remote sensing

MDPI

Article

Multi-Stage Feature Enhancement Pyramid Network for Detecting Objects in Optical Remote Sensing Images

Kaihua Zhang [1] and Haikuo Shen [2,*]

[1] School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Ministry of Education, Beijing 100044, China; 17116375@bjtu.edu.cn
[2] Key Laboratory of Vehicle Advanced Manufacturing, Measuring and Control Technology, School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Ministry of Education, Beijing 100044, China
* Correspondence: shenhk@bjtu.edu.cn; Tel.: +86-1774-356-8910

Abstract: The intelligent detection of objects in remote sensing images has gradually become a research hotspot for experts from various countries, among which optical remote sensing images are considered to be the most important because of the rich feature information, such as the shape, texture and color, that they contain. Optical remote sensing image target detection is an important method for accomplishing tasks, such as land use, urban planning, traffic guidance, military monitoring and maritime rescue. In this paper, a multi stages feature pyramid network, namely the Multi-stage Feature Enhancement Pyramid Network (Multi-stage FEPN), is proposed, which can effectively solve the problems of blurring of small-scale targets and large scale variations of targets detected in optical remote sensing images. The Content-Aware Feature Up-Sampling (CAFUS) and Feature Enhancement Module (FEM) used in the network can perfectly solve the problem of fusion of adjacent-stages feature maps. Compared with several representative frameworks, the Multi-stage FEPN performs better in a range of common detection metrics, such as model accuracy and detection accuracy. The mAP reaches 0.9124, and the top-1 detection accuracy reaches 0.921 on NWPU VHR-10. The results demonstrate that Multi-stage FEPN provides a new solution for the intelligent detection of targets in optical remote sensing images.

Keywords: Multi-stage Feature Enhancement Pyramid Network; Content-Aware Feature Up-Sampling; feature enhancement module; optical remote sensing images; object detection

Citation: Zhang, K.; Shen, H. Multi-Stage Feature Enhancement Pyramid Network for Detecting Objects in Optical Remote Sensing Images. *Remote Sens.* **2022**, *14*, 579. https://doi.org/10.3390/rs14030579

Academic Editor: Chein-I Chang

Received: 27 December 2021
Accepted: 22 January 2022
Published: 26 January 2022

1. Introduction

In recent years, with the development of earth observation technology, object detection of remote sensing images has gradually become a research hotspot. Remote sensing images can be divided into aerial images and satellite images, and the acquisition of images is usually done by shooting, scanning or microwave radar. Optical remote sensing images are remote sensing images captured by cameras in the visible wavelength range (0.38–0.76 microns), which are extremely rich in shape, texture and color information and constitute the most prevalent types of remote sensing images.

Optical remote sensing image target detection uses specially designed algorithms to find and mark targets of interest (hills, lakes, grounds, buildings, vehicles, aircraft, ships etc.) in images and is an important tool for accomplishing tasks, such as land use, urban planning, traffic diversion, military monitoring and maritime rescue, and is of paramount importance in the field of remote sensing image processing [1–3].

Early optical remote sensing image target detection algorithms used the manual design of features. Despite the widely varying designs of algorithms, at a macro level the ideas are broadly the same: First, to determine the candidate region; Secondly, to detect the features that are designed manually according to the characteristics of the target; Thirdly, a classifier is normally used to classify the category of the target to be detected [4–9]. Stankov

improved the hit-or-miss transform (HMT) and proposed Percentage Occupancy HMT (POHMT) for detecting building locations and invoking vegetation masks to eliminate irrelevant factors [4].

Leninisha et al. proposed a geometric active deformation model based on width and color for extracting a road network from remotely sensed images with minimum human interference. However, the model is inadequate for the detection of complex urban roads [5]. Focusing on the shadows cast by buildings, Ok proposed a shadow post-processing method that uses a probabilistic landscape approach to model the directional spatial relationship between buildings and their shadows, which, in turn, automatically detects building targets in ultra-high resolution multispectral images [6].

Du et al. effectively used a robust anomaly degree measure to improve separability between anomalous pixels and other background pixels. They first distinguished the target from the background using popular features and then used metric learning methods to obtain a robust anomaly degree measure [7]. However, these algorithms have many drawbacks.

For instance, determining candidate regions generally requires setting sliding windows on the image, and most of the generated candidate regions are not the final desired regions, resulting in a large number of redundant calculations and high time complexity. Manually designed features are mainly based on the visual information of the target, and the features are easy to understand but poorly expressed, while the robustness and adaptability of the features are low, making them difficult to adopt in various scenarios.

In addition, unlike common optical images taken on the ground, optical remote sensing images are generally taken from high altitudes, and the quality of the images is easily affected by the environment, climate and light. At the same time, due to the long shooting distance, the scale of the target in the optical remote sensing image varies greatly, and the contour and texture information are not as good as in common optical images, as shown in Figure 1. With the continuous innovation of remote sensing technology, remote sensing images have a higher resolution and larger scale, and the information contained in the images is richer. In this case, earlier detection algorithms will be more complicated in design, and the detection effect cannot meet the actual demand.

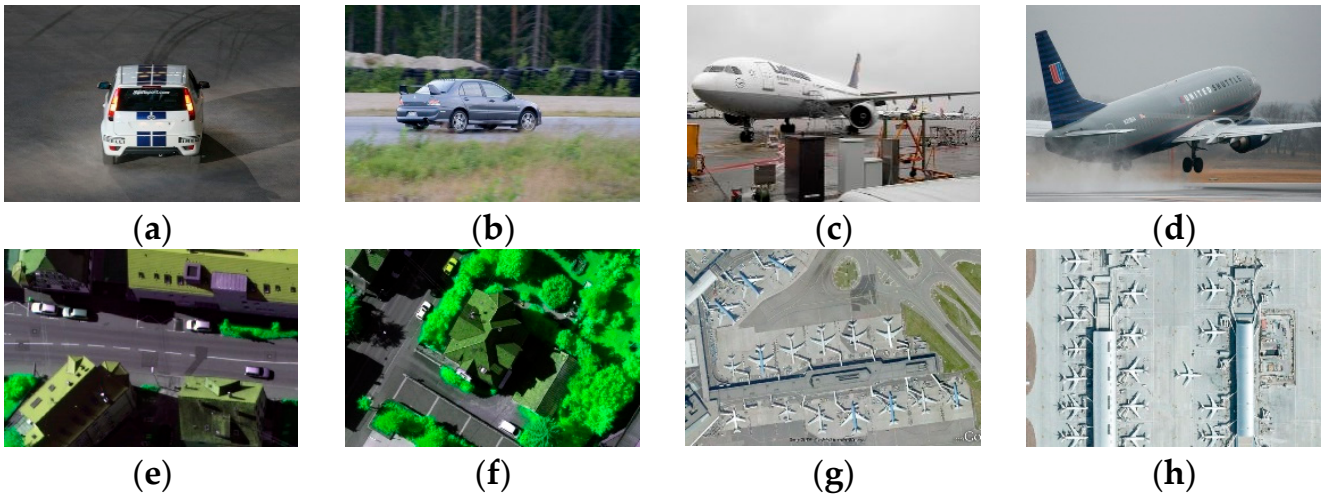


Figure 1. Visual differences between common optical images and optical remote sensing images. The first row is natural images, where the targets in (**a**,**b**) are vehicles and the targets in (**c**,**d**) are aircrafts. The second row is optical remote sensing images, where the targets in (**e**,**f**) are vehicles and the targets in (**g**,**h**) are aircrafts.

In 2006, Hinton proposed a solution to the problem of gradient disappearance during the training of deep neural networks, and since then, deep neural networks have gradually gained recognition. Deep neural networks can extract deeper abstract features, that are not

intuitively understood but have more powerful semantic representations and discriminative properties. Once proposed, deep neural networks dramatically improved the detection accuracy of multi-category targets in images and have received wide attention.

After more than a decade of development, deep neural networks have gradually matured, and many high-level network design solutions have emerged, becoming the mainstream algorithm for solving object detection problems [10–20]. Among these methods, Faster-RCNN [10] provides a new idea to accomplish the task of multi-category target detection for images on an efficient and high accuracy basis. Like other deep neural networks, Faster-RCNN uses convolutional and pooling layers for the down-sampling and computation of feature maps, but instead of Selective Search (SS), Faster-RCNN introduces the 'Anchor' approach for generating proposals and proposes a novel Region Proposal Network (RPN).

The RPN reduces the generation speed of proposal to 10 ms, and thus can focus the training time on feature extraction and classification, which reduces the training cost and improves the detection accuracy. However, the Faster-RCNN contains two networks, making it difficult to meet the requirement of real-time detection. YOLO [11] was proposed to solve the problem of real-time detection with the ability to divide the extracted high-stage feature map into S × S cells directly, and perform bounding box regression for each cell. YOLO is more efficient than Faster-RCNN, which contains only one network.

SSD [12] combines the "Anchor" of Faster-RCNN and the "Cell" of YOLO, using convolutional layers to extract multi-stage features, generating proposals of different sizes and proportions on each convolutional layer for training using the Prior Box algorithm and finally eliminating the redundant boxes using the Non-Maximum Suppression (NMS) algorithm. Compared with Faster-RCNN and YOLO, SSD edges ahead in terms of detection accuracy and detection efficiency.

In contrast to the manual feature design approach, deep neural networks directly integrate identifying candidate regions, learning effective features and feature classifiers to achieve end-to-end detection. Deep learning learns features with stronger semantic characterization ability, driven by large amounts of image data, with greatly improved performance, while avoiding the redundant computation of a large number of windows during the forward propagation of the neural network and improving the detection speed.

The above methods have achieved remarkable results in the task of detecting multi-category targets in images, but they are oriented to common optical images, and the effect of detection in optical remote sensing images is not ideal. As can be seen from Figure 1, optical remote sensing images do not contain as rich of feature information as common optical images, while the scale variation range is large, and there are too many small scale targets. Therefore, how to design special algorithms adapted to the target detection in optical remote sensing images is an urgent problem for experts in this field.

After several years of unremitting efforts, experts around the world have designed a large number of improved algorithms to effectively improve the performance of object detection in the field of optical remote sensing images [21–32]. Cheng et al. added a rotation-invariant layer to the convolutional neural network and proposed a novel rotation-invariant convolutional neural network (RICNN) model. The model achieves rotation invariance by optimizing a new objective function for training, by imposing regularization constraints, and by specifying that the feature representations of the training samples before and after rotation must be mapped to each other [21].

Han et al. found that the proposal generation network and feature classification network of the Faster-RCNN are two separate parts, which are not efficient in training and detection. They compared the acquisition method and annotation method of optical remote sensing images and common optical images and proposed that the annotation of optical remote sensing images is costly. Therefore, to address these problems, Han improved the Faster-RCNN and proposed an efficient and robust integrated geospatial target detection framework named R-P-Faster-RCNN.

R-P-Faster-RCNN shares features in the proposal generation phase and target detection phase and achieves the integration of both, thus improving the network training and detection efficiency. The model uses common optical images for pre-training and optical remote sensing images for fine-tuning during training, thus, solving the problem of expensive optical remote sensing image annotation [22]. Ren et al. argued that the building blocks of standard convolutional neural networks have a fixed geometric structure and are, therefore, limited in geometric transformations.

To eliminate this effect, Ren integrated a deformable convolutional module in the Faster-RCNN. This module is capable of unsupervised learning of the augmented spatial sampling locations in the module. In addition, they generated a single high-level feature map with fine resolution on which predictions can be made using top-down and skip connections. Ren et al. named this Def. Faster-RCNN, and the network shows more significant results on the SORSI and HRRS datasets [23].

Xu et al. proposed a deformable region-based fully convolutional network (Def. R-FCN) to remove the obvious limitation of convolutional neural networks for modeling geometric changes of remote-sensing targets. For training, Xu et al. first pre-trained using natural images and subsequently fine-tuned using ultra-high resolution remotely sensed images. To compensate for the increased number of lines, such as false region suggestions, an aspect-ratio-constrained non maximum suppression (arcNMS) was designed [24].

Li et al. proposed a rotation-insensitive and context-augmented object detection network (RICADet) to solve the problem of rotation change sensitivity and blurred appearance in remote sensing images. The network contains an improved region suggestion network and a local context feature fusion network to solve the above two types of problems, respectively, and a comprehensive evaluation on a publicly available ten object detection datasets demonstrates the effectiveness of the network [25].

Guo et al. proposed a multi-scale convolutional neural network (Multi-Scale CNN) to accomplish geospatial object detection in high-resolution satellite images. The network consists of a multi-scale object suggestion network and a multi-scale target detection network, where high-quality proposals are proposed by the multi-scale object suggestion network, and the proposals are trained using the multi-scale target detection network to generate a good target detector [26].

The above networks are indeed effective for optical remote sensing images; however, there is still much room for improvement in detection accuracy. In particular, they are not ideal for the detection of blurred small-scale targets in optical remote sensing images. In order to further improve the detection accuracy of optical remote sensing image targets, this paper proposes a Multi-stage Feature Enhancement Pyramid Network (Multi-stage FEPN).

This network can generate multi-stage feature maps and effectively fuse adjacent high-stage feature maps with low-stage feature maps to enrich the feature information contained in the feature maps. At the same time, the Multi-stage FEPN introduces a feature enhancement module to highlight useful features and improve the target classification accuracy and localization precision.

We detail the structure and the highlighted design parts of Multi-stage FEPN as well as the datasets and evaluation metrics chosen for this paper in Section 2. In Section 3, we use a unified dataset to compare the above introduced Faster-RCNN [10], YOLO [11], SSD [12], RICNN [21], R-P-Faster-RCNN [22], Def. Faster-RCNN [23], Def. R-FCN [24], RICADet [25], Multi-Scale CNN [26] and the Multi-stage FEPN proposed in this paper to demonstrate that Multi-stage FEPN works better than other networks in optical remote sensing image target detection tasks.

2. Materials and Methods

2.1. Network Architecture

We propose a novel deep convolutional neural network with reference to the design idea of FPN [33]. In the feature map generation stage, as in the FPN, we use a combination

of bottom-to-top branch, up-to-down branch and lateral connection branch to complete the process. The bottom-to-top branch uses the ResNet-101 network to extract multi-stage feature maps group {C1, C2, C3, C4, C5}.

In the up-to-down branch, we first max-pool C5 to obtain P6, then use the Content-Aware Feature Up-Sampling (CAFUS) algorithm to up-sample P6 to the scale of C5, use convolution on C5 to change the number of channels, and finally fuse P6 and C5 to obtain P5. Similarly, we obtain P4, P3 and P2 and combine them with P5 to obtain the feature pyramid {P2, P3, P4, P5}. In order to further enhance the representation ability of the feature maps, we introduce a Feature Enhancement Module (FEM) to modify the features and finally obtain more optimized feature maps group {F2, F3, F4, F5}.

After obtaining the feature map sets, we generate proposals using RPN and fix the proposals into unique scales using RoI Align. We use two fully connected layer branches to compute the category scores and location regression parameters of the modified proposals to finally complete the classification and localization of the targets. To reduce the complexity of training, an adaptive mapping of the generated proposals onto feature maps is performed to calculate the training losses and adjust the network parameters according to the scales of the labeled boxes. The structure of the network proposed in this paper is shown in Figure 2.

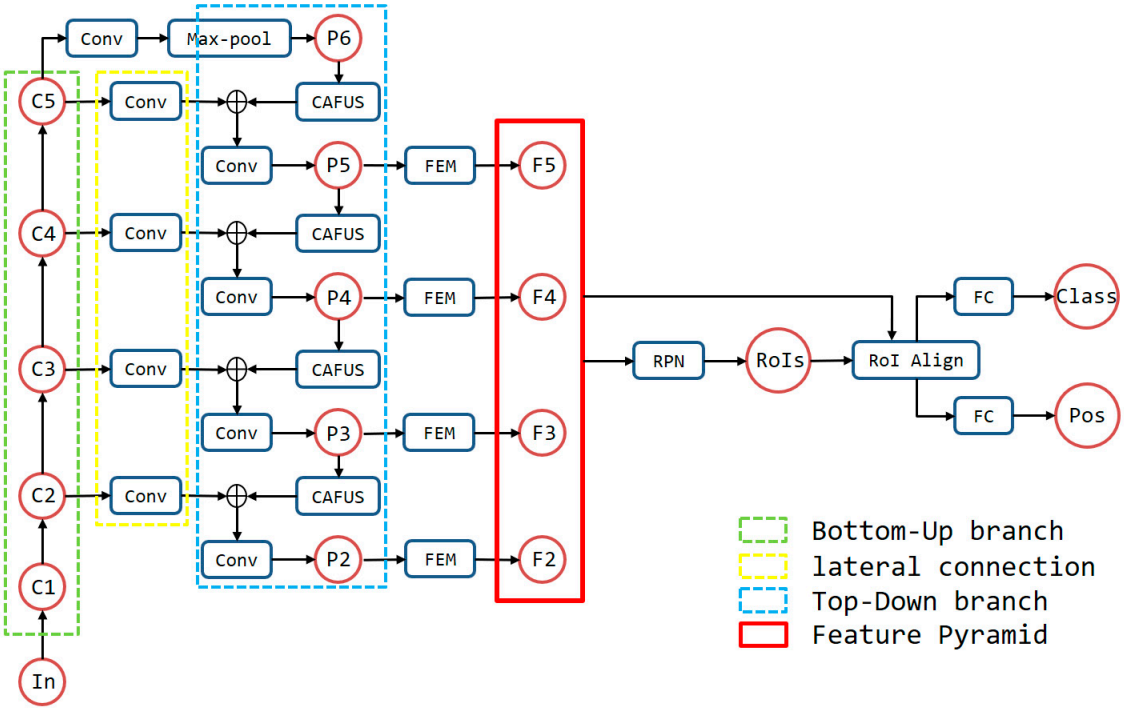


Figure 2. The structure of the Multi-stage FEPN. The green dashed box represents the bottom-to-top branch, the blue dashed box represents the up-to-down branch, and the yellow dashed box represents lateral connection branch. Finally, the feature maps in red box form the feature maps group {F2, F3, F4, F5}.

2.2. Content-Aware Feature Up-Sampling

In the process of feature pyramid generation, the fusion of feature maps of adjacent stages is a problem that needs to be focused on. The feature maps at different stages have different scales, and each feature map expresses different feature information. Therefore, it is extremely important to design an effective feature map up-sampling and fusion algorithm. Traditional feature pyramid networks mainly use interpolation [34] for up-sampling higher-stage feature maps, including nearest neighbor interpolation, bilinear interpolation and bicubic interpolation.

Interpolation is a purely mathematical algorithm that only calculates the new pixels after up-sampling based on the pixel positions and does not fully utilize the semantic

information of the image. At the same time, the perceptual field of the interpolation method is usually small and the computed image can be distorted, while increasing the perceptual field can lead to a significant increase in computational cost. Therefore, the use of interpolation to upsample higher-stage feature maps can only solve the problem of different scales, and the noise introduced in the process will weaken the representation ability of the feature maps to a certain extent.

In addition to the interpolation method, current image up-sampling algorithms include deconvolution [35] and dynamic filters [36]. The deconvolution method can improve the above problems to some extent by learning the kernel parameters through convolutional networks without considering the pixel positions. However, this method does not consider local semantics and uses the same convolutional kernel for each local region, which still cannot effectively restore local feature information, and the computational effort grows exponentially when the convolutional kernel design is too large. The dynamic filtering method designs a convolutional kernel for each position of the image, which is conceivably too large a number of parameters for practical applications.

Content-Aware ReAssembly of Features (CARAFE) [37] is a learnable image up-sampling algorithm based on the input content, which divides the feature map up-sampling into two parts, i.e., up-sampling kernel prediction and feature reassembly. Experiments show that CARAFE has a large perceptual field, and the model is light enough to retain the feature information of the input better. CARAFE will also enhance the feature semantics to an extent. In this paper, we design a new image up-sampling algorithm, named Content-Aware Feature Up-Sampling, as shown in Figure 3.

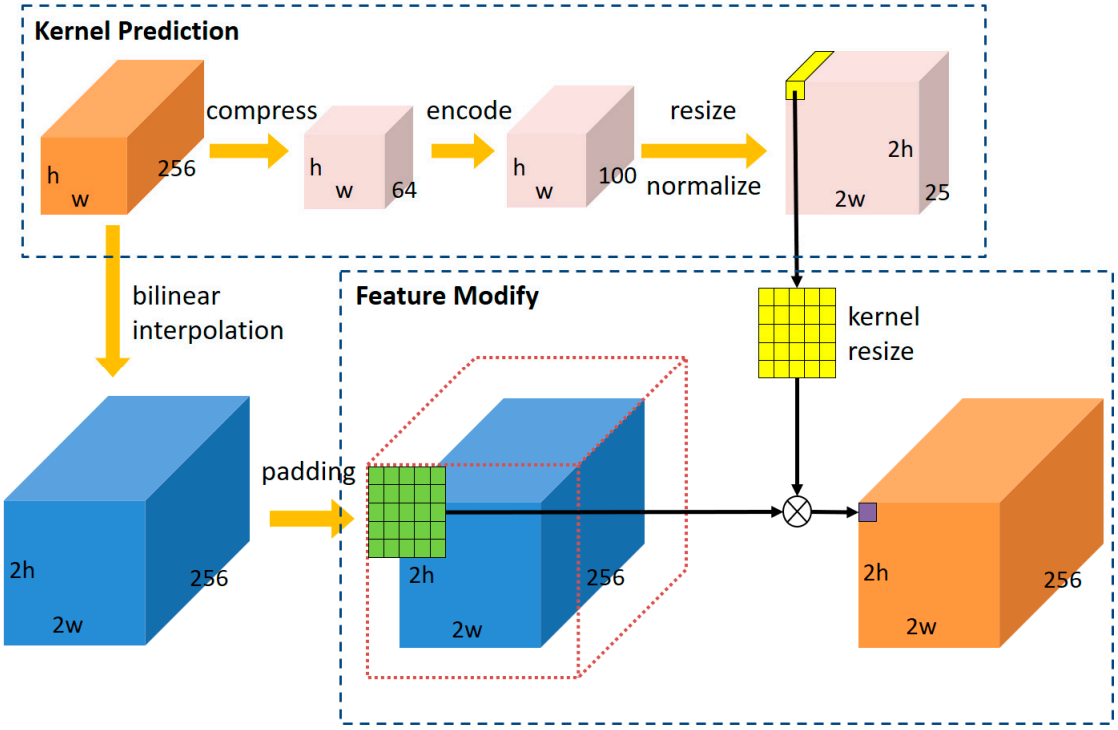


Figure 3. The structure of CAFUS. We chose a prediction kernel size of 5 × 5 and an up-sampling ratio of 2; therefore, the number of channels of the encoded feature map is 2 × 2 × 5 × 5 = 100. As for the choice of the number of compressed channels, we found that increasing the number of compressed channels does not have a significant improvement on the algorithm effect.

Unlike CARAFE, the feature modification kernel of CAFUS is predicted from the input image but does not act directly on the input. Instead, it is applied to the input image after interpolation. CAFUS can make up for the shortcomings of the interpolation method by introducing learnable parameters to fine-tune the interpolated image and maximize the

retention of the semantic information of the higher-stage feature map. At the same time, CAFUS has a simple structure and does not introduce overly many learning parameters, thus, ensuring the training efficiency.

2.3. Feature Enhance Module

Although CAFUS is able to preserve the detail information inside the image during the up-sampling of higher-stage feature maps and enhance the representation ability of the feature maps, the fusion of adjacent-stages feature maps only uses the convolutional "summing" method to fuse multi-channels information. In order to solve this problem, this paper proposes a feature weighting algorithm to further enhance the feature information, i.e., the Feature Enhance Module. The algorithm principle of the FEM is shown in Figure 4. The FEM assigns weights to the feature maps from two directions, thus, enhancing the useful features.

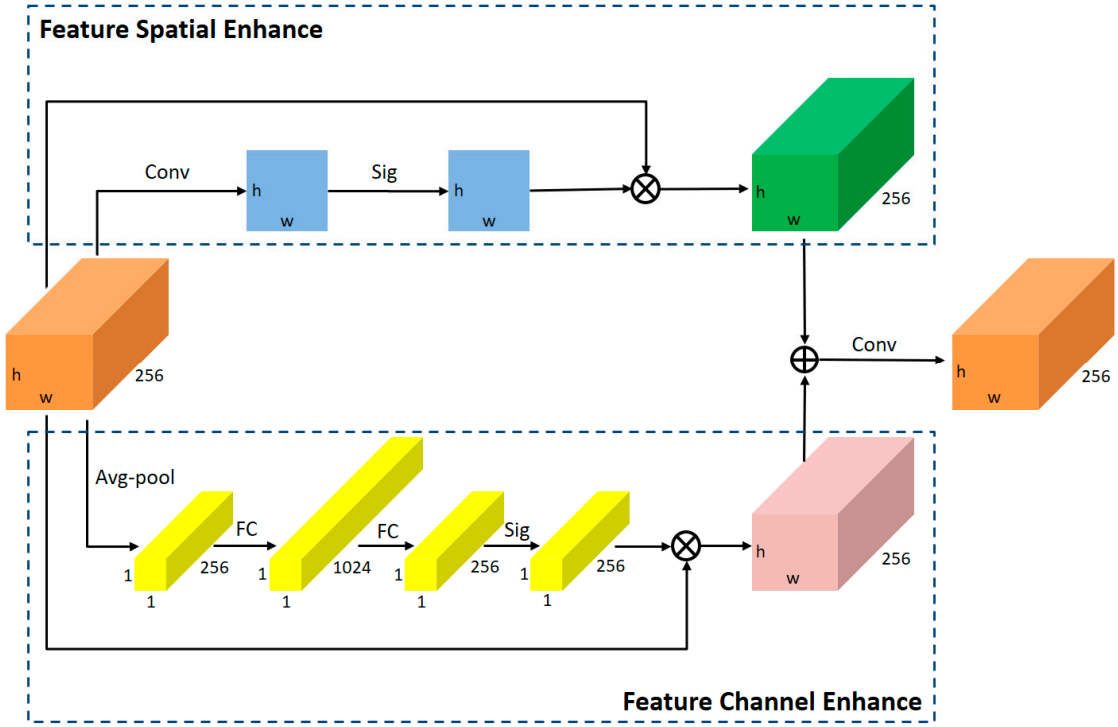


Figure 4. Details of the Feature Enhance Module. Feature spatial enhance (FSE) fuses multi-channel information using a 1×1 convolution kernel and normalizes it using a sigmoid function to obtain two-dimensional spatial weights. Feature channel enhance (FCE) obtains a 1×256 vector using global average pooling. Then, the vector is calculated by two fully connected layers and normalized using the sigmoid function to obtain the one-dimensional channel weights. Enhanced features from the two approaches are superimposed together, followed by further fusion of the local feature information using 3×3 convolution.

In the spatial direction, the fused feature map contains both target features and background information. The feature map is to be modified along the two-dimensional spatial direction to enhance the target features. The specific weight parameters are calculated as shown in Equations (1) and (2). In Equation (2), $\circledast$ represents the multiplication of elements in the corresponding position.

$$\hat{Z}_{(x,y)} = \sigma\left(\sum_{i=1}^{C} P^{i}_{(x,y)} \cdot kernel^{1\times1}_{(x,y)}\right), x \in [1,h],\ y \in [1,w], \tag{1}$$

$$P^{c}_{FSE} = P^{c} \circledast \hat{Z},\ \ c \in [1,C], \tag{2}$$

In the channel direction, as each level of the feature map group contains 256 channels, not every channel of the two-dimensional feature map can contain the target feature information well. Therefore, feature map channel weights are introduced to assign a coefficient to each channel so that the feature maps containing the target features in the feature map sets can be used to greater effect. The feature map channel weight parameters are calculated as shown in Equations (3)–(5). In Equation (4), δ represents a sigmoid function, and σ represents a ReLU activation function.

$$Z^c = \frac{1}{h \times w} \sum_{x=1}^{h} \sum_{y=1}^{w} P_{(x,y)}, \quad c \in [1, C], \tag{3}$$

$$\hat{Z} = \sigma(\delta(W_2 \times \delta(W_1 \times Z))), \tag{4}$$

$$P_{FCE}^c = \hat{Z}^c \times P^c, \quad c \in [1, C], \tag{5}$$

Finally, the features modified in both directions are fused together, as shown in Equation (6), and the fused feature map will contain stronger feature information.

$$P_{FEM} = Conv_{3\times3}(P_{FSE} + P_{FCE}), \tag{6}$$

2.4. Loss Function

The training of deep convolutional neural networks is a process of iterative optimization of network model parameters. Typically, the network first computes the category score loss as well as the position loss for each proposed box in the forward direction, and updates the network model parameters in the reverse direction based on the obtained losses, i.e., back-propagation of losses. Therefore, in the training process of deep neural networks, using a reasonable loss function can achieve better convergence of the network parameters and better detection accuracy of the network. In this paper, different loss functions for the proposal generation part and the feature classification parts are applied to calculate the loss values of each part, as shown in Equations (7)–(11).

For the category loss, the cross-entropy function is used to calculate the loss. Cross-entropy is simple in design and can automatically adjust the learning speed of the network parameters according to the loss size when back-propagating to avoid overfitting, which can effectively calculate the loss for multi-category classification tasks. For location loss, we use the smooth-L1 function for calculation. Smooth-L1 sets a constant 1 as the back-propagation gradient for points with large loss values, while, for points with small loss values, the backpropagation gradient decreases as the loss value decreases, and thus smooth-L1 is insensitive to outliers.

$$L_{RPN\_cls} = \frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} cross\_entro(p_i, p_i^*), \tag{7}$$

$$L_{RPN\_bbox} = \frac{\lambda}{N_{boxes}} \sum_{i=1}^{N_{boxes}} p_i^* \times smooth\_L1(t_i, t_i^*), \tag{8}$$

where N_{batch} represents the number of images in one mini-batch at training, N_{boxes} represents the number of anchors generated in each image, and λ represents the balance factor. p_i is the category score vector of the proposal box, p_i^* represents the label of the proposal box, t_i represents the position parameter of the proposal box, and t_i^* refers to the position parameter of the ground truth box.

$$L_{RCNN\_cls} = cross\_entro(p, u), \tag{9}$$

$$L_{RCNN\_bbox} = smooth\_L1(t^u, v), \tag{10}$$

where p denotes the category score predicted by the network, u represents the score of the ground truth, t^u represents the coordinate of the ground truth, and v represents the coordinate of the predicted box.

$$Loss = L_{RPN\_cls} + L_{RPN\_bbox} + L_{RCNN\_cls} + L_{RCNN\_bbox}, \quad (11)$$

2.5. Dataset Selection and Data Augmentation

As deep neural networks become the dominant algorithm for remote sensing image processing, the demand for remote sensing images is increasing, and therefore many teams have open-sourced remote sensing image datasets for use by other scholars. Among them, some typical remote sensing image datasets include DOTA (a large-scale Dataset for Object deTection in Aerial images) [38], UCAS-AOD (UCAS-High Resolution Aerial Object Detection Dataset) [39], NWPU VHR-10 (NWPU very-high resolution optical remote sensing images with 10 categories) [1,21,40], RSOD-Dataset (Remote sensing object detection) [41,42], the INRIA Aerial Image Labeling Dataset [43] and the TGRS-HRRSD-Dataset [44].

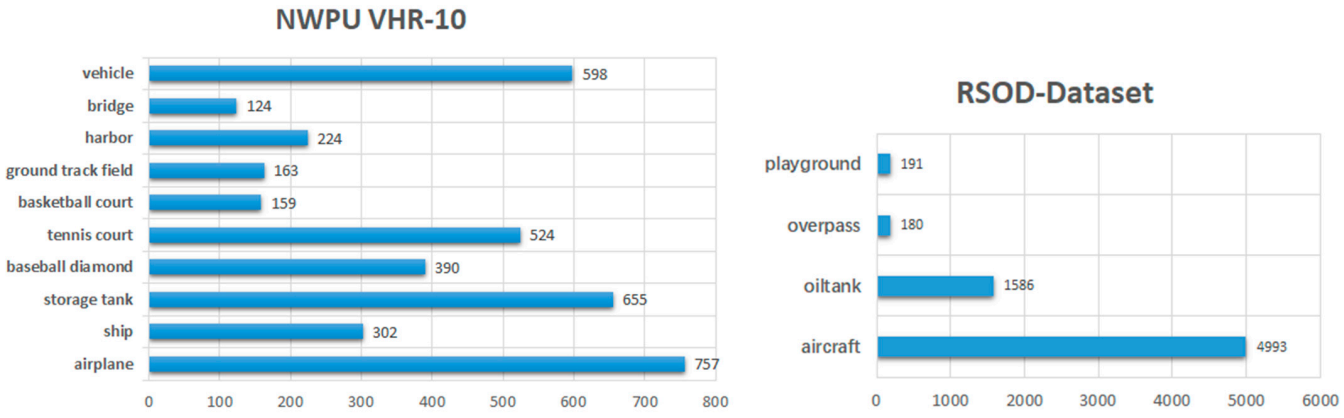
DOTA was co-produced by State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing and School of Electronic Information and Communications, HUST and consists of 2806 remotely sensed images with targets containing a total of 188,282 instances in 15 categories. UCAS-AOD was compiled and labeled by the Pattern Recognition and Intelligent Systems Development Laboratory of the University of Chinese Academy of Sciences. It consists of 910 remote sensing images with only two categories of targets, aircraft and vehicles, with negative background samples totaling 8066 instances.

NWPU VHR-10 contains 800 high-resolution satellite images cropped from the Google Earth and Vaihingen datasets and then manually annotated by NWPU experts. Among them, 650 images carry labeled targets, with a total of 10 categories and 3896 instances. RSOD-Dataset was labeled by a team from Wuhan University, and the dataset contains 976 images divided into four categories with a total of 6950 instances (4993 instances belong to the category of "aircraft", accounting for 71.8%).

The INRIA Aerial Image Labeling Dataset is a dataset for urban building detection, collected and labeled by the French National Institute for Information and Automation. It has only two categories, building and not building, and is semantically segmented at the pixel level. The TGRS-HRRSD-Dataset was collected and labeled by the Xi'an Institute of Optical Precision Machinery, Chinese Academy of Sciences. It contains 13 categories of objects, more than 21,000 images and more than 40,000 instances.

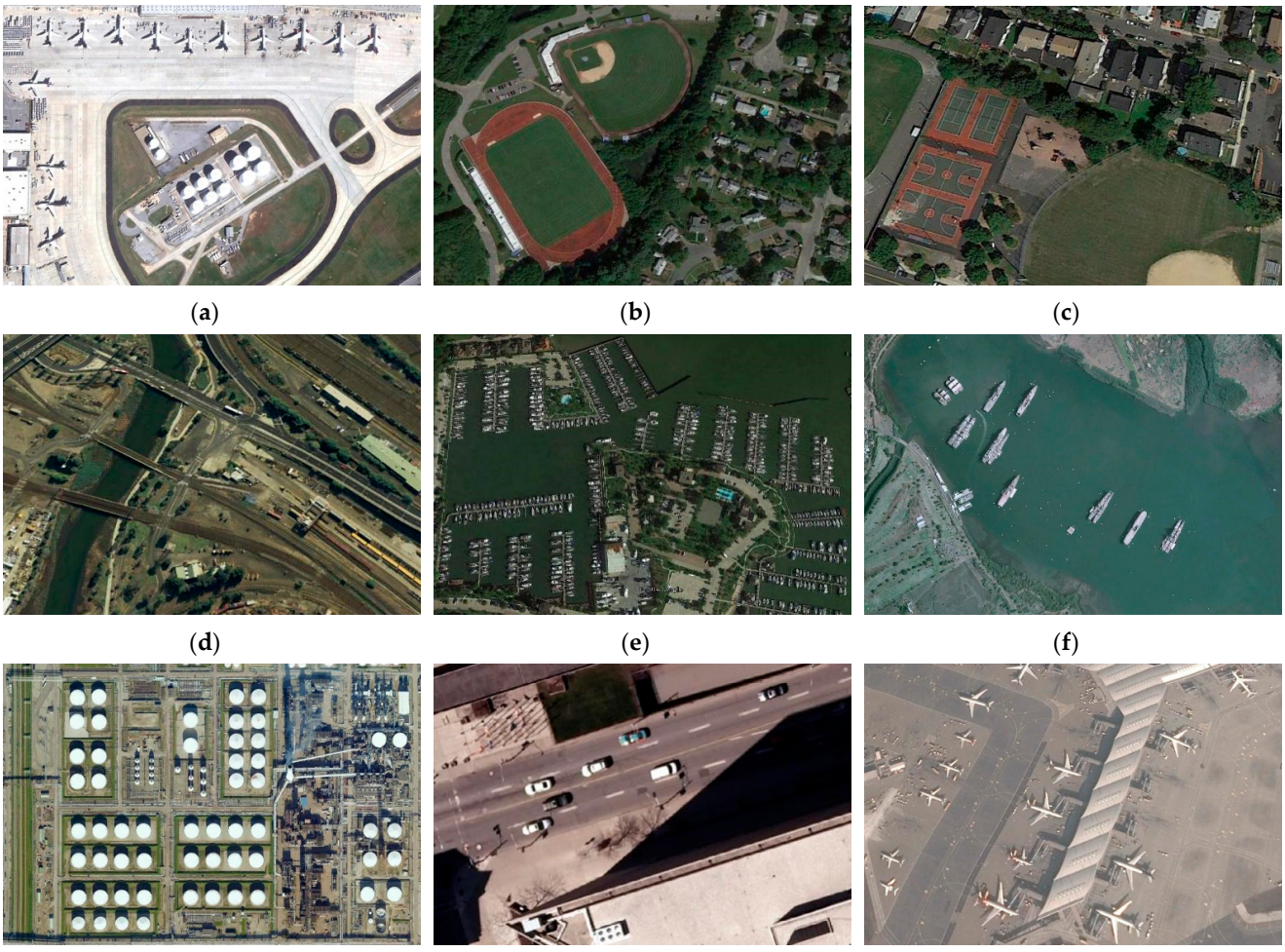
Considering the training cost and data diversity, we decided to select NWPU VHR-10 and RSOD-Dataset as the datasets used for the experiments in this paper. The number of instances of each category in the two datasets is shown in Figure 5, and examples of two datasets corresponding to each category are shown in Figure 6. Compared with RSOD-Dataset, NWPU VHR-10 contains more categories and a relatively even number of instances; therefore, we used NWPU VHR-10 as the main dataset for the comparison of experimental results.

Among the 650 images in the original dataset, we randomly selected 50 images as test images and used the remaining 600 images as training dataset images. In addition, we performed data augmentation on the training set images, including horizontal flip, vertical flip, diagonal flip, random luminance and random contrast, to expand the training set to 3600 images. As for the RSOD-Dataset, 900 images were used as training dataset images, and 36 images were used as test images.



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

Figure 5. The categories and the number of instances of each category in two datasets. (**a**) NWPU VHR-10; (**b**) RSOD-Dataset. Compared with other datasets, NWPU VHR-10 contains more typical and comprehensive object category. Among them, the number of instances with larger scale is smaller compared to those with smaller scale, but the larger the scale of the instances, the more feature information they contain, and the easier they are to train.



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

(**d**) (**e**) (**f**)

(**g**) (**h**) (**i**)

Figure 6. *Cont.*

(**j**) (**k**) (**l**)

Figure 6. Examples corresponding to each category of two datasets. (**a**–**h**) NWPU VHR-10 dataset: (**a**) airplane, (**b**) baseball diamond and ground track field, (**c**) basketball court and tennis court, (**d**) bridge, (**e**) harbor, (**f**) ship, (**g**) storage tank and (**h**) vehicle. (**i**–**l**) RSOD-Dataset: (**i**) aircraft, (**j**) oil tank, (**k**) overpass and (**l**) playground.

2.6. Evaluation Metrics

In image target detection tasks, the most intuitive metric to evaluate the effectiveness of a detection network is the detection accuracy, i.e., whether all targets are detected correctly. The accuracy of an image target detection network can be represented by top-1 and top-5, where top-1 represents the probability that the highest confidence category of each target prediction is correct and top-5 denotes the probability that the top five confidence categories of each target prediction contain the correct category. Early deep neural network target detection frameworks used top-1 detection accuracy and top-5 detection accuracy for combined evaluation due to low detection accuracy, and, with the gradual optimization of the network, currently only top-1 is used to represent the accuracy of the detection network.

However, the samples we tested are of finite size, and there are limitations as the accuracy is obtained by detecting a small sample of test images. Therefore, we need to evaluate a detection network model comprehensively from different perspectives. The detection results of an image target detection network can be classified into four cases, namely TP (True Positive), FP (False Positive), TN (True Negative) and FN (False Negative).

TP indicates when positive samples are predicted to be positive, FP shows when positive samples are predicted to be negative, TN represents when negative samples are predicted to be negative and FN indicates when negative samples are predicted to be positive. When a target detection network model completes the detection task, the number of samples corresponding to these four cases can be counted, and the Precision and Recall of the model can be calculated based on the counted data as shown in Equations (12) and (13).

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{12}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{13}$$

Precision indicates the accuracy of the model's prediction results, with higher values indicating better results. Recall indicates the proportion of correct predictions to the labeled targets, with higher values indicating that the model is more capable of recalling positive samples during training. Ideally, we would like the model to have higher Precision and Recall; however, in reality, the relationship is usually negative. Using Recall as the horizontal coordinate and Precision as the vertical coordinate, the Precision–Recall curve (P-R curve) is obtained by picking different thresholds to count the values of Recall and Precision and plotting the curves.

Following the principle that the larger the two metrics are, the better, we can believe that the larger the area surrounded by the curve and the coordinate axis, the higher the accuracy of the model, and this area is the Average Precision (AP) corresponding to each category. In the multi-category target detection task, we calculate the average of the APs

corresponding to all categories to obtain Mean Average Precision (mAP), and the larger this value is, the higher the combined accuracy of the model can be.

3. Results

To verify that the Multi-stage FEPN proposed in this paper can detect targets in optical remote sensing images more effectively, we selected the commonly used frameworks for common optical image target detection tasks, such as Faster-RCNN, YOLO and SSD, and for optical remote sensing image target detection tasks, such as RICNN, R-P-Faster-RCNN, Def. Faster-RCNN, Def. R-FCN, RICADet and Multi-Scale CNN, for experiments. In terms of hardware, we used a GTX 1050Ti graphics card for the graphical computations. The network parameters and data set assignments used during the experiments are shown in Table 1.

Table 1. Network training parameters and dataset details for Multi-stage FEPN.

	NWPU VHR-10	RSOD-Dataset
Steps of training	60,000	60,000
Initial learning rate	0.01	0.01
Data augmentation	Yes	No
Pre-training	Yes	Yes
Backbone	ResNet-101	ResNet-101
Images of training set	2304	576
Images of validation set	576	144
Images of testing set	720	180

3.1. Evaluation of Proposed Multi-Stage FEPN with NWPU VHR-10 Dataset

During the training process, the loss values generated by Multi-stage FEPN were recorded every 10 iterations during the training process, and the loss curves were plotted as shown in Figure 7. To more clearly represent the direction of change of the loss during the training process, we fitted the data using the least squares method, which is represented by the red curve in Figure 7. From the loss curve, it can be seen that the loss decreases gradually during the training process of the network, with the loss decreasing faster at the beginning of the training and then slower, which is in line with our expected results. Finally, the loss of the network is essentially stable around 0.1, indicating that the training process of Multi-stage FEPN is stable.

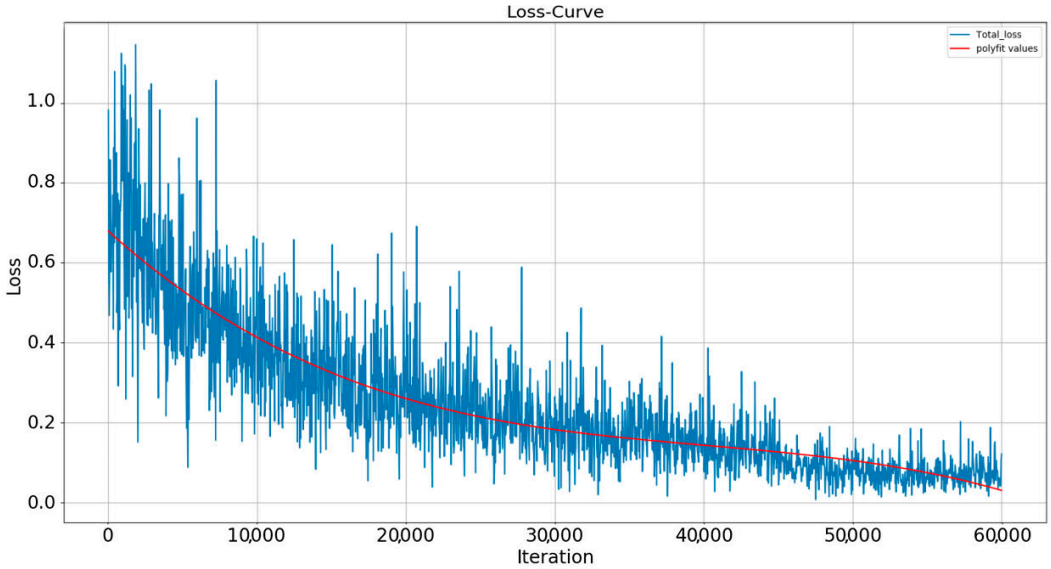


Figure 7. The loss curve and fitted curve of Multi-stage FEPN.

Figure 8 shows the recall rate for each category of targets in the NWPU VHR-10 dataset. The results show that Multi-stage FEPN has a high recall rate for most targets, but the recall rates for storage tanks and vehicles are relatively low, only 0.8364 and 0.8433.

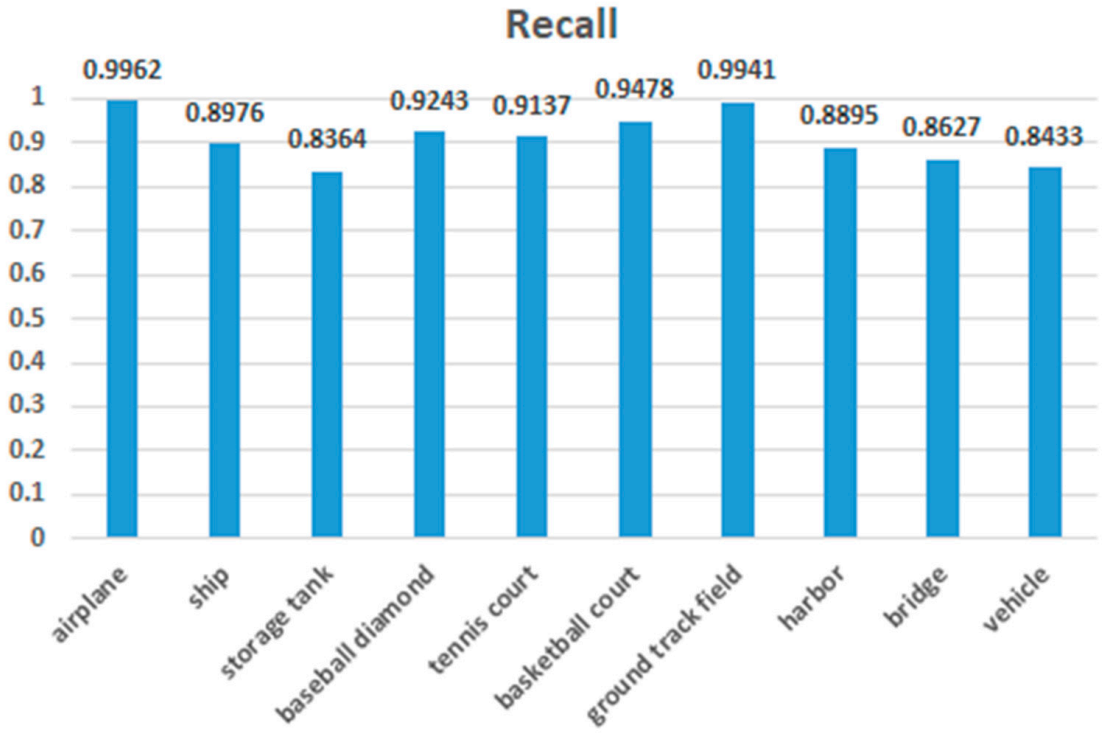


Figure 8. The recall rate for each category of targets of Multi-stage FEPN.

To validate the effect of Multi-stage FEPN, ablation experiment were conducted. We calculated the APs and mAP of FPN, FPN with CAFUS, FPN with FEM and Multi-stage FEPN on the NWPU VHR-10 dataset and plotted the P-R curves, respectively, as shown in Figure 9 and Table 2. By comparison, we can see that CAFUS and FEM are indeed effective for feature fusion and feature enhancement, with the AP for each category of targets improved, and the mAP is also improved by 0.0812 from 0.8312 to 0.9124, indicating a clear advantage.

Table 2. The APs for each category of different methods and mAP in the ablation experiment.

	FPN	FPN with CAFUS	FPN with FEM	Multi-Stage FEPN
Airplane	0.9088	0.9091	0.9091	**0.9995**
Ship	0.8967	0.9001	0.9029	**0.9041**
Storage tank	0.7814	0.7839	0.7647	**0.8024**
Baseball diamond	0.9033	0.9049	0.8888	**0.9131**
Tennis court	0.7598	0.8922	0.8849	**0.8926**
Basketball court	0.8762	0.9022	0.8123	**0.9524**
Ground track field	0.9091	0.9975	0.9984	**0.9992**
Harbor	0.8914	0.8972	0.8961	**0.9079**
Bridge	0.7436	0.8782	0.7991	**0.8853**
Vehicle	0.6416	0.7040	0.7833	**0.8671**
Mean AP	0.8312	0.8769	0.8640	**0.9124**

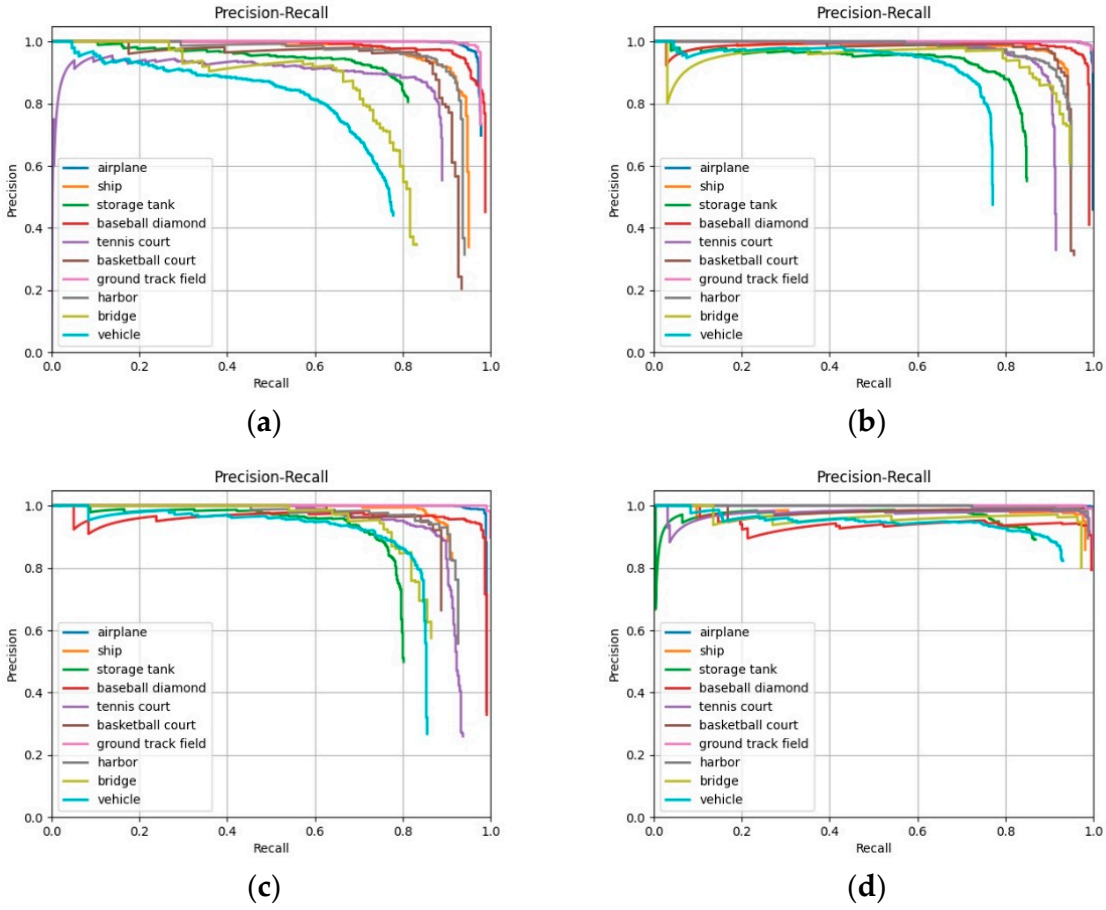


Figure 9. The P-R curves of different methods in the ablation experiment. (**a**) FPN; (**b**) FPN with CAFUS; (**c**) FPN with FEM; (**d**) Multi-stage FEPN.

We plotted the P-R curves for each target detection framework in the comparison experiment separately, as shown in Figure 10. For comparison, we calculated the APs and the mAP of each framework in Table 3, the top-1 detection accuracy for each category of targets and the average top-1 detection accuracy of Multi-stage FEPN in Table 4.

Table 3. The APs for each category and mean AP of different frameworks.

	Faster-RCNN	YOLO	SSD	RICNN	R-P-Faster-RCNN	Def. R-FCN	Def. Faster-RCNN	RICADet	Multi-Scale CNN	Multi-Stage FEPN
Airplane	0.9053	0.9026	0.9078	0.9086	0.9087	0.9082	0.9077	0.9089	0.9086	**0.9995**
Ship	0.8001	0.7491	0.8205	0.8188	0.8372	0.8677	0.8972	0.9011	0.8960	**0.9041**
Storage tank	0.5934	0.4962	0.4981	0.5735	0.5747	0.7896	0.8024	0.7976	0.8041	**0.8078**
Baseball diamond	0.9065	0.9054	0.9087	0.9089	0.9083	0.8927	0.9002	0.8673	0.8819	**0.9131**
Tennis court	0.7694	0.7936	0.8138	0.8091	0.8126	0.8452	0.8672	0.8921	0.8889	**0.8926**
Basketball court	0.8999	0.8991	0.9091	0.9091	0.9064	0.5917	0.9057	0.9056	0.8992	**0.9524**
Ground track field	0.3594	0.4225	0.5379	0.6309	0.6428	0.9091	0.9091	0.9091	0.9091	**0.9992**
Harbor	0.8654	0.7950	0.8691	0.8043	0.9043	0.8858	0.8972	0.9057	**0.9942**	0.9079
Bridge	0.5647	0.6292	0.6497	0.8038	0.7751	0.7246	0.7996	0.8138	0.8102	**0.8853**
Vehicle	0.4215	0.3501	0.5025	0.6513	0.6722	0.7233	0.7073	0.7841	0.7864	**0.8671**
Mean AP	0.7086	0.6943	0.7417	0.7818	0.7942	0.8438	0.8599	0.8685	0.8779	**0.9124**

Table 4. The top-1 detection accuracy for each category and average top-1 detection accuracy of Multi-stage FEPN.

	Airplane	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Field	Harbor	Bridge	Vehicle	Average Top-1
Multi-stage FEPN	0.929	0.904	0.853	0.913	0.922	0.943	0.994	0.924	0.911	0.915	0.921

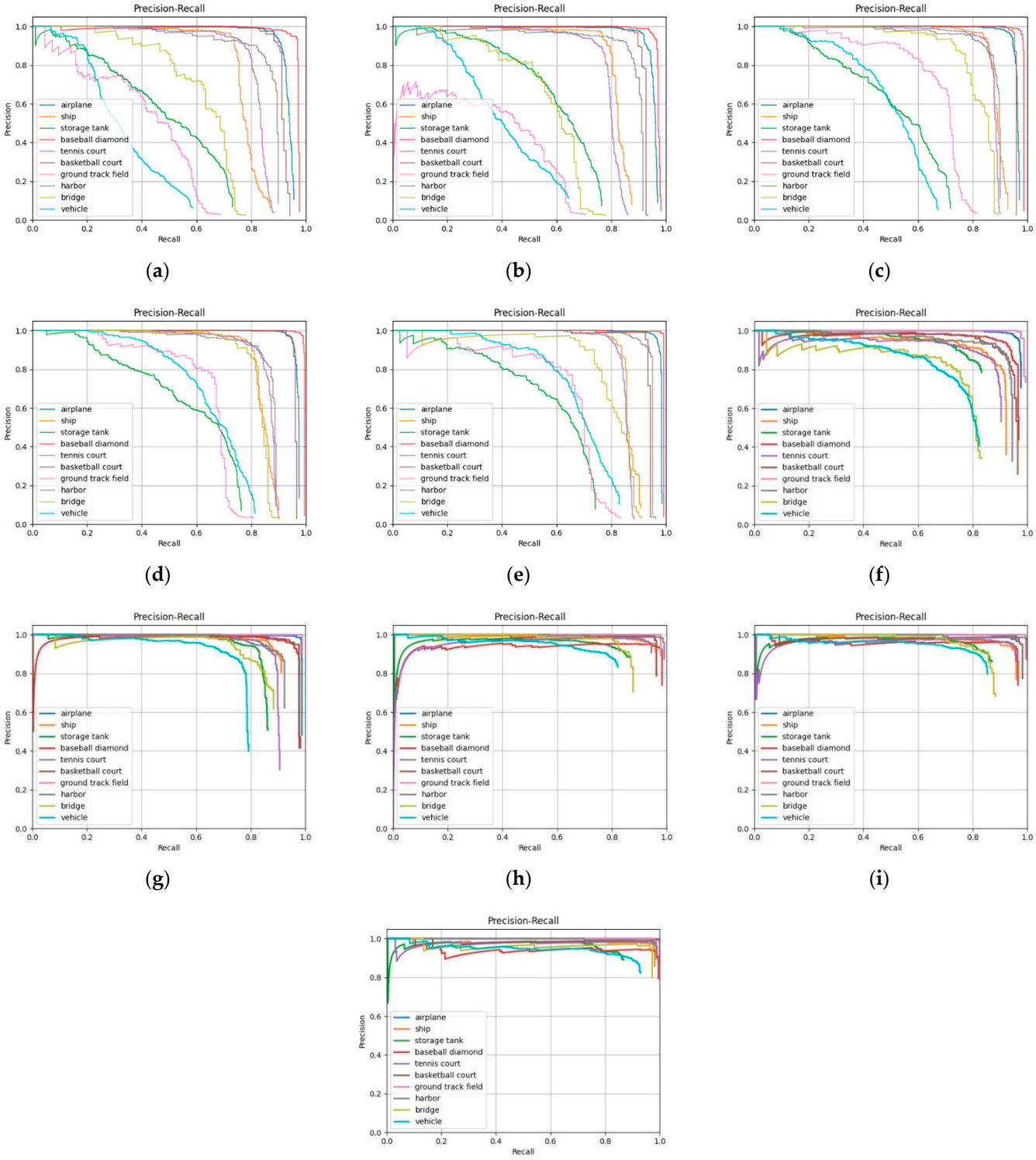


Figure 10. The P-R curves of different frameworks with the NWPU VHR-10 dataset. (**a**) Faster-RCNN; (**b**) YOLO; (**c**) SSD; (**d**) RICNN; (**e**) R-P-Faster-RCNN; (**f**) Def. R-FCN; (**g**) Def. Faster-RCNN; (**h**) RICADet; (**i**) Multi-Scale CNN; (**j**) Multi-stage FEPN.

The experimental results show that the target detection framework that is applicable to common optical images cannot be well adapted to the optical remote sensing image target detection task and is not ideal for the detection of blurred small targets and targets with large scale variation range. Multi-stage FEPN outperformed other optical remote

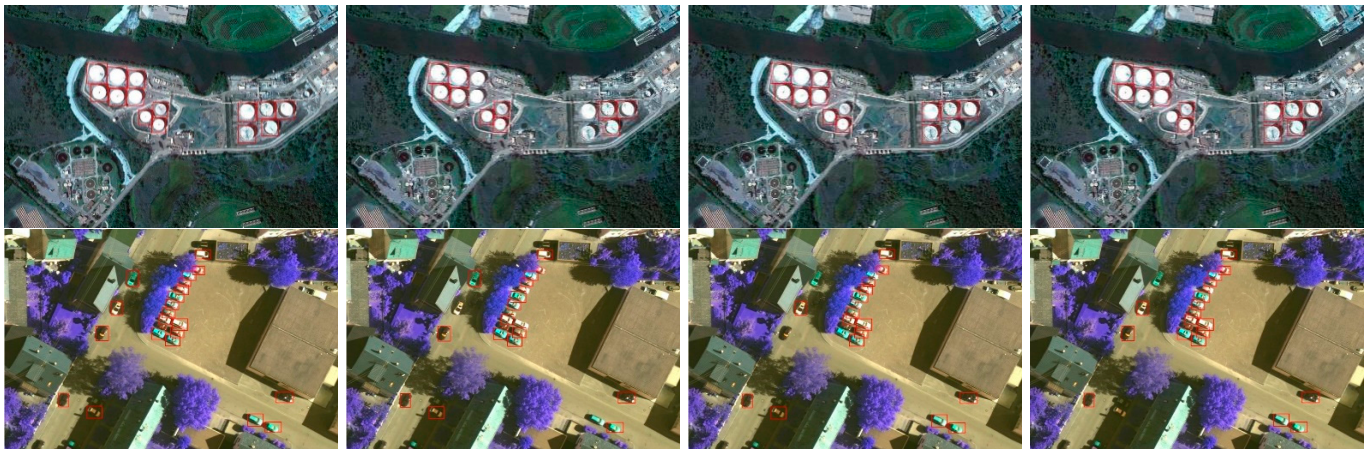
sensing image target detection frameworks in terms of the AP and overall mAP for most categories of targets, especially for blurred small targets, such as aircraft and vehicles.

We recorded the detection times of different frames for the test images, as shown in Table 5. Since Multi-stage FEPN retains multi-stage feature maps and uses algorithms for feature map fusion and feature enhancement in the framework, the testing time was slightly increased compared to some other frameworks.

Table 5. The average running time of different frameworks.

	Faster-RCNN	YOLO	SSD	RICNN	R-P-Faster-RCNN	Def. R-FCN	Def. Faster-RCNN	RICADet	Multi-Scale CNN	Multi-Stage FEPN
Average running time per image (second)	0.28	0.19	0.26	8.77	0.643	1.27	0.726	2.89	0.67	1.16

To more visually demonstrate the detection capability of Multi-stage FEPN for fuzzy small targets, a number of the network's detection results for vehicles and storage tanks are shown in Figure 11. For convenience, we mark only the location of the target in Figure 10 and remove the text labels. The comparison shows that Multi-stage FEPN outperformed other better performing frameworks in detecting small fuzzy targets.



(**a**) (**b**) (**c**) (**d**)

Figure 11. The detection results of different networks for vehicles and storage tanks. (**a**) Multi-stage FEPN; (**b**) Def. Faster-RCNN; (**c**) RICADet; (**d**) Multi-Scale CNN.

Figure 12 shows some detection results of Multi-stage FEPN on the NWPU VHR-10 dataset. In general, the detection accuracy of Multi-stage FEPN is relatively high, and it can perform the task of optical remote sensing image target detection very well.

3.2. Evaluation of Proposed Multi-Stage FEPN with RSOD-Dataset

In Section 2, the RSOD-dataset is briefly described, which contains only four categories of targets, with "aircraft" accounting for 71.4% of the total instances; thus, the dataset does not have a balanced sample. Figure 13 shows the P-R curves of Def. R-FCN, RICADET, Multi-Scale CNN and the proposed Multi-stage FEPN on RSOD-Dataset, and Table 6 shows the APs of different categories and mAP of each framework. The comparison shows that Multi-stage FEPN is better than the other networks in all metrics. Although the unbalanced samples have some influence on Multi-stage FEPN, the detection effect of Multi-stage FEPN is still better when compared with the other frameworks.

Figure 12. Some detection results of the Multi-stage FEPN with NWPU VHR-10.

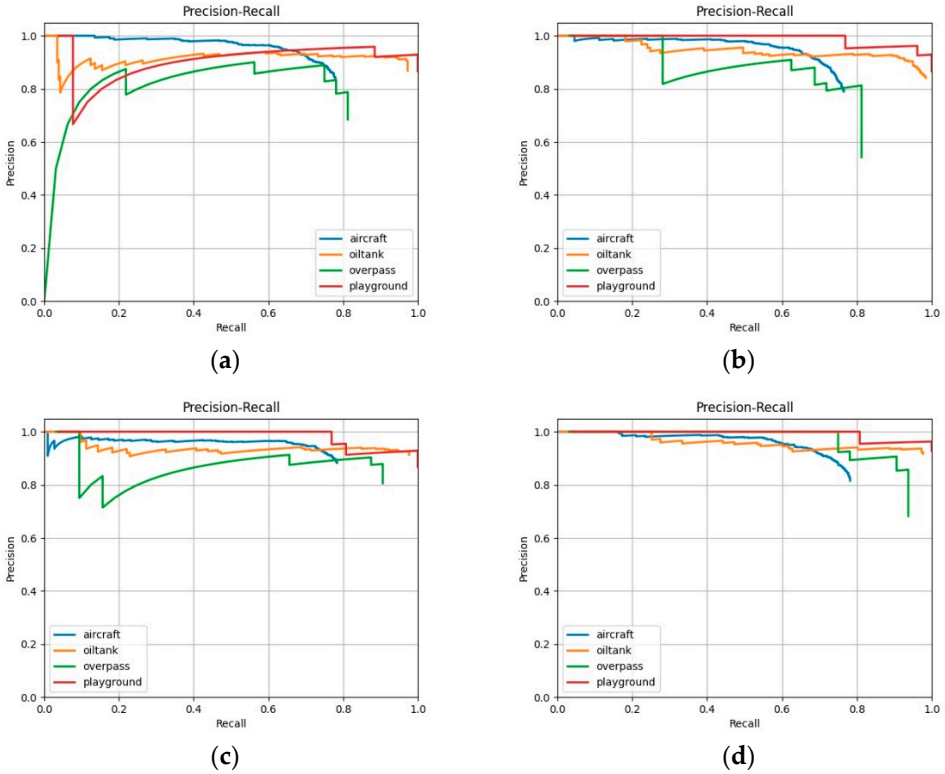


Figure 13. The P-R curves of different frameworks on RSOD-Dataset. (a) Def. R-FCN; (b) RICADet; (c) Multi-Scale CNN; (d) Multi-stage FEPN.

Table 6. The APs for each category and the mAP of different frameworks on RSOD-Dataset.

	Def. R-FCN	RICADet	Multi-Scale CNN	Multi-Stage FEPN
Aircraft	0.7116	0.7089	0.7063	**0.7106**
Oil tank	0.8551	0.8682	0.8629	**0.8782**
Overpass	0.7242	0.7518	0.8330	**0.8920**
Playground	0.9567	0.9865	0.9829	**0.9933**
Mean AP	0.8119	0.8289	0.8463	**0.8685**

4. Discussion

Optical remote sensing images are susceptible to the effects of environment, climate and illumination on the quality of imaging compared to common optical images. Due to the longer distance taken, certain targets are presented in optical remote sensing images at a smaller scale, i.e., small targets. Such small targets consist of fewer pixels and feature information, such as contours and textures, which is one of the pressing issues in the field of target detection in optical remote sensing images. At the same time, there are many kinds of targets to be detected in optical remote sensing images, and there are large scale variations among different kinds of targets, and thus the designed framework needs to consider multi-scale targets and small targets comprehensively to improve the overall detection accuracy.

In recent years, many studies have been proposed to improve the accuracy of small and multi-scale targets in optical remote sensing images [45–49]. Collectively, a common approach to solve the small target detection problem is to use multi-stage feature maps for recognition, which is because multi-stage feature maps can enrich the features of small-scale targets, and models can rely on more feature difference information for target classification.

The methods to solve the multi-scale target detection problem usually introduce multi-scale information in the network, including multi-stage classifiers, multi-stage features quadratic fusion and loss function weighting. Although these methods can deal with the specificities of targets in optical remote sensing images, they can reduce the detection efficiency or the detection accuracy will not be sufficiently high because of insufficient feature usage.

The Multi-stage FEPN proposed in this paper adopts the idea of multi-stage feature maps fusion so that, for small targets, the feature information of low-level feature maps can be used to achieve classification, while for multiple targets with large scale variations, the appropriate stage of feature maps can be reasonably selected to generate corresponding proposals and perform classification. This method performs better in the optical remote sensing image target detection task.

In the experiments of this paper, we demonstrated that the feature map up-sampling algorithm CAFUS used in Multi-stage FEPN can improve the fusion effect of adjacent-stage feature maps, while the feature map enhancement algorithm FEM can also highlight the features effectively through the ablation experimental results in Figure 9 and Table 2. Therefore, the accuracy is high in detecting blurred small-scale targets, as shown in Figure 11. Then, three target detection frameworks (Faster-RCNN, YOLO and SSD) applicable to common optical images and six target detection frameworks (RICNN, R-P-Faster-RCNN, Def. Faster-RCNN, Def. R-FCN, RICADet and Multi-Scale CNN) applicable to remote sensing images were compared in parallel, and the experimental results are recorded in Figure 10 and Tables 3 and 4.

The experimental results show that Multi-stage FEPN was better than the other frameworks in most categories of target detection accuracy, especially for small-scale targets, such as airplanes, ships, storage tanks and vehicles, which have higher detection accuracy improvement. Finally, similar experiments were conducted on RSOD-Dataset. The experimental results show that the detection ability of Multi-stage FEPN decreases on datasets with extreme sample imbalance, which becomes one of the key elements of our future research, i.e., addressing the impact of the sample imbalance problem on deep convolutional neural networks.

In this paper, we only address a small part of the remote sensing image target detection field, i.e., solving the detection problem of blurred small-scale targets and multi-scale targets in optical remote sensing images. The task of remote sensing image target detection has many essential problems that need to be focused on and broken through. For example, optical remote sensing images are sensitive to weather; therefore, it is obvious that how to attenuate or eliminate the influence of weather on the image quality when the weather is cloudy or foggy is a key factor to improve the detection accuracy.

When there are clouds obscuring the targets or the data is satellite remote sensing images, considering the variability of cloud shape and the extremely small scale of the target in the satellite remote sensing image [50–52], the instance segmentation algorithm should be introduced. A large number of studies have confirmed that instance segmentation algorithms has better detection effects compared with target classification for target scale diversity and very small scale targets [53–56].

In summary, in our future work, we will focus on the followings: First, we will introduce the concept of deep separable convolution and the training mechanism of federated learning to reduce the number of parameters of the network, and achieve the lightweight of the model while ensuring the detection accuracy. Secondly, we will introduce the image defogging algorithm to enhance the quality of optical remote sensing images and improve the feature representation of the targets. Third, we will introduce the concept of instance segmentation and design corresponding algorithms to further improve the target detection accuracy of remote sensing images.

5. Conclusions

The research in this paper aimed to accomplish the task of multi-object detection of optical remote sensing images. A Multi-stage Feature Enhance Pyramid Network was proposed, and the detection capability of the framework was experimentally verified. The main contributions of the results in this paper are as follows. First, we made targeted improvements to address the shortcomings of the interpolation method and designed a learnable up-sampling algorithm, named Content-Aware Feature Up-Sampling, which corrects the higher-stage feature map after interpolation and improves the feature representation capability of the feature map after up-sampling.

Secondly, this paper analyzed the problem that the simple fusion process of adjacent-stages feature maps tends to introduce noise and weaken feature information, and we proposed a Feature Enhancement Module to further augment the obtained fused feature maps. The module learns the weights from the spatial and channel directions, which effectively suppresses the influence of noise and thus enriches the useful features of the feature maps. Finally, the framework proposed in this paper is compared with other commonly used frameworks for optical remote sensing images, and through the analysis of various model evaluation metrics, we demonstrated that the Multi-stage Feature Enhance Pyramid Network is more effective in the optical remote sensing image multi-object detection task.

Author Contributions: Conceptualization, H.S.; methodology, K.Z.; software, K.Z.; validation, K.Z. and H.S.; formal analysis, K.Z.; investigation, K.Z.; data curation, H.S.; writing—original draft preparation, K.Z.; writing—review and editing, K.Z.; project administration, H.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

References

1. Cheng, G.; Han, J. A Survey on Object Detection in Optical Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
2. 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]
3. Nie, G.T.; Huang, H. A Survey of Object Detection in Optical Remote Sensing Images. *Acta Autom. Sin.* **2021**, *47*, 1749−1768.
4. Stankov, K. Detection of Buildings in Multispectral Very High Spatial Resolution Images Using the Percentage Occupancy Hit-or-Miss Transform. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *7*, 4069–4080. [CrossRef]
5. Leninisha, S.; Vani, K. Water flow based geometric active deformable model for road network. *ISPRS J. Photogramm. Remote Sens.* **2015**, *102*, 140–147. [CrossRef]
6. Ok, A.O. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS J. Photogramm. Remote Sens.* **2013**, *86*, 21–40. [CrossRef]
7. Bo, D.; Zhang, L. A Discriminative Metric Learning Based Anomaly Detection Method. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 6844–6857.
8. Zhang, L.B.; Zhang, Y.Y. Airport Detection and Aircraft Recognition Based on Two-Layer Saliency Model in High Spatial Resolution Remote-Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 1511–1524. [CrossRef]
9. Gu, W.; Lv, Z.H.; Hao, M. Change detection method for remote sensing images based on an improved Markov random field. *Multimed. Tools Appl.* **2017**, *76*, 17719−17734. [CrossRef]
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
11. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Cham, Switzerland, 2016.
13. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [CrossRef]
14. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 Octorber–2 November 2020; IEEE: Manhattan, NY, USA, 2020.
15. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 Octorber–2 November 2020; IEEE: Manhattan, NY, USA, 2020.
16. He, Y.; Xu, S.; Gao, L.; Zhang, B. Ship Detection Without Sea-Land Segmentation for Large-Scale High-Resolution Optical Satellite Images. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; IEEE: Manhattan, NY, USA, 2018.
17. Shen, Y.; Ji, R.; Wang, Y.; Chen, Z.; Zheng, F.; Huang, F.; Wu, Y. Enabling Deep Residual Networks for Weakly Supervised Object Detection. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020.
18. Qiu, H.; Ma, Y.; Li, Z.; Liu, S.; Sun, J. BorderDet: Border Feature for Dense Object Detection. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020.
19. Chen, Z.M.; Jin, X.; Zhao, B.; Wei, X.S.; Guo, Y. Hierarchical Context Embedding for Region-based Object Detection. In *Computer Vision—ECCV 2020, Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020.
20. Zhang, K.H.; Shen, H.K. Solder Joint Defect Detection in the Connectors Using Improved Faster-RCNN Algorithm. *Appl. Sci.* **2021**, *11*, 576. [CrossRef]
21. Gong, C.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415.
22. Han, X.B.; Zhong, Y.F.; Zhang, L.P. An Efficient and Robust Integrated Geospatial Object Detection Framework for High Spatial Resolution Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]
23. Yun, R.; Zhu, C.; Xiao, S. Deformable Faster R-CNN with Aggregating Multi-Layer Features for Partially Occluded Object Detection in Optical Remote Sensing Images. *Remote Sens.* **2018**, *10*, 1470.
24. Xu, Z.; Xu, X.; Wang, L.; Yang, R.; Pu, F. Deformable ConvNet with Aspect Ratio Constrained NMS for Object Detection in Remote Sensing Imagery. *Remote Sens.* **2017**, *9*, 1312. [CrossRef]
25. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-Insensitive and Context-Augmented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2337–2348. [CrossRef]
26. 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]
27. Chen, S.Q.; Zhan, R.H.; Zhang, J. Geospatial object detection in remote sensing imagery based on multiscale single-shot detector with activated semantics. *Remote Sens.* **2018**, *10*, 820. [CrossRef]

28. Wang, P.J.; Sun, X.; Diao, W.H.; Fu, K. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 3377−3390. [CrossRef]
29. Fu, Y.; Wu, F.; Zhao, J. Context-Aware and Depthwise-based Detection on Orbit for Remote Sensing Image. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018.
30. Wang, C.; Bai, X.; Wang, S.; Zhou, J.; Ren, P. Multiscale Visual Attention Networks for Object Detection in VHR Remote Sensing Images. *IEEE Geoence Remote Sens. Lett.* **2019**, *16*, 310–314. [CrossRef]
31. Pang, J.; Li, C.; Shi, J.; Xiao, Z.; Yu, W.; Havyarimana, V.; Jiao, L. R2-CNN: Fast Tiny Object Detection in Large-Scale Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens. A Publ. IEEE Geosci. Remote Sens. Soc.* **2019**, *57*, 5512–5524. [CrossRef]
32. 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. A Publ. IEEE Geosci. Remote Sens. Soc.* **2019**, *57*, 5146–5158. [CrossRef]
33. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21 July–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2017.
34. Zhong, B.; Lu, Z.; Ji, J. Review on Image Interpolation Techniques. *J. Data Acquis. Process.* **2016**, *31*, 1083–1096.
35. Dumoulin, V.; Visin, F. A guide to convolution arithmetic for deep learning. *arXiv* **2016**, arXiv:1603.07285.
36. Jo, Y.; Oh, S.W.; Kang, J.; Kim, S.J. Deep Video Super-Resolution Network Using Dynamic Upsampling Filters Without Explicit Motion Compensation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18 June–23 June 2018; IEEE: Manhattan, NY, USA, 2018.
37. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware ReAssembly of Features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; IEEE: Manhattan, NY, USA, 2020.
38. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18 June–23 June 2018; IEEE: Manhattan, NY, USA, 2018.
39. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation Robust Object Detection in Aerial Images Using Deep Convolutional Neural Network. In Proceedings of the IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015; IEEE: Manhattan, NY, USA, 2015.
40. 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]
41. Xiao, Z.; Liu, Q.; Tang, G.; Zhai, X. Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *Int. J. Remote Sens.* **2015**, *36*, 618–644. [CrossRef]
42. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
43. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. In Proceedings of the Igarss IEEE International Geoscience & Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017; IEEE: Manhattan, NY, USA, 2017.
44. Zhang, Y.; Yuan, Y.; Feng, Y.; Lu, X. Hierarchical and Robust Convolutional Neural Network for Very High-Resolution Remote Sensing Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5535–5548. [CrossRef]
45. Li, Y.; Huang, Q.; Pei, X.; Chen, Y.; Jiao, L.; Shang, R. Cross-layer Attention Network for Small Object Detection in Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 2048–2146. [CrossRef]
46. Li, C.; Xu, C.; Cui, Z.; Wang, D.; Zhang, T.; Yang, J. Feature-Attentioned Object Detection in Remote Sensing Imagery. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: Manhattan, NY, USA, 2019; pp. 3886–3890.
47. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
48. Zhang, S.; He, G.; Chen, H.B.; Jing, N.; Wang, Q. Scale adaptive proposal network for object detection in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 864–868. [CrossRef]
49. Pham, M.T.; Courtrai, L.; Friguet, C.; Lefèvre, S.; Baussard, A. YOLO-Fine: One-stage detector of small objects under various backgrounds in remote sensing images. *Remote Sens.* **2020**, *12*, 2501. [CrossRef]
50. Chen, Y.; Weng, Q.; Tang, L.; Liu, Q.; Fan, R. An Automatic Cloud Detection Neural Network for High-Resolution Remote Sensing Imagery with Cloud–Snow Coexistence. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 6004205. [CrossRef]
51. Li, Y.; Chen, W.; Zhang, Y.; Tao, C.; Xiao, R.; Tan, Y. Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning. *Remote Sens. Environ.* **2020**, *250*, 112045. [CrossRef]
52. de Albuquerque, A.O.; de Carvalho, O.L.F.; e Silva, C.R.; Luiz, A.S.; Pablo, P.; Gomes, R.A.T.; Guimarães, R.F.; de Carvalho Júnior, O.A. Dealing with Clouds and Seasonal Changes for Center Pivot Irrigation Systems Detection Using Instance Segmentation in Sentinel-2 Time Series. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8447–8457. [CrossRef]
53. Park, H.G.; Yun, J.P.; Kim, M.Y.; Jeong, S.H. Multichannel Object Detection for Detecting Suspected Trees with Pine Wilt Disease Using Multispectral Drone Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8350–8358. [CrossRef]

54. Carvalho, O.L.F.D.; de Carvalho Júnior, O.A.; Albuquerque, A.O.D.; Bem, P.P.D.; Silva, C.R.; Ferreira, P.H.G.; Moura, R.D.S.D.; Gomes, R.A.T.; Guimarães, R.F.; Borges, D.L. Instance segmentation for large, multi-channel remote sensing imagery using Mask-RCNN and a Mosaicking approach. *Remote Sens.* **2021**, *13*, 39. [CrossRef]
55. de Carvalho, O.L.F.; de Moura, R.D.S.; de Albuquerque, A.O.; de Bem, P.P.; Pereira, R.D.C.; Weigang, L.; Borges, D.L.; Guimarães, R.F.; Gomes, R.A.T.; de Carvalho Júnior, O.A. Instance Segmentation for Governmental Inspection of Small Touristic Infrastructure in Beach Zones Using Multispectral High-Resolution WorldView-3 Imagery. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 813. [CrossRef]
56. Dai, Y.; Zhang, J.; He, M.; Porikli, F.; Bowen, L.I.U. Salient object detection from multi-spectral remote sensing images with deep residual network. *J. Geod. Geoinf. Sci.* **2019**, *2*, 101.