Article

Feature Enhancement-Based Ship Target Detection Method in Optical Remote Sensing Images

Liming Zhou ^{1,2} , Yahui Li ^{1,2}, Xiaohan Rao ^{1,2} , Yadi Wang ^{1,2,*}, Xianyu Zuo ^{1,2}, Baojun Qiao ^{1,2} and Yong Yang ³

¹ Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng 475000, China; lmzhou@henu.edu.cn (L.Z.); liyahui@henu.edu.cn (Y.L.); rshenhaibb@henu.edu.cn (X.R.); xianyu_zuo@henu.edu.cn (X.Z.); qbj@henu.edu.cn (B.Q.)

² School of Computer and Information Engineering, Henan University, Kaifeng 475000, China

³ Institute of Plant Stress Biology, State Key Laboratory of Cotton Biology, Department of Biology, Henan University, Kaifeng 475000, China; 104752160097@vip.henu.edu.cn

* Correspondence: yadiwang@henu.edu.cn

Abstract: Ship targets in ORSIs (Optical Remote Sensing Images) have the characteristics of various scales, and most of them are medium and small-scale targets. When the existing target detection algorithms are applied to ship target detection in ORSIs, the detection accuracy is low. There are two main reasons for the above problems, one is the mismatch of the receptive fields, and the other is the lack of feature information. For resolving the problem that multi-scale ship targets are difficult to detect, this paper proposes a ship target detection algorithm based on feature enhancement. Firstly, EIRM (Elastic Inception Residual Module) is proposed for feature enhancement, which can capture feature information of different dimensions and provide receptive fields of different scales for mid- and low-level feature maps. Secondly, the SandGlass-L block is proposed by replacing the ReLu6 activation function of the SandGlass block with the Leaky ReLu activation function. Leaky ReLu solves the problem of 0 output when ReLu6 has negative input, so the SandGlass-L block can retain more feature information. Finally, based on SandGlass-L, SGLPANet (SandGlass-L Path Aggregation Network) is proposed to alleviate the problem of information loss caused by dimension transformation and retain more feature information. The backbone network of the algorithm in this paper is CSPDarknet53, and the SPP module and EIRM act after the backbone network. The neck network is SGLPANet. Experiments on the NWPU VHR-10 dataset show that the algorithm in this paper can well solve the problem of low detection accuracy caused by mismatched receptive fields and missing feature information. It not only improves the accuracy of ship target detection, but also achieves good results when extended to other categories. At the same time, the extended experiments on the LEVIR dataset show that the algorithm also has certain applicability on different datasets.

Keywords: receptive field; feature enhancement; ship target detection; Optical Remote Sensing Images

Citation: Zhou, L.; Li, Y.; Rao, X.; Wang, Y.; Zuo, X.; Qiao, B.; Yang, Y. Feature Enhancement-Based Ship Target Detection Method in Optical Remote Sensing Images. *Electronics* **2022**, *11*, 634. <https://doi.org/10.3390/electronics11040634>

Academic Editors: Miin-shen Yang and Cocianu Catalina-Lucia

Received: 28 January 2022

Accepted: 16 February 2022

Published: 18 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the process of space technology, the amount of ground image data acquired by remote sensing satellites has increased exponentially [1]. The demand for rapid and accurate processing of remote sensing data is becoming increasingly urgent. With the advancement of globalization, ships play an important role in both maritime economy and trade and national defense and security. Therefore, the detection of ship targets in ORSIs is of great significance. Although target detection algorithms have been developing rapidly, target detection in ORSIs still faces many difficulties. ORSIs has the complex background, the target is closely related to the background, and at the same time, it is affected by light and clouds, which makes detection difficult. In particular, small target detection still faces huge challenges. It occupies fewer pixels in ORSIs, and it is easy to lose information during convolution operations or mix with complex backgrounds. Therefore, target detection in remote sensing image still has great room for improvement.

The existing target detection methods mainly include traditional methods and methods based on deep learning. Traditional methods rely on prior knowledge to train by manually designing features [2], then use sliding windows to select candidate frames on the picture, and finally send the candidate frames to the classifier for detection. Deep learning-based methods learn rich feature data for detection through convolutional neural networks [3]. The target characteristics of the traditional method rely on manual design [4], and the design is complex, universality, robustness, and accuracy need to be improved. As deep learning develops, traditional methods are gradually being replaced by deep learning-based methods. There are two types of target detection algorithms based on deep learning: one-stage target detection methods and two-stage target detection methods. The former treats classification and regression as a problem, with low accuracy but fast detection speed. The latter has two steps of region: suggestion generation and region-based suggestion detection [5]. First, a great quantity of candidate regions was generated, and then the candidate regions are secondarily corrected through the backbone network to locate and classify the target object. The two-stage method has slower speed but higher accuracy. One-stage target detection algorithms mainly include YOLO v1 [6], YOLO v2 [7], YOLO v3 [8] and other YOLO series and SSD [9], DSSD [10] and FSSD (Feature Fusion SSD) [11,12], etc. Two-stage target detection algorithms mainly include R-CNN [13], Fast R-CNN [14–16], Faster R-CNN [16–18], Mask R-CNN [19,20], etc.

The detection speed of the YOLO series of algorithms has been widely recognized, but it is not as accurate as the two-stage algorithm. The backbone network of YOLO v1 is based on the GoogleNet image classification model [21], using 1×1 and 3×3 convolutions to replace the inception modules of GoogleNet, and the feature maps extracted from backbone network are detected through two fully connected layers. When predicting, each picture is delimited $s \times s$ grids, and each grid can only predict one target. When multiple targets are located in the same grid, only one of the targets can be detected. Therefore, although YOLO v1 is fast and rarely has false positives, the problem of missed detection is very serious and the types of detection targets are limited. The YOLO v2 backbone network uses Darknet19. The backbone network contains 19 convolutional layers and five max-pooling layers. After convolution, BN (Batch Normalization) is added to improve convergence speed and prevent overfitting. Secondly, the resolution of the input image is improved, and passthrough is introduced to reduce the loss of detailed information. Finally, anchor boxes are used to improve detection accuracy. The largest category of detection has reached 9000. YOLO v3 draws on the ideas of ResNet [22] and proposes Darknet53 as the backbone network, which is composed of 1×1 and 3×3 convolutions and residual connections. The deepening of the network allows it to extract more information and the introduction of residual connections avoids the problem of gradient disappearance in deep convolutional neural networks. The neck network uses FPN (Feature Pyramid Network) [23] to perform feature fusion and multi-scale detection. Assuming that the input image is 416×416 , finally use 13×13 , 26×26 , 52×52 feature maps to detect large, medium and small targets, respectively. YOLO v4 [24] uses CSPDarknet53 as the backbone network. The SPP (Spatial Pyramid Pooling) [25] module acts on the feature map extracted by the backbone network after the fifth downsampling to enlarge the receptive field. The neck network uses PANet [26], which uses upsampling, downsampling and parallel branches to perform feature fusion. The detection head still uses YOLO Head. The balance of accuracy and speed has made YOLO v4 widely recognized and applied, but it still has great room for improvement in the task of target detection in ORSIs, especially for ship target detection.

Aiming at the low accuracy of multi-scale ship detection in ORSIs, a feature-enhanced method for ship target detection in ORSIs is proposed in this paper. The backbone network of this method adopts CSPDarknet53, and SPP module is added at the same time. The EIRM module is used for feature enhancement. SGLPANet is used for feature map fusion of different scales, and at the same time alleviates the problem of loss of detail information caused by dimensional transformation. The detection head uses YOLO Head. The main contributions of this paper are as follows:

1. Aiming at the problems of insufficient receptive field and insufficient feature information of the feature maps extracted from the backbone network, EIRM is proposed for feature enhancement. This module can capture feature information of different dimensions and provide receptive fields of different scales for feature maps, effectively improving the detection accuracy of small and medium-sized ships.
2. In view of the problem that the ReLu6 activation function may cause the gradient to be 0 in the case of negative input, Leaky ReLu is used to replace the ReLu6 activation function in the SandGlass block, and the SandGlass-L block is proposed.
3. SGLPANet is proposed based on SandGlass-L block. Using SandGlass-L to replace ordinary convolution in PANet alleviates the problem of information loss caused by channel transformation, enabling it to retain more semantic information and location information, thereby improving detection accuracy.

2. Related Work

2.1. Object Detection in Natural Images

At present, the main challenge facing the detection task is the low precision of small targets. Thus, scholars have carried out a large number of works to settle this problem. Zhang et al. [27] pointed out that the detection accuracy of small targets can be improved by appropriately improving the receptive field. When the receptive field is too small, it is easy to miss the detection of dense small targets. When the receptive field is too large, the smaller target with too large receptive field will introduce a lot of background noise, making the detection more difficult. Therefore, they propose the E-RFB (Efficient Receptive Field Block) module on the basis of RFB (Receptive Field Block) [28] to appropriately improve the receptive field of small targets, while suppressing background noise, and the accuracy of small targets is significantly improved. Safre et al. [29] found that the lightweight of depth separable convolution [30] has received wide attention, but the impact of its kernel size on accuracy has been ignored. Therefore, mixed depth separable convolution was proposed. In a single convolution, convolution kernels of different sizes are used to introduce different receptive fields to obtain information of different dimensions, and then the information is merged. Only using mixed depth separable convolution instead of ordinary convolution can achieve a great improvement in accuracy. Lim et al. [31] proposed FASSD (Feature Fusion And Attention In SSD), introduces the context feature fusion and residual attention mechanism [32] into SSD. Through the fusion of different levels of information, the network not only pays attention to the information at this level, but the context information is also taken into consideration as an additional feature. At the same time, the residual attention mechanism makes the network pay more attention to small targets. This method significantly increases the detection accuracy of small targets. Ding et al. [33] proposed DBB (Diverse Branch Block). In the training process, a convolution is integrated into four different branches, and then each branch uses different convolution to extract features, and finally the features are merged together. The convolution of each branch can be regarded as a 3×3 Convolution, so only single-branch structure is used in inference. The convolution block can achieve performance improvement without increasing any inference cost.

2.2. Object Detection in ORSIs

Different from natural images, the background of ORSIs is more complex, and most small-scale targets. Although there are problems in the application of target detection algorithm in natural images to target detection in ORSIs, researchers have carried out a lot of work to promote the development of target detection in ORSIs. Sun et al. [34] proposed AFANet (Adaptive Feature Aggregation Network). The feature maps of different layers are adaptively fused to retain only useful target information, which reduces the interference of background noise during information fusion between different layers. Meanwhile, receptive field module is added to capture different receptive fields for the feature maps of each layer. Xu et al. [35] proposed MRFF-YOLO (Multi-Receptive Field Fusion YOLO). In order to detect small targets, a detection layer of 104×104 size feature map was added, and

DenseNet [36] was applied to the detection layer to alleviate the gradient fading. MRFF-YOLO greatly reduces the missed detection rate of small targets. Wang et al. [37] proposed a detection method based on center point prediction, which solves the redundancy problem of anchor point setting in remote sensing image detection, and proposed a deformable feature pyramid network to detect targets in complex backgrounds.

The size characteristics and environmental characteristics of ship targets lead to their poor adaptability in ORSIs target detection algorithms, which also arouses the interest of many researchers. Fu et al. [38] proposed FFPN-RL (Feature Fusion Pyramid Network And Deep Reinforcement Learning) model for multi-scale dense ship target detection. First, the feature fusion pyramid is proposed for feature reuse and extracting more detailed information. Then, reinforcement learning is used to detect ships with tilted angles. Finally, rotating Soft-NMS also solves the problem of missed detection of dense ships. Wu et al. [39] used FPN for feature integration to achieve multi-scale detection, which improved the problem of difficult detection for smaller ships. For the problem of misdetection on land caused by the complex ground environment, the method of combining gradient information and gray information is used to solve the problem of misdetection of land ships. Hou et al. [40] proposed a scale-adaptive neural network, which obtains convolution feature maps through networks of different scales, adaptively learns image features, and then uses a linear SVM classifier to determine the ship's position. Li et al. [41] applied SEAs (Saliency Estimation Algorithms) in deep convolutional neural networks to extract the scale information and position information of ships, and then used the extracted features for ship detection. It has obtained better results in multi-scale ship inspection.

3. Materials and Methods

3.1. The Overall Structure of the Proposed Method

The backbone network of the algorithm in this paper uses CSPDarknet53. CSPDarknet53 includes five CSP modules, and each CSP module superimposes several residual blocks (1, 2, 8, 8, 4), and the backbone network is downsampling for the fifth time. The extracted large-size feature maps are first input into the SPP module, and the small and medium-size feature maps extracted by the fourth and third downsampling are input into EIRM, and then the three-size feature maps are used as the input of SGLPNet for feature fusion, and the feature map after feature fusion enter the detection head for detection. Assuming that the input image size is 416×416 , then three different sizes of feature maps (13×13 , 26×26 , 52×52) are used to detect large, medium and small objects, respectively. Figure 1 shows the proposed method overall architecture. As shown in the figure, the backbone network first performs feature extraction on the input image, the extracted features are enhanced by EIRM and SPP modules, then multi-scale feature fusion is performed by SGLPNet, and finally detected by the YOLO Head detection head. Figure 2 shows the module structure. CBL stands for Convolution, BN, and Leaky ReLU activation functions. CBM stands for Convolution, BN, Mish Activation Function. CSP $\times$ N means that the CSP module contains N residual blocks. The SPP module contains three max-pooling layers with kernel sizes of 5, 9, and 13 and the identity branch.

3.2. EIRM

3.2.1. Inception-ResNet-A Module and Elastic Mechanism

In contrast with the conventional residual module, the IRA(Inception-ResNet-A) [42] module uses a multi-branch structure to increase the depth and width of the network. The three branches have different depths, and each branch uses a different number of convolutions to obtain different receptive fields, extract feature information of different dimensions, and then perform concatenating operations on the feature maps of different branches. After adjusting the dimension through 1×1 convolution, the residual connection is made with the input feature map. The residual connection accelerates the convergence speed of the network and alleviates the problem of gradient disappearance. Figure 3 [42]

shows the Inception-ResNet-A module structure, 1 and 3 represents the convolution kernel size, + represents the residual connection.

Scale transformation is a difficult problem for target detection, especially for the detection of small and medium targets, the scale problem is particularly important. Although there have been many methods attempts to solve multi-scale target detection, such as PANet, these methods are general and fixed strategies, which are not friendly to multi-scale ship detection. To resolve this problems, Wang et al. [43] proposed an elastic mechanism. Elastic mechanism improves the accuracy of multi-scale detection by learning scales from training data instead of manual integration, enabling feature maps of different scales to obtain more suitable receptive fields. The elastic mechanism introduces a residual branch, which uses downsampling and upsampling to perform scaling strategies. At the same time, the module layer has a series of convolution operations to extract feature information, so that each layer of feature map has multiple scales of feature information, and this strategy does not introduce additional parameters and calculations. Figure 4 [43] is a structural diagram of the elastic branch. Operation represents a series of convolution operations.

Figure 1. The overall structure of the proposed method. SGL stands for SandGlass-L block.

Figure 2. Module structure diagram.

Figure 3. Inception-ResNet-A module structure, + stands for residual connection.

Figure 4. Elastic branch structure diagram.

The parameters in the branch are shown in Formula (1). $U_{ri}(x)$ and $D_{ri}(x)$ represent upsampling and downsampling functions, respectively. σ is a non-linear function, q is the count of branches, and $T_i(x)$ is convolution, BN and the activation function.

$$F(x) = \sigma \left(\sum_{i=1}^q U_{ri}(T_i(D_{ri}(x))) \right) \quad (1)$$

3.2.2. EIRM Structure

The Inception-ResNet-A structure uses three different convolution branches to extract different receptive fields and feature information to improve the accuracy of target detection. However, through our experiments, it is found that only using the Inception-ResNet-A module to improve the fixed receptive field is far not enough. When the receptive field of ship targets is large, a large amount of background noise will be introduced, which is not conducive to the detection of small and medium-sized ship targets. When the receptive field of the target is small, only a part of the target can be seen, which may easily lead to missed detection and false detection. Therefore, we proposed EIRM, adding flexible branches to the Inception-ResNet-A module, and adding the multi-branch structure in the Inception-ResNet-A module to the flexible branches. Specifically, the elastic branch first performs downsampling, then extracts features through a three-branch structure, and then concatenating the extracted features, and finally performs a scaling strategy through upsampling. The EIRM adjusts the scales flexibly through training data, and can add receptive fields of different scales to each layer of feature maps to provide suitable receptive fields for small and medium ship targets. At the same time, the multi-branch structure also provides feature maps with different information, improve the detection accuracy of medium and small targets. Figure 5 shows the proposed EIRM. The elastic branch on the left and the Inception-ResNet-A module on the right.

Figure 5. EIRM structure diagram.

Formula (2) can be used to express EIRM:

$$F(x) = E(x) + G(x) \quad (2)$$

The IRA (Inception-ResNet-A) module $G(x)$ can be represented by Formula (3):

$$G(x) = g(x) + x \quad (3)$$

The IRA convolution branch $g(x)$ can be represented by Formula (4):

$$g(x) = f^{1 \times 1} \left(C \left(f^{1 \times 1}(x); f^{3 \times 3} \left(f^{1 \times 1}(x) \right); f^{3 \times 3} \left(f^{3 \times 3} \left(f^{1 \times 1}(x) \right) \right) \right) \right) \quad (4)$$

According to Formulas (1)–(4), the Formula (5) of EIRA can be expressed as:

$$F(x) = U_{ri}(g(D_{ri}(x))) + g(x) + x \quad (5)$$

x represents the input of the module (and residual branch), $E(x)$ represents the elastic branch, $G(x)$ represents the IRA module, $g(x)$ is the IRA convolution branch, $f^{1 \times 1}(x)$ represents 1×1 convolution, C represents Concat, $D_{ri}(x)$ represents the downsampling function, $U_{ri}(x)$ represents the upsampling function. According to the construction formula, it can be seen that EIRM retains the multi-dimensional characteristics of IRA, and at the same time, the existence of elastic mechanism enables it to provide a suitable receptive field.

3.3. SandGlass-L Block

At present, the design of lightweight network is mainly realized by stacking inverted residual blocks [44]. The inverted bottleneck structure makes it possible to extract more information, and the existence of deep separable convolution enables the model to obtain higher accuracy with less parameters. However, the bottleneck structure of the inverted residual block that first raises the dimensionality and then reduces the dimensionality may bring about the problem of information loss. In response to the above problems, Zhou et al. [45] improved the shortcomings of the inverted residual block and proposed SandGlass block. The channel is reduced in low dimensions, and identity mapping and

spatial conversion are performed while extracting information at high latitudes, which alleviates the loss of information and improves the accuracy of target detection. At the same time, the amounts of parameters and calculations have not increased compared with the inverted residual block.

The construction Formula (6) of the SandGlass block is as follows. Figure 6 [45] shows the SandGlass block's structure. The input feature map first undergoes a 3×3 depthwise separable convolution, and then the dimensions are adjusted using two 1×1 convolutions. Finally, a 3×3 depthwise separable convolution is passed. + Stands for residual connection. The operation of the SandGlass block is shown in Table 1 [45]. t represents the ratio of channel expansion and reduction, s is the step size, DSC represents the depth separable convolution. The first two parameters in Input represent the size of the convolution kernel, and the last one is the dimension.

$$G = \phi_e(\phi_r(F)) + F \quad (6)$$

Figure 6. Structural drawing of SandGlass block.

Table 1. The SandGlass block's operation.

Input	Operator	Output
$D_f \times D_f \times M$	3×3 DSC, ReLu6	$D_f \times D_f \times M$
$D_f \times D_f \times M$	1×1 Conv, Linear	$D_f \times D_f \times \frac{M}{t}$
$D_f \times D_f \times \frac{M}{t}$	1×1 Conv, ReLu6	$D_f \times D_f \times N$
$D_f \times D_f \times N$	3×3 DSC, Linear	$D_f \times D_f \times \frac{N}{s}$

Without considering the depthwise convolution and activation function, if the input tensor is $F \in \mathbb{R}^{D_f \times D_f \times M}$, then the output tensor is $G \in \mathbb{R}^{D_f \times D_f \times M}$, where ϕ_e and ϕ_r represent the pointwise convolution of channel expansion and reduction, respectively.

Due to the accuracy limitation of the mobile terminal network, the SandGlass block uses ReLu6 [44] as the non-linear activation function. ReLu6 can maintain good numerical resolution at low precision numerical values, but when the input is negative, it is easy to have a gradient of 0. Therefore, we use Leaky ReLu [46] instead of ReLu6. Leaky ReLu still has a slope when the function input is negative, solving the problem that the gradient is 0 in the case of negative input. Formula (6) is the function of ReLu6, Formula (7) is the function of Leaky ReLu. The operation of the SandGlass-L block is shown in Table 2. The channel reduction rate t is 2, and the step size s is 1.

$$F(x) = \max(0, x)x \leq 6 \quad (7)$$

$$F(x) = \max(\alpha x, x)\alpha = 0.01 \quad (8)$$

Table 2. The operation of the SandGlass-L block.

Input	Operator	Output
$D_f \times D_f \times M$	3×3 DSC, Leaky ReLu	$D_f \times D_f \times M$
$D_f \times D_f \times M$	1×1 Conv, Linear	$D_f \times D_f \times \frac{M}{2}$
$D_f \times D_f \times \frac{M}{2}$	1×1 Conv, Leaky ReLu	$D_f \times D_f \times M$
$D_f \times D_f \times M$	3×3 DSC	$D_f \times D_f \times M$

3.4. SGLPANet

PANet’s high-level feature map has detailed semantic information, and the low-level feature map has location information. Then, the bottom-up and top-down paths are used to propagate semantic information to low-level feature maps and position information to high-level feature maps. Feature maps of different levels are merged to enrich feature information and improve network robustness. PANet uses five convolution blocks to further extract features after each feature fusion. It contains two pairs of 1×1 and 3×3 convolutions, and the last 1×1 convolution is used to adjust the dimension. In the two pairs of convolutions, the 1×1 convolution is used to reduce the dimension, and the 3×3 convolution is used for feature extraction while increasing the dimension. However, using an inverted structure leads to loss of feature information. Therefore, we keep the bottom-up and top-down feature fusion operations, and then replace the two sets of 1×1 and 3×3 convolutions in PANet with two SandGlass-L blocks. The existence of depthwise separable convolution in SandGlass-L can greatly reduce the amount of computation, and at the same time alleviate the problem of information loss caused by convolution operations. Compared with PANet, SGLPANet uses fewer parameters and extracts and retains more feature information.

4. Experimental Results and Analysis

4.1. Dataset and Evaluation Metric

4.1.1. Dataset Description

We conducted an experimental comparison on the NWPU VHR-10 [47] and LEVIR [48] datasets to prove the effectiveness of the method in this paper.

The NWPU VHR-10 dataset contains 10 types of objects: airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track and field, harbor, bridge, and vehicles. There are 800 ultra-high resolution ORSIs, 150 of which do not contain any samples, and the remaining 650 contain positive samples. The images are cropped from the Google Earth and Vaihingen datasets and then manually annotated by experts. We select 650 images that contain samples for experimentation, and randomly divide the test set and training set according to a ratio of 5 to 5.

The LEVIR dataset contains 21,952 pictures, and the images size is 600×800 pixels. The LEVIR dataset contains three categories: airplane, ship, and storage tank. We trained and verified the algorithm according to the original division method. The training set contains 13,171 images and the verification set contains 8781 images.

4.1.2. Evaluation Metric

The evaluation indicators of this experiment are mAP [26] (mean Average Precision) and loss curve. The value of mAP is determined by the recall rate R and the precision P. The calculation of mAP is shown in Formula (12). The loss curve shows the difference between the predicted result and the actual result. The faster the loss curve drops, the lower the loss value, the better the effect of the model.

$$P = \frac{TP}{FP + TP} \quad (9)$$

$$R = \frac{TP}{FN + TP} \quad (10)$$

$$AP = \int_0^1 p(x) dx \quad (11)$$

$$mAP = \frac{\sum_{i=1}^K AP_i}{K} \quad (12)$$

Among them, TP represents positive sample, FP represents negative sample, and FN represents false negative. In this experiment, the undetected target is FN, the IOU higher than 0.5 is a positive sample, and the IOU lower than 0.5 is a negative sample.

The loss function adopts DIOU [49], DIOU adds a penalty term to IOU [50] to minimize the standardized distance between the center points of the two detection frames and accelerate the convergence of loss. IOU is shown in Formula (13). DIOU is shown in Formula (14):

$$L_{IOU} = 1 - IOU(A, B) \quad (13)$$

$$L_{DIOU} = L_{IOU} + \frac{\rho^2(a, a^{gt})}{b^2} \quad (14)$$

Among them, IOU (A, B) is the difference between the predicted bounding box A and the true bounding box B. In the penalty term, a is the coordinate of the center point of the predicted bounding box, a^{gt} is the coordinate of the center point of the true bounding box, and b is the diagonal length of the minimum bounding box.

4.2. Experimental Details

The experiment runs in the hardware environment of RTX3060 GPU and Intel i7 CPU, using the window 10 operating system, using 11.3 version of CUDA and 8.0.5 version of cudnn acceleration. Parameter settings: Momentum is 0.949, initial learning rate is 0.001, learning rate is adjusted to 0.0001 for 8000 iterations, learning rate for 9000 iterations is adjusted to 0.00001, the total number of iterations is 10,000, the input image size is 416×416 , and the batch size is set to 32.

4.3. Experimental Results

4.3.1. Experimental Results and Comparative Analysis

Figure 7 shows the loss image of YOLO V4 and the proposed method on NWPU -VHR 10 dataset. According to the loss image, compared with YOLO v4, the loss curve of the algorithm in this paper decreases more smoothly, the loss value is lower, and the mAP is higher. It shows that the proposed method is greatly improved.

Figure 7. The (left) side is YOLO v4, the (right) side is the algorithm of this paper.

Figure 8 shows the accuracy comparison results of YOLO v4 and the 10 categories of our method. AI, SH, ST, BD, TC, BC, GTF, HA, BR, VE represent airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, vehicle. It can be seen that the overall mAP of this method reaches the advanced level, which is 1.43% higher than the baseline YOLO v4 method. Out of 10 categories, eight of the proposed methods have higher accuracy than the baseline network, and these eight categories include large, medium and small-scale objects. Among them, the ship accuracy is 3.24% higher than the baseline. The experimental results show that the algorithm in this paper can better improve the accuracy of ship target detection. Moreover, the expansion of enrollment to other classes has achieved good results for large, medium and small goals. However, in the harbor class and the vehicle class, the accuracy of the algorithm in this paper is poor. Mainly because harbor is a large target and is densely distributed, one anchor box can often select multiple targets, and the existence of non-maximum suppression algorithm may lead to missed detection of such targets. The improvement of the receptive field and feature information has a limited impact on the detection accuracy of such objects. Vehicles generally belong to very small targets. Although the structure of SGLPANet is lighter and can retain more feature information, the depthwise separable convolution may have channel information loss, which may reduce the accuracy of very small target detection. At present, the detection accuracy of extremely small objects still needs to be improved.

Figure 8. Comparison of the results of each category between YOLO v4 and Ours.

Table 3 shows the comparison between the proposed method and the current advanced methods. AP_1 to AP_{10} are airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle. It can be seen that the overall mAP of this method reaches the state-of-the-art. At the same time, four out of 10 categories achieved the highest level.

Figure 9 shows the comparison between the detection results of proposed method and the YOLO v4. a is the original image, YOLO v4 detection result is b. Our detection result is c. The thick ellipse frame is the missed detection and false detection of YOLO v4. Generally speaking, the high-resolution branch pays more attention to the edge information of the target, and the low-resolution branch is more sensitive to the main information of the target. Insufficient subject information often results in false detection, and lack of feature information may result in missed detection. It can be clearly seen from a, b, c that the missed detection and false detection of the algorithm in this paper are significantly reduced. It is proved that the algorithm in this paper is effective in enhancing feature information and retaining more feature information.

Table 3. Comparison results with other methods.

Methods	AP ₁	AP ₂	AP ₃	AP ₄	AP ₅	AP ₆	AP ₇	AP ₈	AP ₉	AP ₁₀	mAP
R-FCN [51]	99.80	80.82	90.48	97.88	90.69	72.38	98.99	87.18	70.44	88.62	87.74
HitDet [52]	99.36	77.26	90.66	98.46	88.72	75.51	95.89	65.80	60.00	85.93	83.76
FCOS [53]	90.47	73.72	90.36	98.94	89.38	80.82	96.74	87.91	61.92	88.16	85.84
Foveabox [54]	99.49	75.22	89.50	98.14	92.65	50.05	96.58	41.70	63.50	86.27	79.31
Dong et al. [55]	90.8	80.5	59.2	90.8	80.8	90.9	99.8	90.3	67.8	78.1	82.9
MS-FF [56]	95.79	72.50	70.90	97.83	85.62	97.20	98.82	92.40	81.74	64.64	85.64
NEOON [57]	78.29	81.68	94.62	89.74	61.25	65.04	93.23	73.15	59.46	78.26	77.50
HRBM [58]	99.70	90.80	90.61	92.91	90.29	80.13	90.81	80.29	68.53	87.14	87.12
SHDET [59]	100	81.36	90.90	98.66	90.84	82.57	98.68	91.11	76.43	89.82	90.04
YOLO v4	99.9	89.66	98.33	97.14	97.46	93.39	99.82	84.32	75.92	93.41	92.94
Ours	99.92	92.8	98.3	97.49	98.58	96.7	99.9	83.83	83.41	92.75	94.37

4.3.2. Ablation Experiment

We conduct ablation experiments on the NWPU VHR-10 dataset to demonstrate the feasibility and necessity of the proposed module. We compared a total of five combinations, namely YOLO v4, baseline + IRA + PANet, baseline + EIRM + PANet, baseline + SGPANet, baseline + SGLPANet, Ours (baseline + EIRM + SGLPANet). According to the results in Table 4, it can be seen that the accuracy of the improved module in this paper has been improved compared with the initial module. The algorithm in this paper achieves the highest accuracy and the model complexity is reduced compared to YOLO v4. Overall, we achieved an accuracy improvement of 1.43% without increasing model complexity.

Table 4. Ablation experiment. SGPANet represents use SandGlass block. SGLPANet mean uses SandGlass-L block. IRA is the Inception ResNet-A.

Baseline (CSPDarknet53 + SPP)	mAP	Inference Time	BFLOPS
PANet (YOLO v4)	92.94%	0.015	59.628
IRA, PANet	93.87%	0.015	65.720
EIRM, PANet	94.12%	0.018	67.338
SGPANet	93.93%	0.018	48.402
SGLPANet	93.23%	0.018	48.402
Ours	94.37%	0.018	56.112

4.4. Extended Experiment

We conducted extended experiments on the LEVIR dataset. Table 5 shows the comparison results. According to the results, our mAP as well as the accuracy of individual classes are improved compared to YOLO v4. The small and medium-sized targets in the ship class of the LEVIR dataset account for a large proportion and are sparsely distributed, which also verifies that the algorithm in this paper has a better detection effect on small and medium-sized ship targets. At the same time, the aircraft and storage tank categories also have the above characteristics, so the algorithm in this paper also achieves good results. Figure 10 is a comparison of proposed method and YOLO v4 loss images. According to the loss image, it can be seen that the loss value in this paper is lower and can achieve higher accuracy.

Figure 9. Cont.

Figure 9. Comparison of the detection effect of YOLO v4 and Ours. (a1–a6) is the original image, (b1–b6) is the YOLO v4 detection result, (c1–c6) is the detection result of this algorithm.

Table 5. LEVIR dataset comparison results.

	YOLO v4	Ours
airplane	76.61%	77.82%
ship	71.25%	72.74%
storage tank	79.71%	82.05%
mAP	75.86%	77.54%

Figure 11 shows the comparison between the detection results of YOLO v4 and ours. The a is the original image. YOLO v4 detection result is b. Ours detection result is c. The thick ellipse is the missed detection of YOLO v4. It can be seen from c1–c5 that the algorithm in this paper can well detect the ship targets with serious occlusion, small scale and very similar texture to the background texture, and the missed detection situation is greatly improved. It can be seen from c6 and c7 that the algorithm in this paper has also improved the detection accuracy of occluded airplanes and small-sized storage tanks.

Figure 10. YOLO v4 on the (left) and our method on the (right).

Figure 11. Cont.

Figure 11. Cont.

Figure 11. Comparison of the detection effect of our method and YOLO v4. (a1–a7) is the original image, (b1–b7) is the YOLO v4 detection result, (c1–c7) is the detection result of this algorithm.

5. Conclusions

Due to the complex background, diverse target scales and limited resolution, the detection of ship targets in ORSI faces many difficulties. A feature-enhanced ship target detection algorithm is proposed to solve the problem of low detection accuracy of multi-scale ship targets. EIRM is used for feature enhancement, which provides a suitable receptive field for the feature map and captures feature information of different dimensions. The SandGlass block is then improved while proposing SGLPNet to extract and retain more detailed feature information. Experiments on the NWPU VHR-10 dataset show that the proposed algorithm outperforms YOLO v4 and some existing state-of-the-art algorithms. The extended experiments on the LEVIR dataset show that the algorithm in this paper has certain applicability on different datasets. Our algorithm has achieved good results in ship detection, especially for small and medium-sized ships, and has also achieved satisfactory results in the detection of some categories. However, the improvement of the algorithm in this paper is mainly aimed at the multi-scale characteristics of ship targets and mainly for small and medium-scale targets, and does not consider the high aspect ratio characteristics of ships, so the algorithm in this paper has certain universality but weak pertinence. Although the complexity of the algorithm in this paper has been reduced, the network model has become more bloated, and the detection speed needs to be improved. Therefore, in the following work we will further improve the proposed algorithm in two points. Firstly, how to adapt the network to the aspect ratio characteristics of the ship target. Secondly, how to make the network model more lightweight while achieving faster detection speed without losing accuracy.

Author Contributions: Conceptualization, L.Z. and Y.L.; methodology, L.Z.; software, Y.L. and X.R.; validation, L.Z., Y.L. and Y.W.; formal analysis, L.Z. and X.Z.; writing—original draft preparation, L.Z., Y.L. and B.Q.; writing—review and editing, Y.L., Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by grants from National Basic Research Program of China (Grant number 2019YFE0126600); the Major Project of Science and Technology of Henan Province (Grant number 201400210300); the Key Research and Promotion Projects of Henan Province (Grant numbers 212102210393; 202102110121; 222102210151) and Kaifeng science and technology development plan (Grant number 2002001); the Key Scientific and Technological Project of Henan Province (Grant number 212102210496); the National Natural Science Foundation of China (62106066), and the Key Research Projects of Henan Higher Education Institutions (22A520019).

Acknowledgments: The authors sincerely thank anonymous reviewers for critical comments and suggestions on the manuscript.

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

References

1. Han, G.U.O. Gaojing No. 1 is officially commercially available, and China's commercial remote sensing has entered the 0.5 meter era. *Satell. Appl.* **2017**, *5*, 62–63.
2. Zhang, C.; Tao, R. Research progress on optical remote sensing object detection based on CNN. *Spacecr. Recovery Remote Sens.* **2020**, *41*, 45–55.
3. Wang, W. Overview of ship detection technology based on remote sensing images. *Telecommun. Eng.* **2020**, *60*, 1126–1132.
4. Liu, T. Deep learning based object detection in optical remote sensing image: A survey. *Radio Commun. Technol.* **2020**, 624–634.
5. Qu, Z.; Zhu, F.; Qi, C. Remote sensing image target detection: Improvement of the YOLOv3 model with auxiliary networks. *Remote Sens.* **2021**, *13*, 3908. [\[CrossRef\]](#)
6. 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.
7. 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.
8. Redmon, J.; Farhadi, A. Yolo3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
9. 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 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
10. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
11. Li, Z.; Zhou, F. FSSD: Feature fusion single shot multibox detector. *arXiv* **2017**, arXiv:1712.00960.
12. Yang, J.; Wang, L. Feature fusion and enhancement for single shot multibox detector. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 2766–2770.
13. 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 (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
14. Li, X.; Shang, M.; Qin, H.; Chen, L. Fast accurate fish detection and recognition of underwater images with fast r-cnn. In Proceedings of the OCEANS 2015—MTS/IEEE Washington, Washington, DC, USA, 19–22 October 2015; pp. 1–5.
15. Qian, R.; Liu, Q.; Yue, Y.; Coenen, F.; Zhang, B. Road surface traffic sign detection with hybrid region proposal and fast R-CNN. In Proceedings of the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Changsha, China, 13–15 August 2016; pp. 555–559.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; Volume 28, pp. 91–99.
17. Mhalla, A.; Chateau, T.; Gazzah, S.; Ben Amara, N.E. Scene-specific pedestrian detector using monte carlo framework and faster r-cnn deep model: Phd forum. In Proceedings of the 10th International Conference on Distributed Smart Camera, New York, NY, USA, 12–15 September 2016; pp. 228–229.
18. Zhai, M.; Liu, H.; Sun, F.; Zhang, Y. Ship detection based on faster R-CNN network in optical remote sensing images. In Proceedings of the 2019 Chinese Intelligent Automation Conference, Jiangsu, China, 20–22 September 2019; pp. 22–31.
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
20. Zhao, T.; Yang, Y.; Niu, H.; Wang, D.; Chen, Y. Comparing U-Net convolutional network with mask R-CNN in the performances of pomegranate tree canopy segmentation. In Proceedings of the Multispectral, Hyperspectral, and Ultraspectral Remote Sensing Technology, Techniques and Applications VII, Honolulu, HI, USA, 24–26 September 2018; Volume 10780, p. 107801J.
21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
22. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

23. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
24. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [[CrossRef](#)]
26. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
27. Zhang, J.; Zhao, Z.; Su, F. Efficient-receptive field block with group spatial attention mechanism for object detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 3248–3255.
28. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
29. Tan, M.; Le, Q.V. Mixnet: Mixed depthwise convolutional kernels. *arXiv* **2019**, arXiv:1907.09595.
30. Sifre, L.; Mallat, P.S. Rigid-Motion Scattering for Image Classification. Ph.D. Thesis, Ecole Polytechnique, Palaiseau, France, 2014.
31. Lim, J.S.; Astrid, M.; Yoon, H.J.; Lee, S.I. Small object detection using context and attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), Jeju Island, Korea, 13–16 April 2021; pp. 181–186.
32. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
33. Ding, X.; Zhang, X.; Han, J.; Ding, G. Diverse branch block: Building a convolution as an inception-like unit. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10886–10895.
34. Sun, W.; Zhang, X.; Zhang, T.; Zhu, P.; Gao, L.; Tang, X.; Liu, B. Adaptive feature aggregation network for object detection in remote sensing images. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 1635–1638.
35. Xu, D.; Wu, Y. MRFF-YOLO: A multi-receptive fields fusion network for remote sensing target detection. *Remote Sens.* **2020**, *12*, 3118. [[CrossRef](#)]
36. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
37. Wang, J.; Yang, L.; Li, F. Predicting arbitrary-oriented objects as points in remote sensing images. *Remote Sens.* **2021**, *13*, 3731. [[CrossRef](#)]
38. Fu, K.; Li, Y.; Sun, H.; Yang, X.; Xu, G.; Li, Y.; Sun, X. A ship rotation detection model in remote sensing images based on feature fusion pyramid network and deep reinforcement learning. *Remote Sens.* **2018**, *10*, 1922. [[CrossRef](#)]
39. Wu, Y.; Ma, W.; Gong, M.; Bai, Z.; Zhao, W.; Guo, Q.; Chen, X.; Miao, Q. A coarse-to-fine network for ship detection in optical remote sensing images. *Remote Sens.* **2020**, *12*, 246. [[CrossRef](#)]
40. Hou, X.; Xu, Q.; Ji, Y. Ship detection from optical remote sensing image based on size-adapted CNN. In Proceedings of the 2018 Fifth International Workshop on Earth Observation and Remote Sensing Applications (EORSA), Xi'an, China, 18–20 June 2018; pp. 1–5.
41. Li, Z.; You, Y.; Liu, F. Analysis on saliency estimation methods in high-resolution optical remote sensing imagery for multi-scale ship detection. *IEEE Access* **2020**, *8*, 194485–194496. [[CrossRef](#)]
42. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
43. Wang, H.; Kembhavi, A.; Farhadi, A.; Yuille, A.L.; Rastegari, M. Elastic: Improving cnns with dynamic scaling policies. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2258–2267.
44. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
45. Zhou, D.; Hou, Q.; Chen, Y.; Feng, J.; Yan, S. Rethinking bottleneck structure for efficient mobile network design. In Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 680–697.
46. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. *Computer Science*. **2013**, *30*, 3.
47. 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](#)]
48. Zou, Z.; Shi, Z. Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. *IEEE Trans. Image Processing A Publ. IEEE Signal Processing Soc.* **2018**, *27*, 1100–1111. [[CrossRef](#)]
49. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12993–13000.
50. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. UnitBox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, New York, NY, USA, 15–19 October 2016.

51. Dai, J.; Li, Y.; He, K.; Sun, J. Object detection via region-based fully convolutional networks. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016.
52. Guo, J.; Han, K.; Wang, Y.; Zhang, C.; Yang, Z.; Wu, H.; Chen, X.; Xu, C. Hit-detector: Hierarchical trinity architecture search for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11405–11414.
53. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 9627–9636.
54. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. Foveabox: Beyond anchor-based object detection. *IEEE Trans. Image Processing* **2020**, *29*, 7389–7398. [[CrossRef](#)]
55. Chen, H.; Zhang, L.; Ma, J.; Zhang, J. Target heat-map network: An end-to-end deep network for target detection in remote sensing images. *Neurocomputing* **2019**, *331*, 375–387. [[CrossRef](#)]
56. Zhang, W.; Jiao, L.; Liu, X.; Liu, J. Multi-scale feature fusion network for object detection in vhr optical remote sensing images. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 330–333.
57. Xie, W.; Qin, H.; Li, Y.; Wang, Z.; Lei, J. A novel effectively optimized one-stage network for object detection in remote sensing imagery. *Remote Sens.* **2019**, *11*, 1376. [[CrossRef](#)]
58. 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](#)]
59. Zhu, D.; Xia, S.; Zhao, J.; Zhou, Y.; Niu, Q.; Yao, R.; Chen, Y. Spatial hierarchy perception and hard samples metric learning for high-resolution remote sensing image object detection. *Appl. Intell.* **2021**, *52*, 3193–3208. [[CrossRef](#)]