metric-based learning and meta-learning methods, as indicated in the table. Compared to baseline TFA [31], our method improves it by 25%,

9%, and 5%, respectively, in 3-shot, 5-shot, and 10-shot settings. In addition, compared with PAMS-Det, our method improves mAP by 17% in 3-shot, 9% in 5-shot and 4% in 10-shot settings.

Table 4. Comparison of the base and novel classes results on the NWPU VHR-10 dataset.

Class	Shot	YOLO v5	Faster R-CNN	RepMet ‡	FSRW †	FSODM †	TFA ⋆	PAMS-Det ⋆	Ours ⋆
Base Classes Results									
ship		0.80	0.88	-	0.77	0.72	0.90	0.88	0.91
storage tank		0.52	0.49	-	0.80	0.71	0.90	0.89	0.88
basketball court		0.58	0.56	-	0.51	0.72	0.91	0.90	0.91
ground track field		0.99	1.00	-	0.94	0.91	0.99	0.99	0.99
harbor		0.67	0.66	-	0.86	0.87	0.79	0.84	0.80
bridge		0.56	0.57	-	0.77	0.76	0.80	0.80	0.87
vehicle		0.70	0.74	-	0.68	0.76	0.81	0.89	0.89
mean		0.69	0.70	-	0.76	0.87	0.87	0.88	**0.89**
Novel Classes Results									
airplane	3	0.06	0.09	0.19	0.13	0.15	0.12	0.21	0.52
	5	0.10	0.19	0.20	0.24	0.58	0.51	0.55	0.67
	10	0.18	0.20	0.22	0.20	0.60	0.60	0.61	0.71
baseball diamond	3	0.14	0.19	0.36	0.12	0.57	0.61	0.76	0.79
	5	0.20	0.23	0.36	0.39	0.84	0.78	0.88	0.88
	10	0.28	0.35	0.39	0.74	0.88	0.85	0.88	0.88
tennis court	3	0.12	0.12	0.12	0.11	0.25	0.13	0.16	0.31
	5	0.15	0.17	0.14	0.11	0.16	0.19	0.20	0.37
	10	0.15	0.17	0.18	0.26	0.48	0.49	0.50	0.53
mean	3	0.11	0.13	0.23	0.12	0.32	0.29	0.37	**0.54**
	5	0.15	0.20	0.23	0.24	0.53	0.49	0.55	**0.64**
	10	0.20	0.24	0.26	0.40	0.65	0.65	0.66	**0.70**

† marks meta-learning based methods. ‡ marks metric-learning based methods. ⋆ marks the average of the experiments on 10 random seeds.

For the airplane category, the mAPs of all methods do not exceed 21% in the three-shot setting. In contrast, our approach can achieve an impressive 50%. It is clear that our method performs well with fewer samples. Although the TFA-based PAMS-Det uses inner convolution and PAM, it does not solve the foreground and background imbalance caused by low confidence in the novel classes, so their approach remains limited compared to ours.

Figures 9 and 10 show the qualitative inference results of FSODM [23], TFA [31] and our proposed method on the NWPU VHR-10 dataset. As shown in Figure 9a, the aircraft rotation angles in the dataset are more pronounced and diverse, which makes it harder to distinguish them from the background. Even with few samples, our method is able to use rich contextual information and improved RPN to identify the airplanes well. When there are some dense and small objects in the scene, as shown in Figure 10, such as storage tank, vehicle and ship, CIR-FSD can still use the refinement of contextual information to locate and classify them accurately. As can be seen from the figure, the detection performance of our method is significantly better than that of the compared methods.

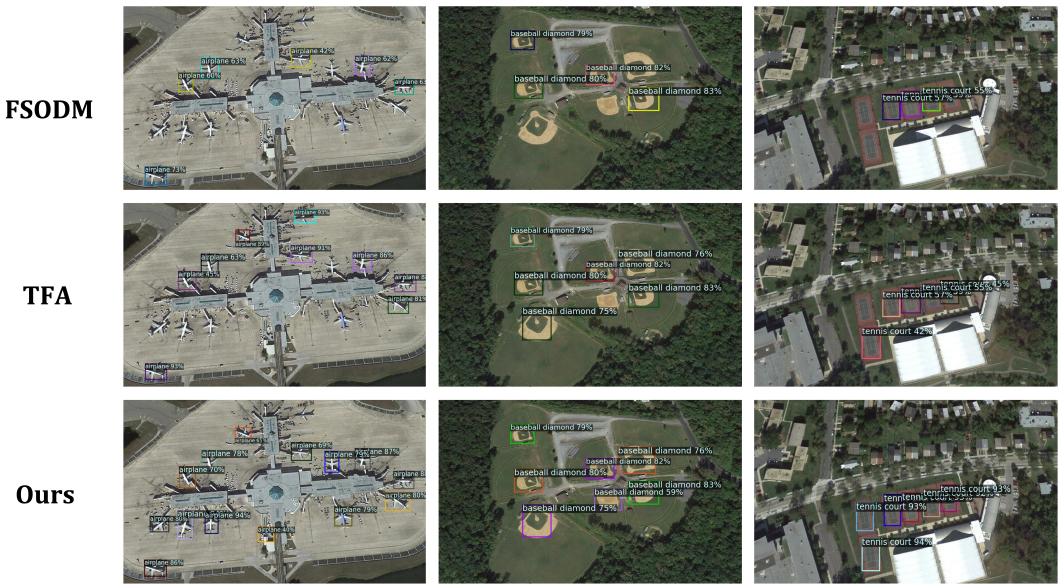

Figure 9. Qualitative inference results of novel categories on the NWPU VHR-10 dataset. (**a**) airplane, (**b**) baseball diamond, (**c**) tennis court.

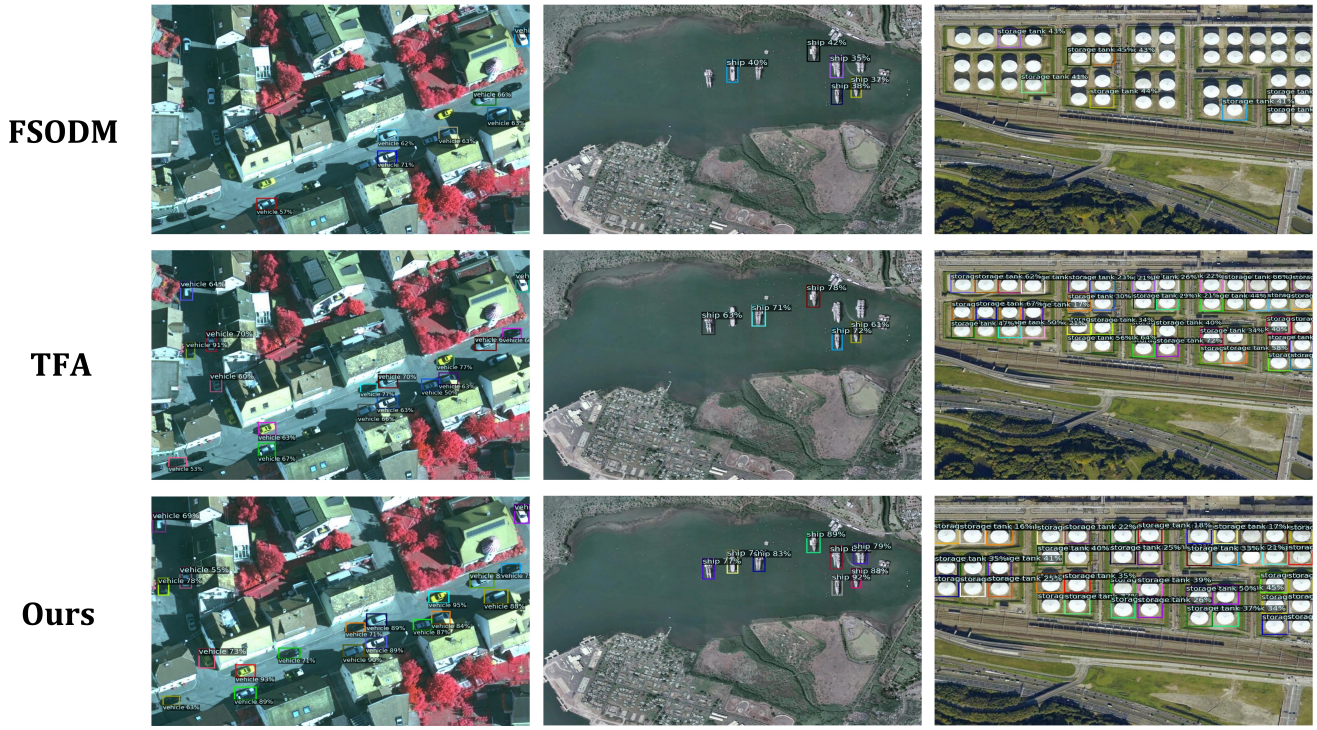

Figure 10. Qualitative inference results of base categories on the NWPU VHR-10 dataset. (**a**) vehicle, (**b**) ship, (**c**) storage tank.

3.6. Ablation Experiments

In order to further validate the effectiveness of the main components of our CIR-FSD, we performed ablation experiments on the two datasets mentioned above. As shown in

Tables 5 and 6, after adding CIR, our method achieves 3–5% improvement on the DIOR dataset, and 1–11% improvement on the NWPU VHR-10 dataset, and achieves more remarkable gain in fewer shots. By fine-tuning RPN and relaxing NMS, the mAPs for novel categories increase by 1.5–3% on the DIOR dataset and 4.6–14% on NWPU VHR-10 dataset, indicating that improved RPN alleviates the suppression for novel categories. In addition, we separately validated the performance of CIR for base categories. As shown in Table 7, after adding CIR, our method achieves significant improvement on both DIOR and NWPU VHR-10 datasets.

Table 5. Ablation experiment on the DIOR Dataset.

TFA [31]	CIR	F-RPN	R-NMS	5 Shot	10 Shot	20 Shot
✓				0.250	0.310	0.370
✓	✓			0.294	0.359	0.401
✓	✓	✓		0.310	0.362	0.415
✓	✓	✓	✓	**0.325**	**0.375**	**0.427**

Table 6. Ablation experiment on the NWPU VHR-10 Dataset.

TFA [31]	CIR	F-RPN	R-NMS	3 Shot	5 Shot	10 Shot
✓				0.290	0.490	0.650
✓	✓			0.400	0.565	0.651
✓	✓	✓		0.476	0.580	0.655
✓	✓	✓	✓	**0.539**	**0.641**	**0.697**

Table 7. Ablation experiment of Base categories.

TFA [31]	CIR	mAP on DIOR	mAP on NWPU VHR-10
✓		0.630	0.870
✓	✓	**0.738**	**0.893**

To evaluate the parameters and computational complexity of the proposed method relative to the baseline, we use the number of model parameters and floating-point operations per second (FLOPS) as evaluation metrics. The calculation of parameters and FLOPS are measured with the analysis tool of Detectron2 [43], and the experiments are performed on two datasets, DIOR and NWPU VHR-10. The input size of the datasets is 800 × 600. As shown in Table 8, our proposed method has only a small increase in model parameters and computational complexity compared to baseline. It indicates that the performance improvement brought by our method over the baseline is worth the few extra parameters.

Table 8. Model parameters and computational complexity.

Method	DIOR		NWPU VHR-10	
	#Params (M)	FLOPS (G)	#Params (M)	FLOPS (G)
TFA	58.22	154.72	58.21	154.70
Ours	63.35	159.10	63.35	158.98

To more visually verify the effect of CIR, Figure 11 visualizes the feature maps and detection results generated by the proposed method. Two types of detection objects in complex scenarios are selected: storage tanks with small and dense characteristics and airplanes with arbitrary orientation characteristics. As shown in the figure, CIR highlights the foreground and suppresses noise in the complex background as seen in the feature map of the two examples. It is seen that many potential objects that were missed in TFA are re-identified by the proposed method according to the detection results. The accurate detection of storage tanks and airplanes in the complex background demonstrates the ability of our proposed CIR to learn discriminative context features.

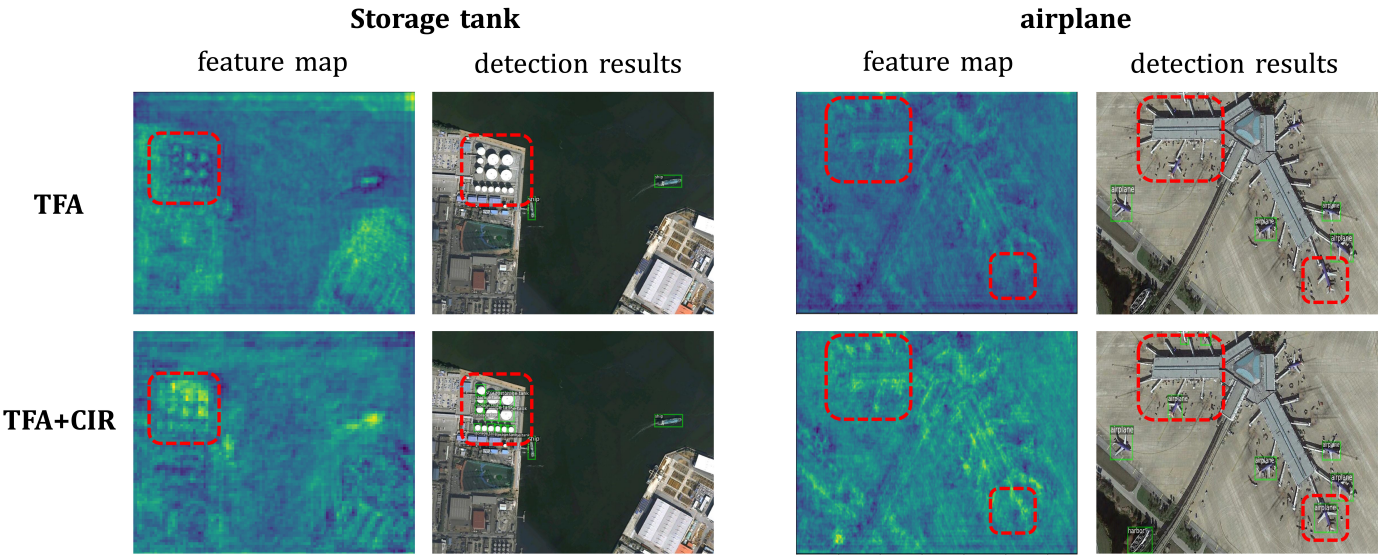

Figure 11. Feature map and the detection results of the proposed network under two examples. **Top**: feature map and detection results without CIR. **Bottom**: feature map and detection results with CIR. The green boxes are made by the detector and the red boxes are the areas we focus on.

4. Discussion

The proposed CIR-FSD was evaluated in experiments and compared with the state-of-the-art FSOD methods. Experimental results demonstrate the proposed method's efficiency on the DIOR and NWPU VHR-10 datasets.

According to the ablation experiments in Table 5, compared with the TFA [31], TFA + CIR improves mAPs for novel categories on the DIOR [49] dataset by 4.4%, 4.9%, and 3.1% in 5-shot, 10-shot, and 20-shot settings, respectively, which fully indicates that CIR extracts more robust context features conducive to object detection. After adding F-RPN, the mAPs for novel categories in 5-shot, 10-shot, and 20-shot are improved by 1.6%, 0.3% and 1.4%, respectively. Finally, after adding R-NMS, the mAPs of our proposed method are improved by 1.5%, 1.3% and 1.2% for novel categories in 5-shot, 10-shot, and 20-shot, respectively. With the above improvements, compared with the TFA, the proposed method improves mAPs for novel categories on the DIOR dataset by 7.5%, 6.5% and 5.7% in 5-shot, 10-shot, and 20-shot settings, respectively. Therefore, CIR, F-RPN and R-NMS are efficient and indispensable in the proposed method.

The comparison results of base classes are shown in Tables 3 and 4. Our method, TFA and PAMS-det [34] are based on fine-tuning, while FSRW [22] and FSODM [23] are based on meta-learning. For base categories, in most cases, the mAPs of the fine-tuning-based few-shot methods are better than that of all meta-learning-based few-shot methods. Especially on the larger dataset DIOR, the mAPs of the fine-tuning-based methods are much better than that of all meta-learning-based methods. The above analysis shows that compared with the meta-learning-based methods, the fine-tuning-based methods sacrifice less accuracy on base categories.

As can be seen from Tables 3 and 4, our CIR-FSD exceeds all competitive methods. The results of comparative experiments demonstrate the advantages of our proposed method, which are discussed separately below.

First of all, it can be seen from Tables 3 and 4 that the performance of YOLO v5 [7] and Faster R-CNN [10] without few-shot-based settings is much worse than that of the few-shot-based methods. For the DIOR dataset, YOLO v5 and Faster R-CNN can only achieve a mAP of 0.15 and 0.18 in the 20-shot setting, which is worse than the mAP of 0.33 obtained by our proposed method in 5 shots. Similarly, for NWPU VHR-10 [50] dataset, YOLO v5 and Faster R-CNN can only achieve mAP of 0.20 and 0.24 in the 10-shot setting, which is even lower than our proposed method's mAP of 0.54 in the 3-shot setting. YOLO

v5 and Faster R-CNN perform terribly in detecting objects of novel classes, demonstrating that few-shot-based methods can effectively address the challenge of novel-class object detection without a sufficient number of bounding box annotations.

Secondly, as can be seen from Tables 3 and 4, RepMet, FSRW and TFA are designed for detecting common objects in optical pictures (such as bicycles, cars, and chairs), and their performance is inferior than that designed for RSIs object detection in most cases. The fundamental reason for this is that objects in RSIs have greater scale variation and spatial resolution than that in optical images, which makes object detection with only a few annotated samples more challenging. Compared with FSRW, FSODM designs a multi-scale feature extraction module and a novel FSOD architecture to address the inherent scale variances problem in RSIs. Compared with TFA, PAMS-Det improves classification by using the involution operator and shape bias, and it creates a multi-scale module to better localization in remote sensing images.

Thirdly, FSODM, PAMS-Det and our proposed method are specially optimized for detecting objects in RSIs. PAMS-Det and our proposed CIR-FSD are based on fine-tuning, while FSODM is based on meta-learning. As you can see from Tables 3 and 4, the fine-tuning-based methods are superior to the meta-learning-based method. Compared with two methods based on fine-tuning TFA and PAMS-det, our proposed method shows superior performance. TFA and PAMS-Det only fine-tune box classifier and regressor, while our proposed method fine-tunes RPN, finetune box classifier and regressor. More importantly, the CIR module we designed extracts more robust context features, which can capture rich context from different receptive fields and enhance the object information. Take the three-shot setting as an example, the mAP of our CIR-FSD is 8% higher than TFA and 5% higher than PAMS-Det on the DIOR dataset, while the mAP of our CIR-FSD is 25% higher than TFA and 17% higher than PAMS-Det on the NWPU dataset.

To sum up, in this study, we design a novel context information refinement few-shot detector for remote image object detection. Detailed experiments and analyses show the advantages of the proposed method. Although our method can bring high improvement to horizontal region detection, it cannot address the rotation detection boundary problem. In the future, we will make our method solve the rotation detection boundary problem while dealing with the horizontal region detection.

5. Conclusions

In this study, a novel context information refinement few-shot detector on remote sensing images is proposed. We design a CIR module with the dilated convolutions and dense connections in this method. It is found that the dilated convolutions can expand the receptive fields of convolutional neural networks, and the dense connections can strengthen feature reuse; therefore, CIR can capture rich context information from different receptive fields. The designed CIR module also uses a binary map obtained from ground truth as a supervision label to refine context information, so as to enhance the discriminative capability of context features. In addition to the box classifier and regressor, we fine-tune RPN on novel categories and relax the constraint of NMS on the confidence of anchors. By fine-tuning RPN and relaxing NMS, more positive anchors for novel categories can be obtained, thus alleviating the imbalance between foreground and background. Experiments on two public benchmark data sets demonstrate that our proposed detector achieves state-of-the-art object detection performance for objects of different scales and cluttered objects in complex backgrounds.

Author Contributions: Conceptualization: Z.L. and Y.W.; methodology: C.X. and Z.L.; writing—original draft preparation: Z.L. and C.X.; writing—review and editing: Y.W. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 62171295, in part by the Liaoning Provincial Natural Science Foundation of China under Grant

2021-MS-266, and in part by the Shenyang Science and Technology Innovation Program for Young and Middle-aged Scientists under Grant RC210427.

Data Availability Statement: The experiments are evaluated on publicly open data sets. The access manner of the data sets can refer to the corresponding published papers.

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

References

1. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
2. Jiao, L.; Zhang, F.; Liu, F.; Yang, S.; Li, L.; Feng, Z.; Qu, R. A survey of deep learning-based object detection. *IEEE Access* **2019**, *7*, 128837–128868. [CrossRef]
3. 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.
4. 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.
5. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
6. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
7. Glenn, J.; Wang, C.Y.; Liao, H.Y.M. *Yolov5: v3.1—Bug Fixes and Performance Improvements*; Zenodo: Geneva, Switzerland , 2020; [CrossRef]
8. 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.
9. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
10. Ren, S.; He, K.; Girshick, R.B.; 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]
11. Wang, Y.; Bashir, S.M.A.; Khan, M.; Ullah, Q.; Wang, R.; Song, Y.; Guo, Z.; Niu, Y. Remote sensing image super-resolution and object detection: Benchmark and state of the art. *Exp. Syst. Appl.* **2022**, *197*, 116793. [CrossRef]
12. Wu, X.; Hong, D; Tian, J; Chanussot, J; Li, W; Tao, R. Orsim Detector: A Novel Object Detection Framework in Optical Remote Sensing Imagery Using Spatial-Frequency Channel Features. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5146–5158. [CrossRef]
13. 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]
14. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021.
15. 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 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019.
16. Zheng X.; Gong, T.; Lu, X. Generalized Scene Classification From Small-Scale Datasets With Multitask Learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
17. Chen, H.; Wang, Y.; Wang, G.; Qiao, Y. Lstd: A low-shot transfer detector for object detection. In Proceedings of the AAAI conference on artificial intelligence, Hilton New Orleans Riverside, New Orleans, LO, USA, 2–7 February 2018.
18. Xu C.; Zheng, X.; Lu, X. Multi-Level Alignment Network for Cross-Domain Ship Detection. *Remote Sens.* **2022**, *14*, 2389. [CrossRef]
19. Wu, X.; Li, W.; Hong, D.; Tian, J.; Tao, R.; Du, Q. Multi-Level Alignment Network for Cross-Domain Ship Detection. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 39–53. [CrossRef]
20. Köhler, M.; Eisenbach, M.; Gross, H.M. Few-Shot Object Detection: A Survey. *arXiv* **2021**, arXiv:2112.11699.
21. Huang, G.; Laradji, I.; Vazquez, D.; Lacoste-Julien, S.; Rodriguez, P. A Survey of Self-Supervised and Few-Shot Object Detection. *arXiv* **2021**, arXiv:2110.14711.
22. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019.
23. Li, X.; Deng, J.; Fang, Y. Few-Shot Object Detection on Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
24. Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; Lin, L. Meta r-cnn: Towards general solver for instance-level low-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019.
25. Cheng, G.; Yan, B.; Shi, P.; Li, K.; Yao, X.; Guo, L.; Han, J. Prototype-CNN for Few-Shot Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–10. [CrossRef]
26. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HIi, USA, 21–26 July 2017.

27. Hsieh, T.I.; Lo, Y.C.; Chen, H.T.; Liu, T.L. One-shot object detection with co-attention and co-excitation. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019 ; Volume 32.
28. Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; Bronstein, A.M. Repmet: Representative-based metric learning for classification and few-shot object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
29. Yang, Y.; Wei, F.; Shi, M.; Li, G. Restoring negative information in few-shot object detection. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 3521–3532.
30. Weinberger, K.Q.; Saul, L.K. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* **2009**, *10*, 207–244
31. Wang, X.; Huang, T.E.; Darrell, T.; Gonzalez, J.E.; Yu, F. Frustratingly Simple Few-Shot Object Detection. In Proceedings of the International Conference on Machine Learning (ICML), Virtual Event, 13–18 July 2020.
32. Wu, J.; Liu, S.; Huang, D.; Wang, Y. Multi-scale positive sample refinement for few-shot object detection. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
33. Zhang, W.; Wang, Y.X. Hallucination Improves Few-Shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021.
34. Zhao, Z.; Tang, P.; Zhao, L.; Zhang, Z. Few-Shot Object Detection of Remote Sensing Images via Two-Stage Fine-Tuning. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
35. Huang, X.; He, B.; Tong, M.; Wang, D.; He, C. Few-Shot Object Detection on Remote Sensing Images via Shared Attention Module and Balanced Fine-Tuning Strategy. *Remote Sens.* **2021**, *13*, 3816. [CrossRef]
36. Li, Y.; Zhu, H.; Cheng, Y.; Wang, W.; Teo, C.; Xiang, C.; Vadakkepat, P.; Lee, T. Few-Shot Object Detection via Classification Refinement and Distractor Retreatment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021.
37. Sun, B.; Li, B.; Cai, S.; Yuan, Y.; Zhang, C. FSCE: Few-shot object detection via contrastive proposal encoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021.
38. Luo, W.; Li, Y.; Urtasun, R.; Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
39. Cao, J.; Chen, Q.; Guo, J.; Shi, R. Attention-guided context feature pyramid network for object detection. *arXiv* **2020**, arXiv:2005.11475.
40. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT , USA, 18–21 June 2018.
41. Huang, Z.; Li, W.; Xia, X.; Wu, X.; Cai, Z.; Tao, R. A Novel Nonlocal-Aware Pyramid and Multiscale Multitask Refinement Detector for Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–20. [CrossRef]
42. Shivappriya, S.N.; Priyadarsini, M.J.P.; Stateczny, A.; Puttamadappa, C.; Parameshachari, B.D. Cascade Object Detection and Remote Sensing Object Detection Method Based on Trainable Activation Function. *Remote Sens.* **2021**, *13*, 200. [CrossRef]
43. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 1 June 2021) .
44. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
45. Wang, J.; Wang, X.; Shang-Guan, Y.; Gupta, A. Wanderlust: Online continual object detection in the real world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021.
46. Fan, Z.; Ma, Y.; Li, Z.; Sun, J. Generalized few-shot object detection without forgetting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–25 June 2021.
47. Guirguis, K.; Hendawy, A.; Eskandar, G.; Abdelsamad, M.; Kayser, M.; Beyerer, J. CFA: Constraint-based Finetuning Approach for Generalized Few-Shot Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual Event, 19–24 June 2022.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
49. 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]
50. 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]