remote sensing

MDPI

Article

Few-Shot Object Detection on Remote Sensing Images via Shared Attention Module and Balanced Fine-Tuning Strategy

Xu Huang [1], Bokun He [2], Ming Tong [2], Dingwen Wang [1,*] and Chu He [2]

[1] School of Computer Science, Wuhan University, Wuhan 430072, China; huangxu@whu.edu.cn
[2] School of Electronic Information, Wuhan University, Wuhan 430072, China; bokun.he@whu.edu.cn (B.H.); tongming@whu.edu.cn (M.T.); chuhe@whu.edu.cn (C.H.)
* Correspondence: wangdw@whu.edu.cn

Abstract: Few-shot object detection is a recently emerging branch in the field of computer vision. Recent research studies have proposed several effective methods for object detection with few samples. However, their performances are limited when applied to remote sensing images. In this article, we specifically analyze the characteristics of remote sensing images and propose a few-shot fine-tuning network with a shared attention module (SAM) to adapt to detecting remote sensing objects, which have large size variations. In our SAM, multi-attention maps are computed in the base training stage and shared with the feature extractor in the few-shot fine-tuning stage as prior knowledge to help better locate novel class objects with few samples. Moreover, we design a new few-shot fine-tuning stage with a balanced fine-tuning strategy (BFS), which helps in mitigating the severe imbalance between the number of novel class samples and base class samples caused by the few-shot settings to improve the classification accuracy. We have conducted experiments on two remote sensing datasets (NWPU VHR-10 and DIOR), and the excellent results demonstrate that our method makes full use of the advantages of few-shot learning and the characteristics of remote sensing images to enhance the few-shot detection performance.

Keywords: object detection; few-shot learning; remote sensing images; attention mechanism

Citation: 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. https://doi.org/10.3390/rs13193816

Academic Editor: Paolo Addesso

Received: 16 August 2021
Accepted: 20 September 2021
Published: 23 September 2021

1. Introduction

1.1. Background

Attribute to the rapid development of aerospace technology, the research on object detection on remote sensing images is much more than ever before, which brings both convenience and challenge to the computer vision field of remote sensing images (RSI). Object detection on remote sensing images is of great significance in many fields, such as weather forecast, disaster monitoring, urban planning, and missile navigation.

Recently, deep learning has achieved great success in object detection, since it performs much better than conventional methods, in terms of accuracy and efficiency. In the general object detection field, a large number of object detection methods based on convolutional neural networks (CNN) have been proposed.

However, most of the existing methods in the remote sensing field simply apply general object detection methods on remote sensing images, with a little improvement for detecting the small objects or rotation angles. However, these methods are not well adapted to the characteristics of remote sensing images. Moreover, a well-designed deep learning neural network can achieve good detection performance when a large number of training samples are available, but it is easy to overfit when the training set is small. Therefore, this leads the object detection on remote sensing images to a new research direction: few-shot learning, which aims at training CNN with only a few samples. This is exactly the same as the problem faced by object detection on remote sensing images, so the research on few-shot object detection on remote sensing images is much needed and indeed meaningful.

1.2. Related Work

General object detection. The research on object detection has a long history in the field of computer vision. About a decade or so ago, most studies on object detection emphasized the use of the sliding window method [1,2]. With computing power continuing to increase, the deep CNN-based methods are widely proposed in recent years. The existing CNN-based methods can be divided into two categories: proposal-free and proposal-based. Proposal-free methods regard object detection as a regression analysis of target location and category information. The most classic method is YOLO [3–5]. YOLO is the first single-stage object detection algorithm that achieves good results in terms of both detection speed and accuracy. single shot multibox detector (SSD) [6] is another proposal-free object detection algorithm that has received a lot of attention. It utilizes feature maps at multiple scales, which improves the algorithm's ability to detect objects at different scales. As for the proposal-based methods which use the proposals to improve the accuracy in object detection, R-CNN [7] is the first algorithm proposed to use the region proposal approach to solve the problem of localization in CNN. Instead of the serial feature extraction method in R-CNN, fast R-CNN [8] extracts the feature maps directly with a neural network, which improves the speed of the algorithm. However, on the basis of the former, faster R-CNN [9] achieves end-to-end training by proposing the region proposal network (RPN), and it becomes the most commonly used algorithm in object detection nowadays.

Compared with general optical images, the development of object detection on remote sensing images is much slower. Most of them are based on the former, with some adaptations for the structural characteristics of remote sensing images, such as smaller scales [10] or oriented bounding boxes [11]. The article [12] provides a comparative evaluation of recent popular CNN-based object detection models on remote sensing images, which is a detailed summary. However, for the object detection on remote sensing images with few training samples, the studies are rather few.

Few-shot learning. Few-shot learning is a recently emerging research orientation to train CNN with only a few samples. Early methods known as metric learning mainly focus on the classification problem, such as Siamese neural network [13], matching network [14], and prototypical network [15]. When faced with the challenge of the few-shot object detection, which is a more difficult task because it not only aims at classification but also localization, there are two main methods: meta-learning method and fine tuning-based method, both of them partition the training set into base and novel classes. The meta-learning methods design a framework composed of meta-training and meta-testing to train a meta-learner to transfer knowledge from base classes to novel classes. Meta R-CNN [16] aims at meta-learning over region of interest (RoI) and proposes a predictor-head remodeling network (PRN) to infer class-attentive vectors to detect novel classes. MetaDet [17] trains the category-agnostic parameters with base class samples and a weight prediction meta-model to learn category-specific parameters from few-shot samples. FSFR [18] designs a meta feature learner and a feature reweighting module based on YOLOv2, and FSODM [19] applies it in remote sensing images based on YOLOv3. F The fine tuning-based method divides the training process into base training stage and few-shot fine tuning stage. LSTD [20] is a novel framework based on transfer learning, in which proposed transfer knowledge (TK) and background depression (BD) are regularized. TFA [21] fixes the feature extractor components and fine-tunes only the last layer of detectors in faster R-CNN. MPSR [22] manually increases the scale of training sets by selecting positive samples and scaling them to different sizes. Built on the distance metric learning (DML) theory, RepMet [23] designs a network with embedding loss to train a DML embedding together with the models of the class posterior distribution.

However, as for the remote sensing images, although there have been some few-shot object classification methods such as [24–27], few studies have begun to design a network specifically for few-shot object detection. In our work, considering the characteristics of remote sensing images, we aim at this challenging task in the remote sensing field.

1.3. Problems and Motivations

Although deep CNN-based few-shot object detection methods can automatically extract more features than traditional methods, they show good performance on accuracy. However, when directly applying them on remote sensing images, there are still many problems faced, as follows:

(1) On the one hand, compared with general images, remote sensing images are much more difficult to obtain. Since conventional CNN-based methods are prone to over-fitting with only few samples, we find that the existing fine tuning-based training strategy is an effective way to solve this problem. This strategy aims at extracting category-agnostic features through the base training stage and adapting to the category-specific features through the few-shot fine tuning stage. However, when applying them on remote sensing few-shot object detection, we still have to make some adaptations for the characteristics of remote sensing images.

(2) On the other hand, remote sensing images have a different imaging mechanism than general images. Their camera angles are fixed, but due to different distances from the ground and different spatial resolutions, even the scales of the objects from the same class vary greatly. This makes it difficult for CNN-based feature extractors to learn effective features. However, considering that the remote sensing images provide richer ground information, we can design a mechanism to make better use of this information to help the backbones learn to extract more abstract and high-level features.

(3) In addition to the challenges caused by the characteristics of remote sensing images, there is another problem when using the fine tuning-based training strategy under the settings of few-shot learning: the number of novel class samples in the few-shot fine-tuning stage is much smaller than the number of base class samples in the base training stage. Moreover, this leads to the problem of class imbalance, which will also affect the convergence of the CNN and thus reduce the detection accuracy.

1.4. Contributions and Structure

In response to the problems analyzed in the previous section, we propose a few-shot object detection method designed for remote sensing images. Our network extracts the prior knowledge from the base classes to help the feature extractors adapt quickly to the novel classes. Thus, more effective features from novel classes are brought out. Then, some adjustments are made to the few-shot fine tuning stage for enhancing the performance of detecting the novel class objects with few samples. The main contributions of our article can be summarized as follows:

(1)With analyzing the reasons leading to the poor performances of the original fine tuning based methods on remote sensing images, we propose a new few-shot fine tuning based method that is more appropriate to the characteristics of remote sensing images to accommodate the few-shot object detection on remote sensing images.

(2) In order to make better utilization of the rich ground information and to adapt to the large variations in object sizes of the remote sensing images, we propose a shared attention module (SAM), with shared multi-attention maps between the base training stage and the few-shot fine-tuning stage, which facilitates the use of the category-agnostic prior knowledge extracted from the base classes to help detect the novel class objects.

(3) To solve the problem of class imbalance caused by few-shot settings, we design a new few-shot fine-tuning stage with a balanced fine tuning strategy (BFS), which adjusts the maximum number of proposals in the region proposal network (RPN) to bring out more novel class proposals. It also replaces the original loss function with a more class-balanced loss function in the few-shot fine-tuning stage to mitigate the imbalance between the number of base classes and novel classes.

As for the organization of this article, Section 2 analyses the characteristics of remote sensing images and briefly introduces the preliminaries of the fine-tuning based few-shot object detection method. In Section 3, we introduce the framework of our work and describe the proposed method in detail. Next, the description of the experimental setup and results

are shown in Section 4, and then we discuss our method and the corresponding results in detail in Section 5. Finally, the conclusion of this article is summarized in Section 6.

2. Preliminaries

2.1. Characteristics of Remote SENSING Images

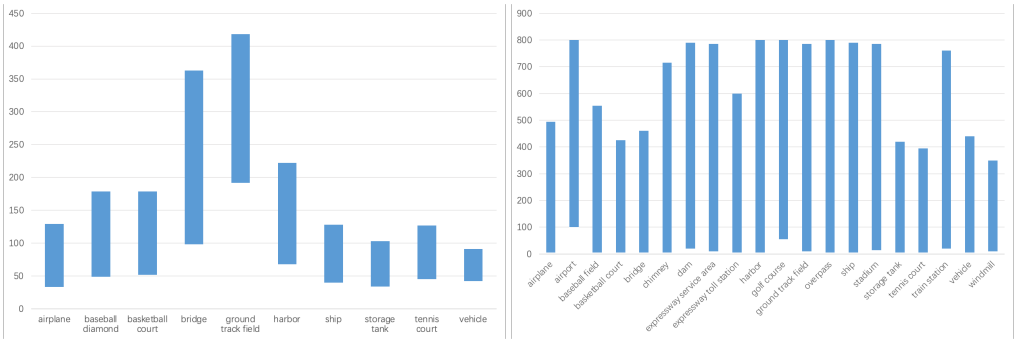
Object detection on general optical images may have problems because of different camera angles. For example, photos of vehicles taken from the front and the side look very different. However, when people take pictures with a camera, due to the distance between the targets and the cameras being about the same, the sizes of objects of the same category in the general optical images are usually similar, so the anchors in the region proposal network (RPN) with predefined sizes can predict proposals accurately.

However, the situation faced by the remote sensing images is exactly the opposite. Since remote sensing images are mostly taken from top to bottom, the difference in their camera angles is relatively small. However, due to the different shooting distances from the ground and the different spatial resolutions, even the sizes of the objects from the same class vary greatly in the remote sensing images. This problem limits the performance of anchors with predefined sizes and thus affects the accuracy of detection.

In order to better implement few-shot object detection on remote sensing images, we first need to be familiar with its characteristics. With this consideration, we analyze two widely used remote sensing image datasets NWPU VHR-10 [28] and DIOR [29]. Through specific analysis, we believe that there are two main challenges that distinguish them from few-shot object detection on general optical datasets:

(1) The size of the remote sensing image datasets is usually smaller than the general optical datasets. Thus, the number of base class samples is not enough for deep CNNs to learn the generalization ability of extracting effective category-agnostic features.

(2) As shown in Figure 1, remote sensing image datasets (e.g., NWPU VHR-10 and DIOR) have a large range of object size variations, not only between different classes but also between different objects in the same class, which makes it more difficult for the detector to distinguish between objects from different classes with few annotated samples.



(**a**) NWPU VHR-10 (**b**) DIOR

Figure 1. Object size (pixel) range per class in NWPU VHR-10 and DIOR.

2.2. Few-Shot Fine-Tuning

The fine tuning-based method is an important branch in the field of few-shot object detection first proposed by the two-stage fine tuning approach (TFA) [21]. As shown in Figure 2, this method defines the classes to be detected as novel classes and the others as base classes, and divides the training process into a base training stage and a few-shot fine-tuning stage. In the base training stage, a large number of base class images are fed into the network for training. Moreover, in the few-shot fine-tuning stage, only K-annotated images from each class (including base classes and novel classes) are randomly selected for training.

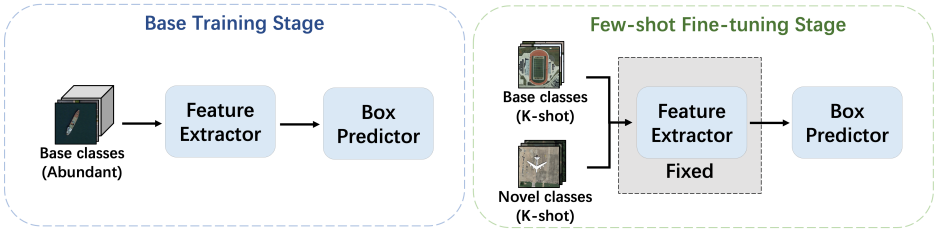


Figure 2. Framework of the Two-stage Fine tuning Approach (TFA).

TFA aims at extracting category-agnostic features from the base classes through the base training stage and adapting to the category-specific features of the novel classes through the few-shot fine tuning stage. However, when applied to few-shot object detection on remote sensing images, because of the insufficient base class samples and the large range of object size variations analyzed in the previous section, its strategy of fixing the feature extractor components in the few-shot fine tuning stage lead to limited performance. Therefore, with taking the characteristics analyzed in the previous section into consideration, we decide to design a more appropriate method for few-shot object detection on remote sensing images.

3. Proposed Method

As analyzed in the previous sections, the key challenge is that no matter how robust the algorithm is, a model trained on a limited set of samples still has difficulty detecting the novel classes that have only been seen few times. With the consideration that increasing the number of samples is against the original purpose of the few-shot learning, and to make better utilization of the characteristics of remote sensing images, the main contribution of this article is to design a network with a shared attention module (SAM) to help the model extract multi-dimensional prior knowledge from the base classes at the base training stage, thus helping the model to focus more accurately on what is worth learning in the few-shot fine tuning stage and improve its adaptability to the novel class objects. In addition, a balanced fine tuning strategy (BFS) is proposed to mitigate the imbalance between the number of samples in the base classes and the novel classes due to the few-shot settings. As a result, in our experiments, the SAM and the BFS are integrated based on the faster R-CNN [9], and their advantages are complementarily utilized to detect novel class objects on remote sensing images, with only a few annotated samples. The overall framework of the proposed method is shown in Figure 3, and the specific implementation of our method is described in the following subsections.

3.1. Shared Attention Module

In order to adapt to the characteristics of the remote sensing images analyzed in Section 2, we need to make better utilization of the rich ground information provided by remote sensing images. Considering that the channel and spatial information used for localization is category-agnostic, we propose a shared attention module (SAM) that extracts the multi-attention maps from the base classes in the base training stage. Then, the attention maps will be shared with the backbone in the few-shot fine tuning stage to enhance its capability of feature extraction. The specific structure of SAM is shown in Figure 4.

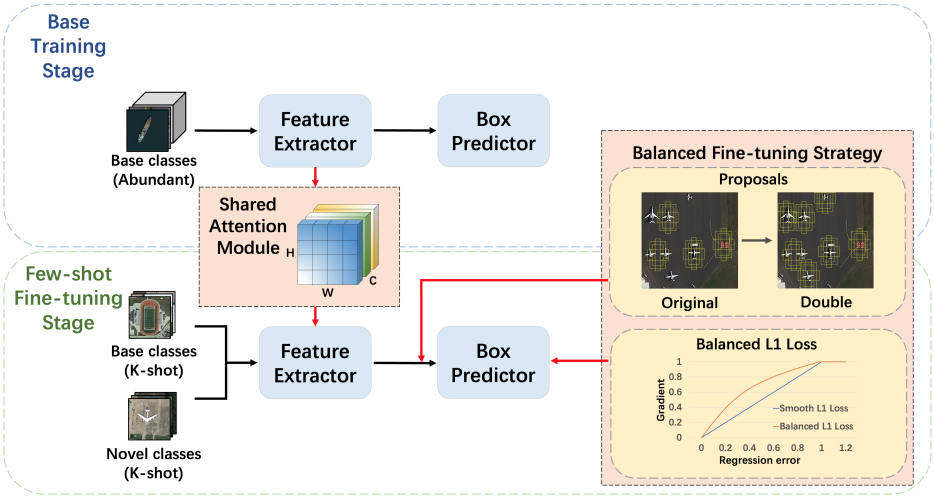


Figure 3. Framework of the proposed method for few-shot object detection on remote sensing images. Our method consists of two main components: a shared attention module (SAM) between the backbones and a balanced fine-tuning strategy (BFS). The SAM extracts the multi-dimensional attention map from the base training stage and shares it to the few-shot fine-tuning stage as prior knowledge to improve the localization of novel class objects. With the help of it, the BFS improves the performance of few-shot object classification on remote sensing images by solving the imbalance between the number of samples from base classes and novel classes.

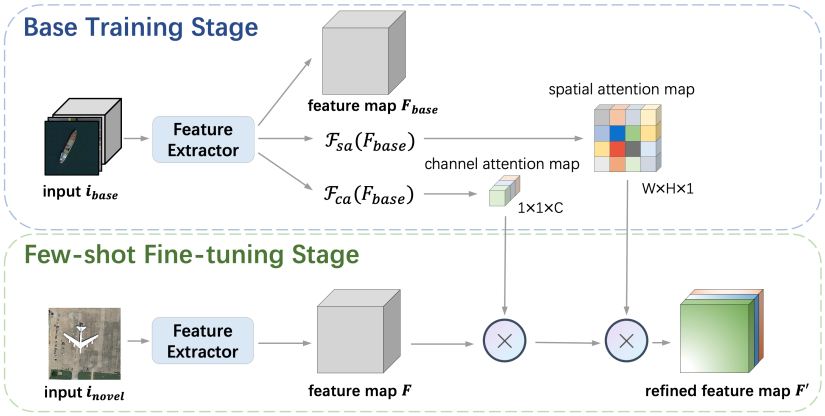


Figure 4. The architecture of our shared attention module (SAM).

As shown in Figure 5, the channel attention extractor $F_{ca}()$ performs both average-pooling and max-pooling on the input feature map F_{base}. Average-pooling places more emphasis on downsampling the overall feature information, which helps to gather more complete information of a dimension and learn the extent of the target object effectively. As for max-pooling, it gathers important features of distinctive objects by selecting features with higher recognition. Then, the average-pooled features and max-pooled features are applied in a shared full convolution module, which contains a downscales weight matrix W_d and an upscales weight matrix W_u. Next, we merge the output feature maps using an element-wise summation denoted as *Concat*. Finally, we get the channel attention map of size $1 \times 1 \times C$ after the calculation by a sigmoid function, where C denotes the number of channels.

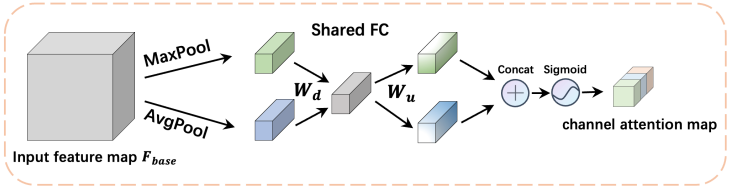


Figure 5. Diagram of channel attention extractor $F_{ca}()$.

The diagram of the spatial attention extractor is shown in Figure 6. In the spatial attention extractor $F_{sa}()$, average-pooling and max-pooling are performed on the input feature map F_{base} at first. Then, we concatenate the average-pooled features and max-pooled features using an element-wise summation. After this, we apply a convolution layer followed by a sigmoid function to generate the spatial attention map of size $W \times H \times 1$, where W and H denote the width and height of the feature map, respectively.

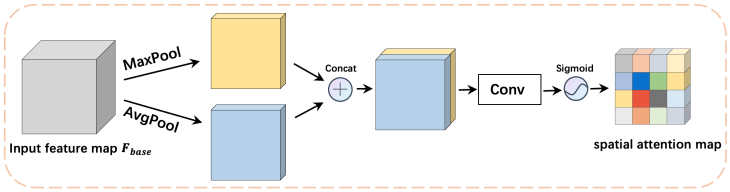


Figure 6. Diagram of spatial attention extractor $F_{sa}()$.

Inspired by CBAM [30], we set the formula of the channel attention extractor as:

$$F_{ca}(F) = sig(FC(AvgPool(F)) \oplus FC(MaxPool(F))) \tag{1}$$

where sig denotes the sigmoid function and $\oplus$ denotes the concatenate operation, and FC is a full convolution module, which contains a downscales weight matrix W_d and an upscales weight matrix W_u. The formula of the spacial attention extractor is set as:

$$F_{sa}(F) = sig(Conv(AvgPool(F) \oplus MaxPool(F))) \tag{2}$$

in which sig denotes the sigmoid function and $\oplus$ denotes the concatenate operation, and a convolution layer denoted as $Conv$ is applied after concatenate.

We combine the channel attention extractor and spatial attention extractor into ResNet [31] sequentially. In the base training stage, the features extracted from the abundant base class samples by ResNet are fed into SAM and then calculated to generate a channel attention map of size $1 \times 1 \times C$ and a spatial attention map of size $W \times H \times 1$, where C denotes the number of channels, W and H denote the width and height of the feature map, respectively. Then, in the few-shot fine-tuning stage, only K-shot samples of each class are randomly selected for training, which is definitely not enough for training a deep CNN. So, the SAM shares the channel attention map $F_{ca}(F_{base})$ and the spatial attention map $F_{sa}(F_{base})$ extracted from the base class samples with the feature extractor in the few-shot fine-tuning stage, and multiplies the feature map F of novel input i_{novel} by the multi-attention maps to get the refined feature map F' as Equation (3), in which $\otimes$ denotes the element-wise multiplication.

$$F' = F \otimes F_{ca}(F_{base}) \otimes F_{sa}(F_{base}) \tag{3}$$

This refined feature map F' containing the rich ground information of channel and spatial attention is then fed into the RPN to help with the adaptation to the small number of novel class samples. Due to the large object size variations, conventional RPN is difficult to make effective proposals for novel class targets which may have scales it has never seen before. With the help of the channel and spatial attention maps extracted from abundant base class samples, which have the same problem of the large object size variations, our

model is more adaptable to this issue. With the prior knowledge of rich ground information, the backbone in the few-shot fine tuning stage is able to extract more effective features, thus improving our model's adaptability to localizing the novel class objects under the few-shot settings.

3.2. Balanced Fine-Tuning Strategy

In the previous subsection, a module named SAM is proposed to improve the accuracy of localization, and in this subsection, we will introduce another important part of our method termed balanced fine tuning strategy (BFS), which is designed to solve the problems in classification. Under the few-shot settings, abundant samples from base classes are used in the base training stage, but in the few-shot fine-tuning stage, the network is trained with only k images from each class. This leads to a severe imbalance between the number of samples from the base classes and the novel classes, resulting in generally lower confidence scores for the novel classes. However, the maximum number of the proposals generated by RPN is fixed, which means that many proposals of novel class objects will be filtered out because they do not have a high enough confidence score. Thus, in the few-shot fine tuning stage with our BFS, the maximum number of RPN's proposals is doubled to bring more candidate boxes that may belong to the objects of novel classes, as shown in Figure 7.

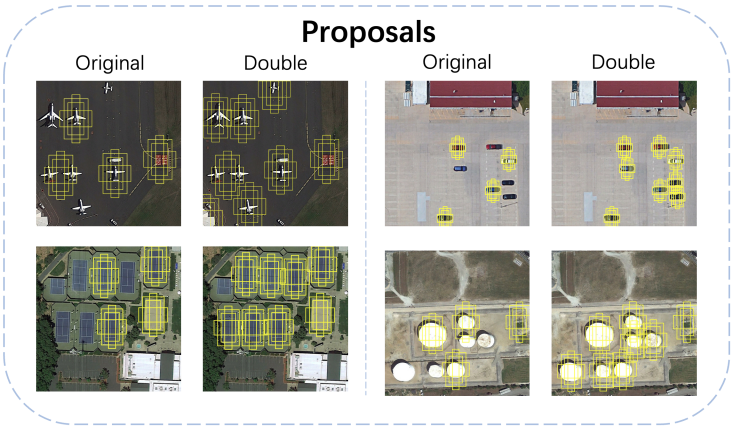


Figure 7. A schematic diagram of double proposals.

With the help of the doubled maximum proposals number, RPN is able to propose a sufficient number of foreground proposals of novel class objects. However, we observe that these novel foreground proposals are mostly accurately localized, but often incorrectly classified to confusable base classes. Through some analytical experiments, we figure out that this is caused by the severe imbalance between the number of the base class samples and the novel class samples. The original smooth L1 loss in the faster R-CNN uses the difference between the prediction bounding box and ground truth to limit the gradient.

$$L_s(x) = \begin{cases} 0.5x^2 & if|x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{4}$$

As the formula shown in Equation (4), the gradient of smooth L1 loss is guaranteed not to be too large when the difference is large. Furthermore, when the difference is small, the gradient can also be small enough. However, under the control of smooth L1 loss, all the classes have an equal influence on the value of the loss. This leads to a new problem when faced with the class imbalance: with the effect of the abundant base class samples, the loss of the novel class samples is almost negligible. As a result, the novel foreground proposals are more likely to be misclassified as base classes than novel classes.

Therefore, in order to solve this problem, we propose to boost the influence of the novel class objects on the loss, which have a small number of samples. Inspired by [32], we replace the smooth L1 loss in the few-shot fine-tuning stage with a more appropriate loss

function: balanced L1 loss. In Figure 8, we compare balanced L1 loss to several commonly used loss functions, such as L1 loss, L2 loss and smooth L1 loss. The key idea of balanced L1 loss is to boost the regression gradient of the crucial samples. To be specific, it defines the samples with the value of loss less than 0.1 as inliers and the remaining samples as outliers. Because the regression targets are un-bounded, directly promoting the localization loss will make the model more sensitive to outliers. However, the balanced L1 loss rebalances the involved samples with its parameters, thus achieving a more balanced training. With the help of it, we increase the gradient from the novel class samples defined as inliers. The formula of balanced L1 loss is defined as:

$$L_b(x) = \begin{cases} \frac{\alpha}{b}(b|x|+1)ln(b|x|+1) - \alpha|x| & if|x| < 1 \\ \gamma|x| + C & otherwise \end{cases} \quad (5)$$

in which γ controls overall promotion magnification, α increases more gradient for inliers, and b is used to ensure that $L_b(x = 1)$ has the same value for both formulations in Equation (5). The parameters γ, α and b are constrained by:

$$\alpha ln(b+1) = \gamma \quad (6)$$

The default values of parameters are set as $\alpha = 0.5$ and $\gamma = 1.5$ in our experiments.

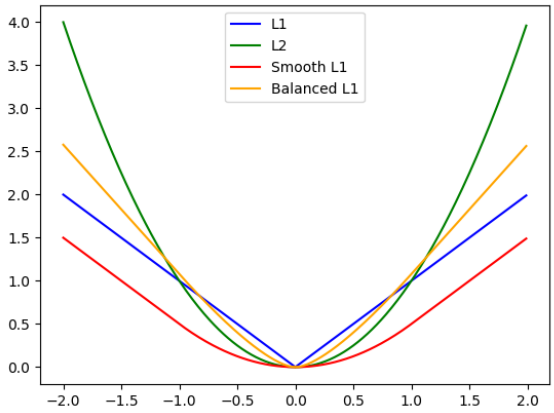


Figure 8. Diagram of several loss functions.

With the above modifications, our BFS implements a more class-balanced training than the regular fine-tuning strategy by combining adjustment to the RPN with improvement to the loss function. As a result, our method brings more foreground proposals and corrects the misclassification of them, thus solving the class imbalance problem under the few-shot settings and improving the accuracy of detection.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Dataset

NWPU VHR-10 [28] contains 800 VHR optical remote sensing images manually annotated by experts. In this dataset, 715 color images are cropped from Google Earth with the spatial resolution ranging from 0.5 to 2 m, and 85 infrared images are acquired from the ISPRS Vaihingen dataset [33] with the spatial resolution of 0.08 m. There are 10 categories in NWPU VHR-10 as shown in Figure 9: airplane, baseball diamond, basketball court, bridge, ground track field, harbor, ship, storage tank, tennis court and vehicle. These images are rectangles with long sides ranging from about 500 to 1200 pixels and are divided into two sets. The negative image set contains 150 images that do not contain any objects of the given object categories, and the positive image set contains 650 images. Each image contains at least one object to be detected.

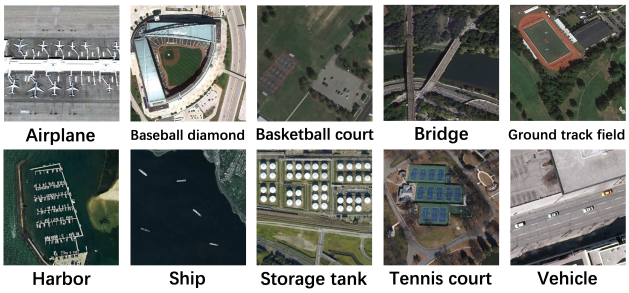


Figure 9. A set of sample images for each class in NWPU VHR-10.

DIOR [29] is a large-scale dataset for object detection with 23,463 optical remote sensing images and 192,472 instances, where the train set contains 5862 images, the valuation set contains 5863 images and the test set contains 11,738 images. The images are all squares in the size of 800 × 800 pixels and the spatial resolutions range from 0.5 to 30 m. DIOR contains 20 categories, including airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf field, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle and windmill, as shown in Figure 10.

Figure 10. A set of sample images for each class in DIOR.

4.1.2. Implementation Details

To evaluate the detection performance on novel classes, we follow the same division on base and novel classes according to the existing work [19], i.e., 3 novel classes (airplane, baseball diamond, and tennis court) in NWPU VHR-10, and 5 novel classes (airplane, baseball field, tennis court, train station, and windmill) in DIOR. On the NWPU VHR-10 dataset, we base-train on all images that do not contain any object of novel classes, and then randomly select a very small train set containing both base and novel classes for fine-tuning, in which each class only has K-annotated images, where K equals 1, 2, 3, 5 and 10. Then, we test our method on all images containing at least one novel class object except those already used for training, and this test set consists of about 300 images. Similarly, on the DIOR dataset, we use all images of base classes in its train set for base-training. Considering that DIOR is a large-scale dataset, the value of K in the train set for fine-tuning is set to 5, 10, and 20. Then, the performance is tested on the entire DIOR's valuation set, which consisted of 5863 images.

As for the parameter settings for training, the learning rate is set to 0.02 and the step strategy is used (i.e., multiply the learning rate by 0.1 at 16 and 22 epochs, respectively). The images in the NWPU VHR-10 dataset are resized to 1024 × 1024 pixels, where the

length of the long side is 1024 pixels, and the images are kept at the original ratio. For the images in the DIOR dataset, sizes remain the same as the original 800 × 800 pixels. To visualize the results, we evaluate the K-shot mean average precision (mAP) on novel classes of each method.

4.1.3. K-Shot Evaluation Metrics

When evaluating few-shot object detection algorithms, the conventional approach of computing mAP for all classes is not appropriate. Regarding the base class images as auxiliary information, few-shot object detection algorithms are only concerned with the performance of the novel class objects. In the few-shot fine tuning stage, few-shot object detection algorithms are fine-tuned with a small image set containing both base and novel classes, in which each class only has K annotated images. We term these small image set as K-shot image sets, where K is typically equal to 1, 2, 3, 5, 10, or 20 depending on the size of the dataset.

In the K-shot evaluation metrics, we have the following definitions, which are the same as the conventional evaluation metrics for detection:

True Positives (*TP*): Correctly predicted positive samples.
False Positives (*FP*): Falsely predicted positive samples.
True Negatives (*TN*): Correctly predicted negative samples.
False Negatives (*FN*): Falsely predicted negative samples.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

Moreover, in order to better evaluate the few-shot object detection algorithms, the models trained on K-shot image sets are tested on the test sets. As a result, we compute the mAPs of the novel classes with different K values and term them K-shot mAPs. Similar to the conventional mAP, we define the formula for the K-shot mAP as:

$$mAP^K = \frac{\sum_{i=1}^{n} AP_n^K}{n} \tag{9}$$

in which n denotes the number of novel classes and AP^K denotes the average precision of K-shot model, which is computed as:

$$AP^K = \int_0^1 p^K(r)\mathrm{d}r \tag{10}$$

in which r denotes the recall and p^K denotes the precision of K-shot model.

4.1.4. Comparing Methods

Since our method is based on faster R-CNN [9], we first set up two baselines without fine-tuning built on faster R-CNN, and the total training epoch of these two methods is consistent with ours.

Scratch: The first baseline is termed scratch. We directly train the faster R-CNN on a K-shot image set, in which both the base classes and the novel classes have only K images.

Joint: Another baseline joint simply trains the faster R-CNN on the same image set as our method, which contains large amounts of base class images and only K images from each novel class.

TFA: In addition, we implement another faster R-CNN-based method which only fine-tunes the last layers of the detector, which is defined in [21], and we term it TFA. We set both the base training epoch and the fine tuning epoch to 24 to be the same as ours.

FR: FR [18] is the first few-shot object detection method based on YOLOv2 [4], which proposed a meta-learning module named feature reweighting. It is re-implemented on remote sensing images by the work [19].

FSODM: To prove the excellence of our method, we also compare our method with the current state-of-the-art few-shot object detector based on the feature reweighting module in the remote sensing field named FSODM [19].

4.2. Results

For those methods based on Faster R-CNN, we conduct experiments adopting ResNet50 and ResNet101 as backbone respectively on the open-source framework MMDetection [34]. We only list the results of comparing methods using ResNet101 because they perform better than those using ResNet50 with the same training epoch. We display our performance on both ResNet50 and ResNet101 to show the excellent adaptability of our method.

Due to the lack of source code and the limitations of our experimental equipment, we directly quote the results of FSODM and the original feature reweighting method FR from the work [19]. Because it does not conduct the 1-shot and 2-shot experiments on the NWPU VHR-10 dataset, we do not list them in the table either. However, the results of 1-shot and 2-shot in our method show that our algorithm works well, even with very few samples.

Results on NWPU VHR-10. As shown in Table 1, the results of scratch indicate that a few-shot training set is too small for the faster R-CNN to converge. Therefore, the mAPs are significantly low. When applied to the remote sensing images dataset NWPU VHR-10, the small number of base classes samples and the large inter-class differences make the results of TFA even worse than that of joint. Our method performs excellently, although, without comparing the mAPs in 1-shot and 2-shot settings, it outperforms the existing state-of-the-art method FSODM in 3-shot, 5-shot, and 10-shot settings (+14.7%, +8.9%, and +9.6%, respectively).

Results on DIOR. We increase the value of K to 5, 10, and 20 when testing our method on DIOR, with the consideration that DIOR is a large dataset with tens of thousands of remote sensing images. Table 2 shows that TFA performs better when the training set is large. Sufficient base class samples enable faster R-CNN's backbone to be trained adequately, and the process of fine-tuning allows the detector of TFA to adapt to the novel class objects speedily. However, we still demonstrate the superiority of our method. We get 13.7% mAP higher in the 5-shot setting than FSODM, 14.7% and 13.4% mAP higher in the 10-shot and 20-shot setting than TFA.

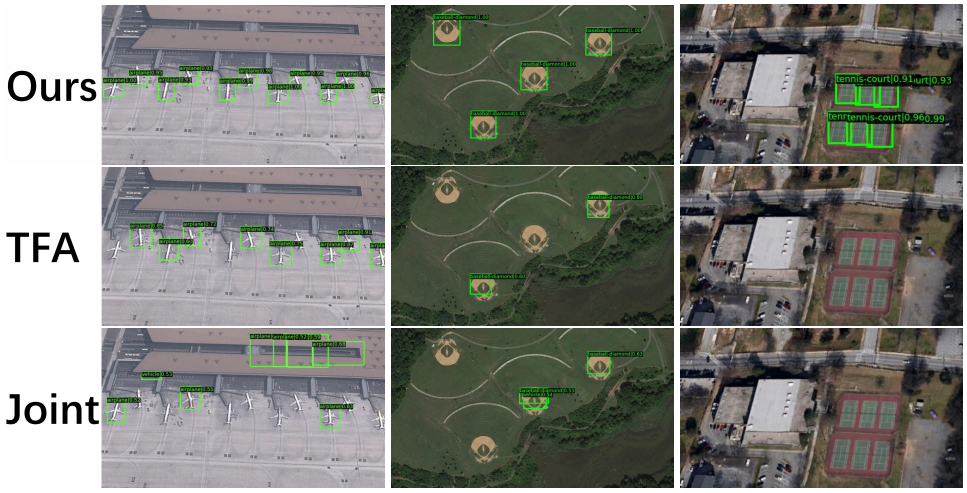
We visualize some comparisons between TFA, joint and our method on detecting novel class objects in Figures 11 and 12. Furthermore, in order to compare the results of each few-shot object detection algorithm more visually, we display the mAP values of K-shot in the form of a bar chart. As shown in Figure 13, our method achieves a great performance with the highest mAP at different values of K. More notably, the improvement is more obvious at smaller K values. This proves that our method can adapt well to the task of few-shot object detection on remote sensing images.

Table 1. Few-shot detection performance on novel classes of the NWPU VHR-10 dataset.

Method	Backbone	mAP				
		1-Shot	2-Shot	3-Shot	5-Shot	10-Shot
Scratch	ResNet101	0.0	1.7	6.3	15.1	34.3
Joint	ResNet101	4.3	13.3	20.9	33.0	48.8
TFA	ResNet101	3.3	10.1	14.1	16.1	25.3
FR [19]	YOLOv2	-	-	12.0	24.7	40.0
FSODM [19]	YOLOv3	-	-	32.3	52.7	65.3
Ours	ResNet50	16.2	33.3	45.6	60.4	**75.1**
Ours	ResNet101	**16.4**	**36.3**	**47.0**	**61.6**	74.9

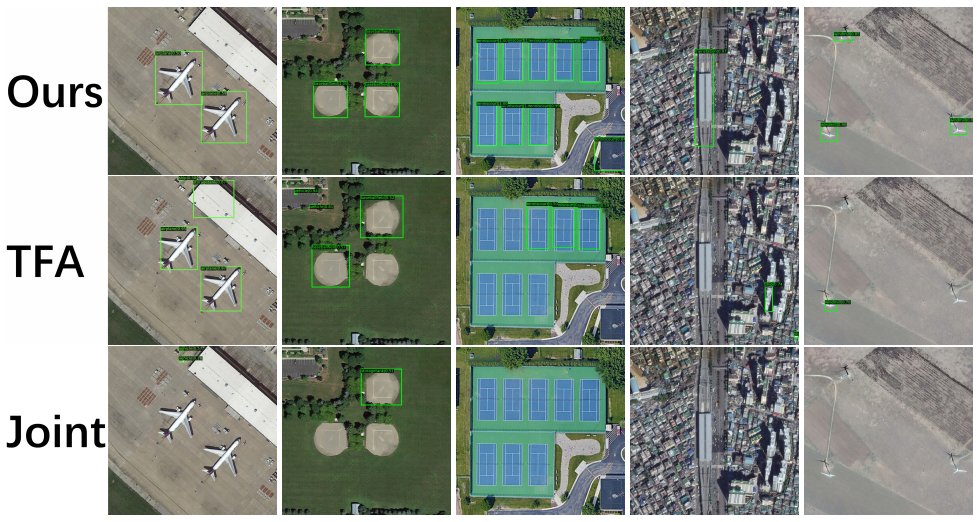
Table 2. Few-shot detection performance on novel classes of the DIOR dataset.

Method	Backbone	mAP		
		5-Shot	10-Shot	20-Shot
Scratch	ResNet101	8.2	16.2	28.4
Joint	ResNet101	17.7	28.1	32.9
TFA	ResNet101	24.5	32.6	37.5
FR [19]	YOLOv2	22.2	27.8	34.0
FSODM [19]	YOLOv3	24.6	32.0	36.6
Ours	ResNet50	**38.4**	46.5	50.1
Ours	ResNet101	38.3	**47.3**	**50.9**



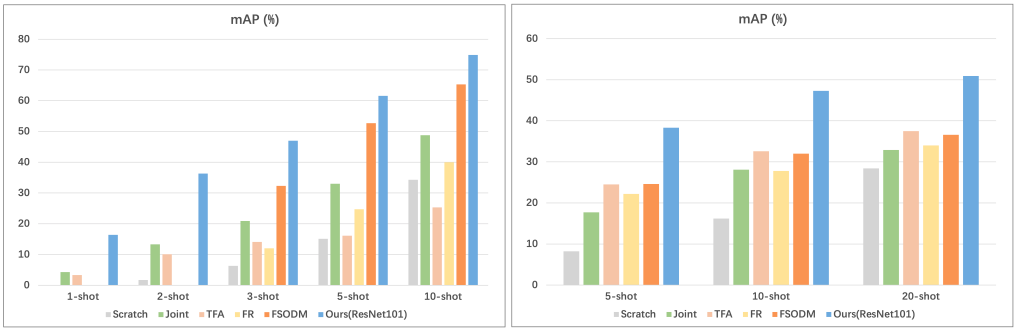
(**a**) Airplane (**b**) Baseball diamond (**c**) Tennis court

Figure 11. Results of 10-shot models on novel classes in the NWPU VHR-10 dataset. (From top to bottom: Ours, TFA and Joint).



(**a**) Airplane (**b**) Baseballfield (**c**) Tennis court (**d**) Train station (**e**) Windmill

Figure 12. Results of 20-shot models on novel classes in the DIOR dataset. (From top to bottom: Ours, TFA and Joint).



(**a**) results on NWPU VHR-10 (**b**) results on DIOR

Figure 13. Few-shot detection performance on novel classes.

There are some examples of our detection performance on the novel classes in the NWPU VHR-10 dataset and the DIOR dataset. It can be seen in Figures 14 and 15, despite a slight offset when localizing the objects with too large or too small size (such as train station and windmill), our method works well in detecting the objects of novel classes with only a few samples.

(**a**) Airplane (**b**) Baseball diamond (**c**) Tennis court

Figure 14. Results of our 10-shot model on novel classes in the NWPU VHR-10 dataset.

4.3. Ablation Studies

To demonstrate the effectiveness of each part in our method more concretely, we have conducted three ablation experiments for the main components of our method:

(1) Whether to use the single-stage training strategy or the few-shot fine-tuning strategy (FS-ft) for the few-shot training.

(2) Insert the shared attention module (SAM).

(3) Modify the few-shot fine-tuning stage with the balanced fine tuning strategy (BFS).

The results in Tables 3 and 4 show that the few-shot fine tuning strategy is significantly better than the single-stage training strategy, and on the basis of this, our SAM brings a great performance improvement, especially when the value of K is small (such as 1 and 2). In addition, the BFS demonstrates its capability by increasing each mAP by 0.5% to 4.5%. The results of the ablation experiments demonstrate that each part of our method plays an important role in improving the performance of few-shot object detection on remote sensing images.

(**a**) Airplane (**b**) Baseballfield (**c**) Tennis court (**d**) Train station (**e**) Windmill

Figure 15. Results of our 20-shot model on novel classes in the DIOR dataset.

Table 3. Ablation study on the NWPU VHR-10 dataset.

FS-ft	SAM	BFS	mAP				
			1-Shot	2-Shot	3-Shot	5-Shot	10-Shot
			4.3	13.3	20.9	33.0	48.8
✓			3.8	19.1	32.0	45.4	62.0
✓	✓		15.9	33.2	42.5	57.5	71.5
✓	✓	✓	**16.4**	**36.3**	**47.0**	**61.6**	**74.9**

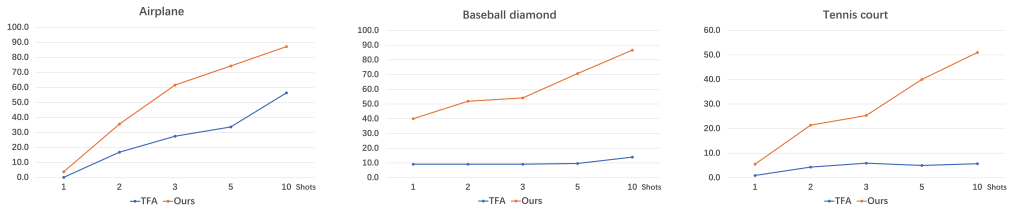
Table 4. Ablation study on the DIOR dataset.

FS-ft	SAM	BFS	mAP		
			5-Shot	10-Shot	20-Shot
			17.7	28.1	41.9
✓			33.0	43.6	46.9
✓	✓		35.5	44.5	49.8
✓	✓	✓	**38.3**	**47.3**	**50.9**

5. Discussion

5.1. Performance on Novel Classes

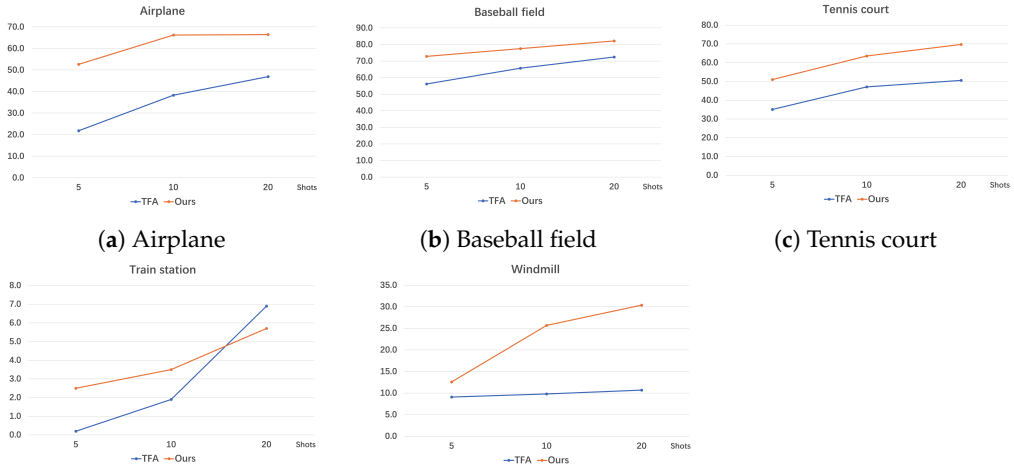
We specifically test the performances of TFA [21] and our method on novel class object detection. Moreover, for better presentation, we display them as line charts shown in Figures 16 and 17. By analyzing the detecting precision of each novel class, we can summarize that for the targets like airplanes, which have a distinct appearance and are easily distinguished from the background, both our method and TFA have good accuracies in the two datasets, and our method obviously outperforms TFA a lot.



(**a**) Airplane (**b**) Baseball diamond (**c**) Tennis court

Figure 16. Precision of few-shot detection on each novel class in NWPU VHR-10.

As for the baseball field and the tennis court, which are easily confused with other sports fields, we find that TFA performs poorly in NWPU VHR-10. We believe that the main reason is that the number of base class samples in NWPU VHR-10 is really small, which is definitely not enough for TFA to train a robust enough feature extractor. So, its strategy of the fixed feature extractor leads to poor performance. On the contrary, our method performs well, both in NWPU VHR-10 and DIOR. This is due to the multi-attention maps brought by the shared attention module (SAM), which help the feature extractor in our method to accurately focus on the novel class targets, even under the condition of few samples.



(**a**) Airplane (**b**) Baseball field (**c**) Tennis court

(**d**) Train station (**e**) Windmill

Figure 17. Precision of few-shot detection on each novel class in DIOR.

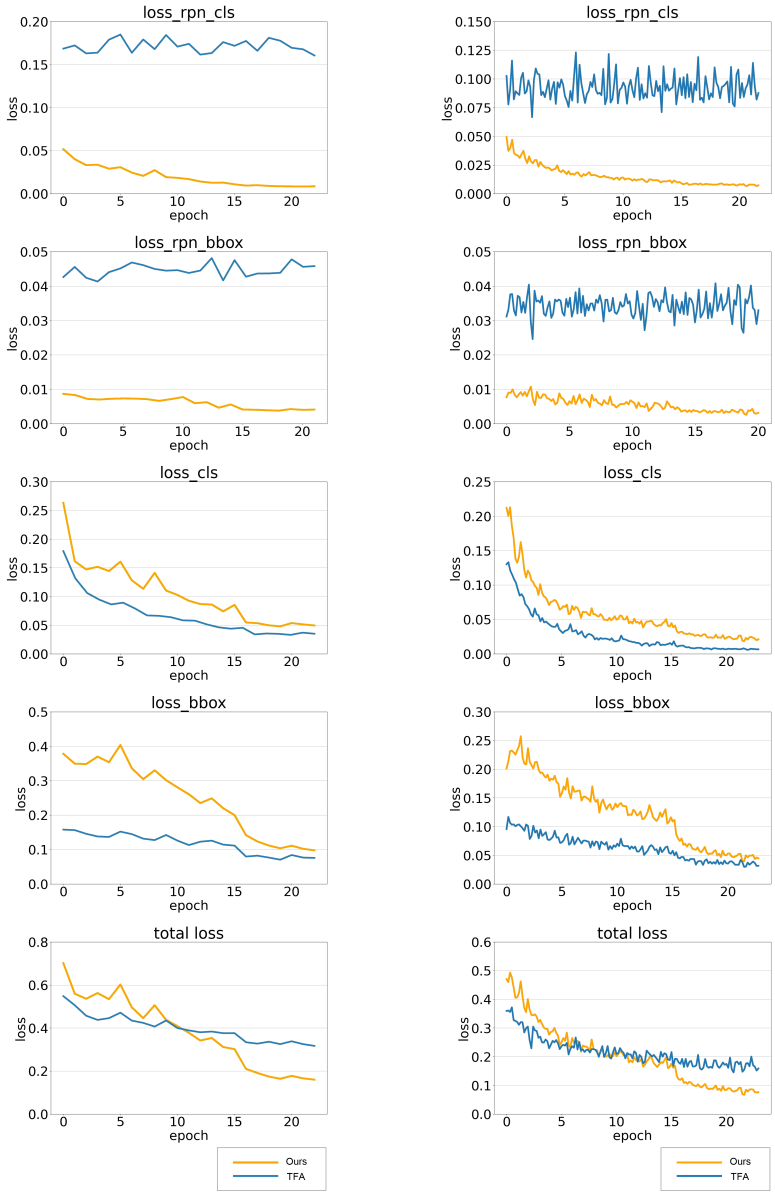
Moreover, the detection of windmills is a difficult task in few-shot object detection on remote sensing images. Windmills usually have small sizes and confusing backgrounds. TFA suffers from serious noise interference in detecting objects like windmills, resulting in poor precision. However, the excellent performance of our method demonstrates that the SAM helps us adapt well to the large variations of object sizes. In addition, the BFS also works well with a doubled number of proposals and a more balanced focus on the influence on gradients of the samples from novel classes.

However, the detection of objects such as the train station is still a weakness of the few-shot object detection algorithm. Train stations are generally located between dense objects, such as tracks, containers and trains. This makes the object detection on them susceptible to interference. When viewed from the top, they look very similar to the roofs of general buildings, while the objects from other categories usually have more distinctive features. Moreover, due to the large difference between its appearance with the base class samples, the detecting precision of both TFA and our method is relatively low, which requires further study.

5.2. Visualization of the Loss Curves

In faster R-CNN [9], the loss function works in two parts: RPN and fast R-CNN. Both of them compute the loss of classification and localization, which are denoted as

loss_rpn_cls, *loss_rpn_bbox*, $loss_{cls}$ and $loss_{bbox}$, respectively. By analyzing the loss curves of TFA [21] and our method shown in Figure 18, we can discover the problems faced by TFA when applied to remote sensing images and clearly see how our balanced fine tuning strategy (BFS) improves the performance of few-shot object detection.



(**a**) NWPU VHR-10 (**b**) DIOR

Figure 18. Loss curves of 10-shot models for NWPU VHR-10 and 20-shot models for DIOR. (From top to bottom: loss of classification in RPN, loss of localization in RPN, loss of classification in fast R-CNN, loss of localization in fast R-CNN, and total loss).

We display the loss curves during the few-shot fine-tuning stage of TFA and our method in the above figures. Considering the large number of base class samples in general optical images, TFA proposes a fine-tuning strategy with a fixed feature extractor. To be specific, its backbone and RPN are frozen in the few-shot fine tuning stage, which gives it the relatively low initial values of *loss_cls* and *loss_bbox*. However, on the contrary, the RPN in TFA cannot propose enough novel class proposals due to the insufficient base training, which leads to its *loss_rpn_cls* and *loss_rpn_bbox* hardly decreasing. While in our method, with the help of the shared attention module (SAM), our feature extractor can

learn more effective features to train the RPN and the box predictor. Thus, our *loss_rpn_cls* and *loss_rpn_bbox* keep decreasing smoothly during the few-shot fine tuning stage.

In addition, after modifying to a more class-balanced loss function, our performances of novel class classification and localization have both been improved. This makes our *total_loss* decrease more rapidly than TFA. Finally, after 24 epochs of few-shot fine tuning, we get a value of *total_loss* that is almost half of TFA, which means that our method is more appropriate for few-shot object detection on remote sensing images.

6. Conclusions

In this paper, a two-stage few-shot fine tuning method with the shared attention module (SAM) and the balanced fine tuning strategy (BFS) is proposed in order to realize few-shot object detection on remote sensing images. We firstly analyze the characteristics of remote sensing images and the reasons leading to the poor performances of existing methods in detecting the remote sensing objects. Then, to better adapt to the remote sensing objects with few samples and large size variations, we propose an effective module SAM, which learns multi-attention maps from abundant base class samples and shares it as a prior knowledge with the few-shot fine tuning stage to help extract more effective features from novel class objects. In addition, we design a new few-shot fine tuning stage with the BFS, which helps to mitigate the severe imbalance between the number of novel class samples and base class samples caused by the settings of few-shot learning. Thus, the accuracies of both localization and classification have been improved. As a result, the SAM and the BFS are integrated into faster R-CNN, and we have conducted experiments on two commonly used remote sensing image datasets NWPU VHR-10 [28] and DIOR [29]. The results indicate that our method effectively utilizes the rich ground information of remote sensing images and can achieve good performance of few-shot object detection on remote sensing images.

Author Contributions: Conceptualization, X.H. and C.H.; Methodology, X.H.; Writing—original draft, X.H.; Writing—review and editing, X.H. and B.H.; Software, X.H. and M.T.; Supervision, D.W.; Project administration, D.W. and C.H.; Funding acquisition, C.H. and D.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2016YFC0803000) and the National Natural Science Foundation of China (No. 41371342).

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

Abbreviations

The following abbreviations are used in this manuscript:

RSI	Remote sensing image
TFA	Two-stage fine tuning approach
CNN	Convolutional neural network
RPN	Region proposal network
RoI	Region of interest
mAP	Mean average precision
SAM	Shared attention module
BFS	Balanced fine tuning strategy

References

1. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I.
2. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
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; pp. 779–788.

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; pp. 7263–7271.
5. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
7. 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.
8. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
10. Guo, W.; Yang, W.; Zhang, H.; Hua, G. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. *Remote Sens.* **2018**, *10*, 131. [CrossRef]
11. Li, C.; Luo, B.; Hong, H.; Su, X.; Wang, Y.; Liu, J.; Wang, C.; Zhang, J.; Wei, L. Object detection based on global-local saliency constraint in aerial images. *Remote Sens.* **2020**, *12*, 1435. [CrossRef]
12. Alganci, U.; Soydas, M.; Sertel, E. Comparative research on deep learning approaches for airplane detection from very high-resolution satellite images. *Remote Sens.* **2020**, *12*, 458. [CrossRef]
13. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015; Volume 2.
14. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3630–3638.
15. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4080–4090.
16. 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; pp. 9577–9586.
17. Wang, Y.X.; Ramanan, D.; Hebert, M. Meta-learning to detect rare objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9925–9934.
18. 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; pp. 8420–8429.
19. Li, X.; Deng, J.; Fang, Y. Few-Shot Object Detection on Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**. [CrossRef]
20. 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, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
21. Wang, X.; Huang, T.; Gonzalez, J.; Darrell, T.; Yu, F. Frustratingly Simple Few-Shot Object Detection. In Proceedings of the International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 9919–9928.
22. 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; pp. 456–472.
23. 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; pp. 5197–5206.
24. Li, Y.; Shao, Z.; Huang, X.; Cai, B.; Peng, S. Meta-FSEO: A Meta-Learning Fast Adaptation with Self-Supervised Embedding Optimization for Few-Shot Remote Sensing Scene Classification. *Remote Sens.* **2021**, *13*, 2776. [CrossRef]
25. Zeng, Q.; Geng, J.; Huang, K.; Jiang, W.; Guo, J. Prototype Calibration with Feature Generation for Few-Shot Remote Sensing Image Scene Classification. *Remote Sens.* **2021**, *13*, 2728. [CrossRef]
26. Li, H.; Cui, Z.; Zhu, Z.; Chen, L.; Zhu, J.; Huang, H.; Tao, C. RS-MetaNet: Deep Metametric Learning for Few-Shot Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6983–6994. [CrossRef]
27. Zhang, P.; Bai, Y.; Wang, D.; Bai, B.; Li, Y. Few-shot classification of aerial scene images via meta-learning. *Remote Sens.* **2021**, *13*, 108. [CrossRef]
28. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
29. 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]
30. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; Lin, D. Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 821–830.

33. Niemeyer, J.; Rottensteiner, F.; Soergel, U. Contextual classification of lidar data and building object detection in urban areas. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 152–165. [CrossRef]
34. 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.