remote sensing

MDPI

Article

A Multi-Feature Fusion and Attention Network for Multi-Scale Object Detection in Remote Sensing Images

Yong Cheng [1], Wei Wang [2], Wenjie Zhang [3,*], Ling Yang [1], Jun Wang [1], Huan Ni [4], Tingzhao Guan [1], Jiaxin He [2], Yakang Gu [1] and Ngoc Nguyen Tran [5,6]

[1] School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China
[2] School of Automation, Nanjing University of Information Science & Technology, Nanjing 210044, China
[3] School of Geographical Sciences, Nanjing University of Information Science & Technology, Nanjing 210044, China
[4] School of Remote Sensing & Geomatics Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China
[5] School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi 100803, Vietnam
[6] School of Life Science, University of Technology Sydney, Ultimo 2007, Australia
* Correspondence: zhangwenjie@nuist.edu.cn

Abstract: Accurate multi-scale object detection in remote sensing images poses a challenge due to the complexity of transferring deep features to shallow features among multi-scale objects. Therefore, this study developed a multi-feature fusion and attention network (MFANet) based on YOLOX. By reparameterizing the backbone, fusing multi-branch convolution and attention mechanisms, and optimizing the loss function, the MFANet strengthened the feature extraction of objects at different sizes and increased the detection accuracy. The ablation experiment was carried out on the NWPU VHR-10 dataset. Our results showed that the overall performance of the improved network was around 2.94% higher than the average performance of every single module. Based on the comparison experiments, the improved MFANet demonstrated a high mean average precision of 98.78% for 9 classes of objects in the NWPU VHR-10 10-class detection dataset and 94.91% for 11 classes in the DIOR 20-class detection dataset. Overall, MFANet achieved an mAP of 96.63% and 87.88% acting on the NWPU VHR-10 and DIOR datasets, respectively. This method can promote the development of multi-scale object detection in remote sensing images and has the potential to serve and expand intelligent system research in related fields such as object tracking, semantic segmentation, and scene understanding.

Keywords: remote sensing images; multi-scale object detection; multi-feature fusion and attention network; multi-branch convolution; attention mechanism; loss function

Citation: Cheng, Y.; Wang, W.; Zhang, W.; Yang, L.; Wang, J.; Ni, H.; Guan, T.; He, J.; Gu, Y.; Tran, N.N. A Multi-Feature Fusion and Attention Network for Multi-Scale Object Detection in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 2096. https://doi.org/10.3390/rs15082096

Academic Editors: Jukka Heikkonen, Fahimeh Farahnakian and Pouya Jafarzadeh

Received: 21 February 2023
Revised: 7 April 2023
Accepted: 12 April 2023
Published: 16 April 2023

1. Introduction

The multi-scale object feature recognition of remote sensing images plays a vital role in many fields, including military and civilian. In the military field, remote sensing images can be used to detect and identify ships at sea and then analyze the locations of ship objects to ensure naval defense security [1,2]. In the civilian sector, they can help predict changes in animal habitats and environmental quality [3,4]. Object detection technology in remote sensing images is significant for ocean monitoring, weather monitoring, military navigation, urban planning, and layout. Therefore, how to further improve the multi-scale object detection of remote sensing images has become the focus of research.

Multi-scale object detection in remote sensing images is generally performed in high-resolution images, which provides high-definition information on object features. Among multi-scale objects, large and medium ones are more feature-rich and easier to detect. However, small objects are generally difficult to be characterized and detected effectively. Early

object detection methods in remote sensing images mainly used manual feature modeling combined with classifiers for classification to determine the object class. However, traditional detection methods were inefficient and costly in terms of time and labor, making it difficult to meet the needs of practical applications [5]. In recent years, the wide application of deep learning in the fields of video processing [6], image set classification [7,8], image encryption [9], and image recognition [10] has made rapid and accurate object feature recognition possible. Convolutional neural networks, such as RCNN [11] and Faster RCNN [12], have been able to extract high-level semantics from images with better robustness and expressiveness than artificial features, significantly improving detection accuracy. Zhu et al. [13] proposed a multi-layer feature fusion model based on Faster RCNN to improve small object characterization effectively. Shivappriya et al. [14] introduced the additive activation function into Faster RCNN to solve model overfitting problems and improve recognition performance. Although these algorithms improved the detection accuracy of remote sensing images to some extent, Faster RCNN had more parameters and a slower detection speed because it was a two-stage detection algorithm that needed to extract the candidate regions first and then classify and regress the candidate regions.

Single-stage detection algorithms are generally preferred, with representative algorithms such as You Only Look Once (YOLO) [15–19]. This algorithm obtains prediction results directly from the input image. It transforms the object detection problem into a regression problem, significantly improving the detection speed and meeting the demand for real-time detectability of remote sensing images. However, the semantic information in the YOLO about shallow features is weak. After high-fold features compress the input image in the deeper convolution layers, some object information is lost and the sensitivity of detecting objects will gradually decrease. Therefore, Laban et al. [20] proposed the method of an anchor expansion based on YOLOv3 to improve the ability to detect small category goals. Hong et al. [21] improved the anchor frame based on the K-means algorithm with linear scaling while introducing Gaussian parameters into YOLOv3 to enhance the accuracy of multi-scale object detection. Zhou et al. [22] introduced a frequency channel attention network in YOLOv5 to detect small targets in remote sensing images. For object multi-scale variation and dense object distribution characteristics, Wang et al. [23] designed the SPB module and PANet sampling strategy based on YOLOv5. The mean average precision (mAP) was improved by 5.3% compared with the baseline. To address the difficulty of small target object detection in remote sensing images, Han et al. [24] improved YOLO by increasing the residual connection and cross-layer attention to enhance the detection ability of the model for small targets in remote sensing images. Wu et al. [25] proposed the combination of a transformer encoder and a reparameterized backbone based on YOLOX, which effectively improved dense oil tank detection and classification. To enhance the feature learning ability of the network, Yang et al. [26] used efficient channel attention in YOLOX and combined adaptively spatial feature fusion with the neck network to finally achieve high-accuracy object detection in remote sensing images.

YOLOX, the latest version, has improved detection accuracy and speed, and its performance has reached new heights [19]. The anchorless-based network does not require manual anchor scale and aspect ratio setting. It is more suitable for remote sensing images and multi-target detection, for which YOLOX is chosen to conduct further research in this paper. YOLOX uses ordinary convolution, which is imperfect for verifying high-dimensional semantic information, for feature extraction. The problem of partial loss of object information in deep features can lead to lower detection accuracy. The improved feature map scaling and design feature fusion achieved good detection results in the multi-scale object detection task [27]. Therefore, this paper explored further research using YOLOX to improve multi-scale object detection accuracy. To achieve this, the paper proposed a multi-feature fusion and attention network (MFANet) based on YOLOX that effectively detected multi-scale objects while considering the detection accuracy of small objects. The study showed that multi-scale high-level feature extraction and multi-layer pyramidal feature fusion are effective for more accurate target detection. The proposed MFANet incorporated

RepVGG, detail channels, Res-RFBs, and CA modules to help the model extract remote sensing objects more accurately. The paper presented ablation and comparison experiments on two publicly available datasets, NWPU VHR-10 and DIOR. It showed that MFANet achieves 96.63% and 87.88%, respectively, and could handle multi-scale object detection tasks on remote sensing images better than existing methods.

This paper proposed a multi-feature fusion and attention network (MFANet) based on YOLOX to enhance multi-scale object detection accuracy in remote sensing images. The main contributions of this paper included:

(1) To enhance feature extraction of multi-scale objects in remote sensing images, MFANet introduced a structurally reparameterized VGG-like technique (RepVGG) to reparameterize a new backbone and improve multi-object detection accuracy without increasing computation time.
(2) Detailed enhancement channels were introduced in path aggregation feature pyramid networks (PAFPN) to express a great deal of object information. Combining with residual connections, this paper formed a new multi-branch convolutional module (Res-RFBs) to improve the recognition rate of multi-scale objects in remote sensing images. The coordinate attention (CA) mechanism was introduced to reduce the interference of background information and enhance the perception of remote sensing objects by the neural network.
(3) To address the shortcomings of the baseline in the object localization and identification problem, generalized intersection over union (GIoU) was used to optimize the loss, speed up the convergence of the model, and reduce the target miss rate.

2. Methods

2.1. The Structure of the Network

In 2021, Megvii Inc. (Beijing, China) proposed a new object detection network, YOLOX, which exceeds the performance of YOLOv3 and has certain advantages compared with YOLOv5 [19]. The algorithm does not use anchor points and performs dynamic sample matching for objects of different sizes, integrating the previous data enhancement and decoupled head. Its detection speed and effectiveness are improved. First, an image of size 640 × 640 is used as the input layer, and data enhancement is performed using Mosaic and Mixup. The pre-processed image is then fed into the CSPDarknet53 backbone for feature extraction, resulting in three feature layers with different resolutions derived from Dark3, Dark4, and Dark5. These layers are then fused using the path aggregation feature pyramid network (PAFPN) to enhance the information content. The fused feature layers, P3, P4, and P5, are obtained through upsampling, downsampling, and enhanced feature extraction of the three-resolution images and passed on to the three decoupled heads for accurate object prediction in images.

YOLOX excels at object detection. Among the YOLOX-derived models, YOLOX-s has the advantages of a low number of parameters and easy deployment. Therefore, this study chose to improve on YOLOX-s. The improved structure of the network is shown in Figure 1.

2.2. RepVGG Block

RepVGG is a simple and superior convolutional network. It decouples the model's training and inference time structures using structural reparameterization [28], fully balancing speed and accuracy, and is suitable for real-time detection in remote sensing images. The model's overall structure is a stack of more than 3 × 3 convolutional layers, divided into 5 parts. The first layer of each part is a downsample with stride = 2. Each convolutional layer uses Relu [29] as the activation function. During training, the RepVGG block is mainly obtained by adding the 3 deviation vectors to obtain the final deviation, extending the fused 1 × 1 conv and identity with complementary zeros to 3 × 3 conv, and then adding the 3 3 × 3 conv to obtain the final 3 × 3 convolutional layer, as shown in Figure 2. Based

on this study, this paper used the RepVGG block to optimize the backbone of the baseline to improve the model for multi-feature extraction of the object.

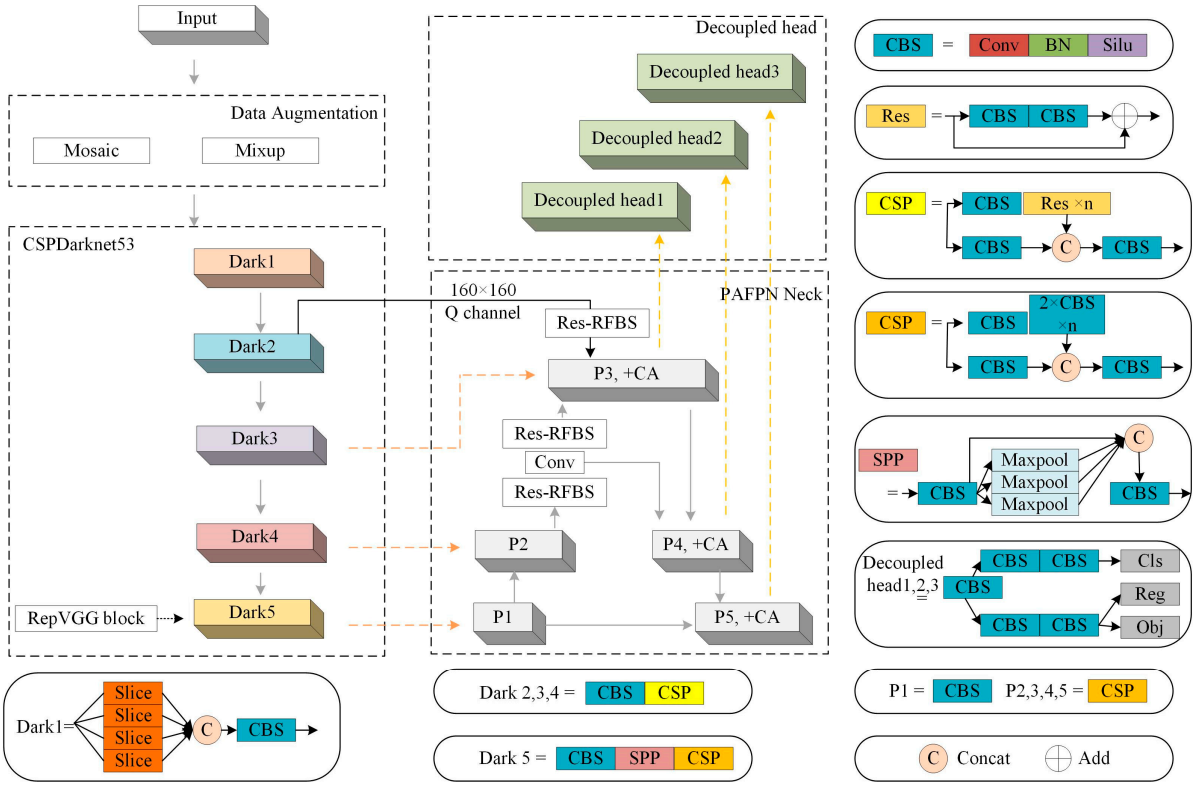

Figure 1. Improved network structure.

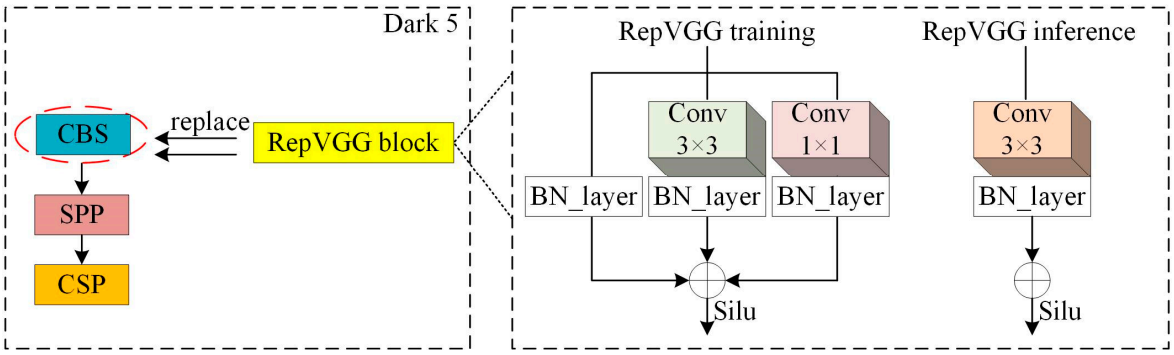

Figure 2. RepVGG block visualization.

In addition, Silu [30] is used instead of the Relu activation function. The Relu activation function is set to zero when the negative gradient is negative, causing some neurons to "necrotize" and affecting network convergence. Silu has the characteristics of no upper bound and lower bound, smooth, and non-monotonic, avoiding negative gradient zeroing to reduce neuronal "necrosis", and better gradient descent than Relu. A comparative plot of the Relu and Silu activation functions is shown in Figure 3. It can be seen that when the function is in a negative gradient, the Relu is set to zero, causing the neural network to fail to learn useful knowledge and neuron "necrosis". In contrast, the Silu function avoids zeroing the negative gradient, retains part of the buffer to reduce neuron "necrosis", and has a better overall gradient descent than Relu. Therefore, this section introduced an improved RepVGG block instead of the partial convolutional layer to optimize the backbone and improve the model's extraction of multi-scale target features.

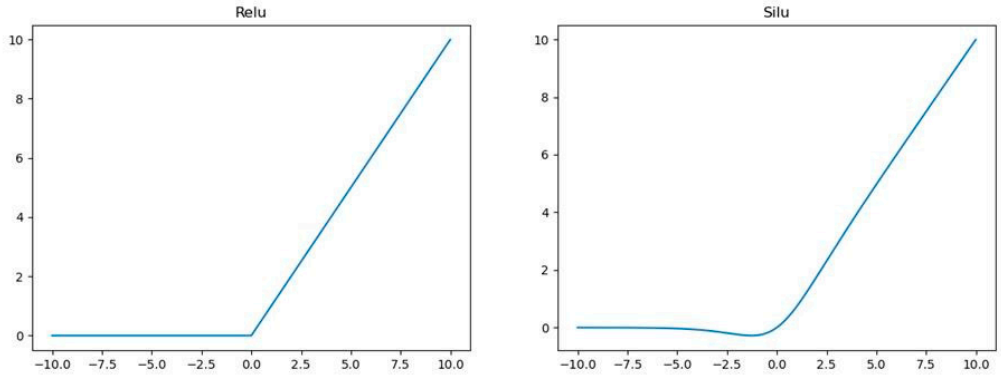

Figure 3. Relu and Silu activation function graphs.

2.3. Improved Feature Detection

The PAFPN pools the feature maps into 80 × 80, 40 × 40, and 20 × 20 resolutions, which are used to detect objects of different sizes. Low resolution detects large objects, medium resolution detects medium-sized objects, and high resolution detects small objects. The resolution of YOLOX for small object detection changes from 640 × 640 to 80 × 80 after convolution, and the lower resolution does not easily capture small object information, resulting in a weaker ability to detect small objects. To avoid the loss of important details in the transmission process of PAFPN, this paper introduced a smaller feature detection channel, 160 × 160, based on the traditional feature detection channel, which will directly input more small object feature information into the feature detection and fuse more network features, which was noted as the Q channel in this paper.

Meanwhile, as the convolution continues, the perceptual field gradually becomes smaller and the detection of multi-scale objects declines progressively. In PAFPN, a new multi-branch convolution module named Res-RFBs was proposed in combination with residual connections, which enhanced the screening of valuable features in this part of the network, as shown in Figure 4.

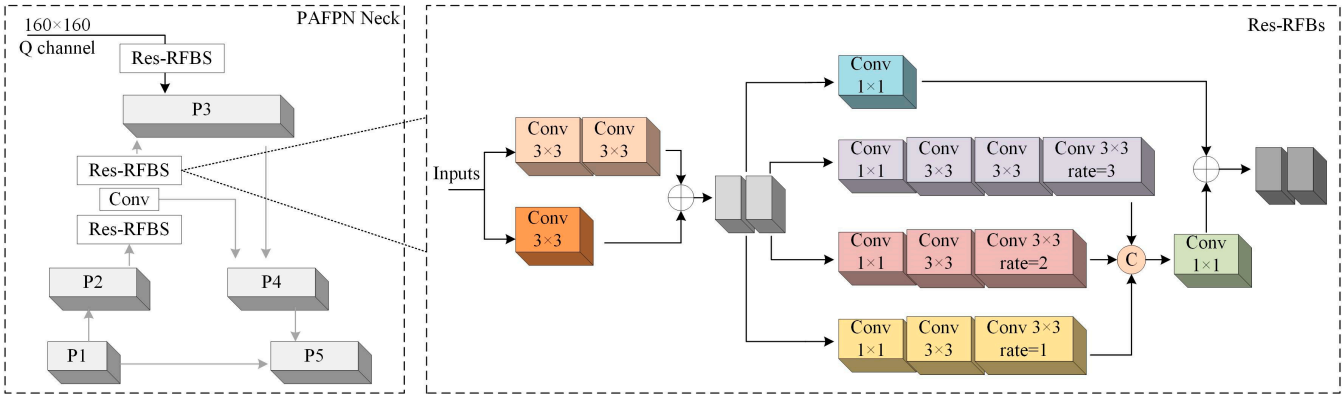

Figure 4. Improved feature detection structure.

ResNet is a deep neural network architecture that effectively solves the gradient disappearance problem in deep neural networks by adding cross-layer connections [31]. ResNet Block is designed to pass the input through two convolutional layers to obtain an output and then add this output to the input, which can further learn deeper features. For this reason, the residual connection was introduced in this paper to improve the network. The improved residual connection was divided into 2 paths: 1 goes through a 3 × 3 and 3 × 3 convolution, and the other is directly shorted with a 3 × 3 convolution. The two are added together and then output. Introducing residual connections into successive convolutions could enhance feature reuse, on the one hand, and avoid the problem of

deep network degradation on the other. The receptive field block (RFB) [32] imitates the receptive field of human vision and enhances the feature expression ability of the network. Adding dilated convolution based on inception increases the receptive field and fuses more information from the image. The RFB structure is mainly composed of three branches, which are interconnected to achieve the fusion of different features. Based on the original RFB, each branch added a layer of 3 × 3 convolutions and replaced the 5 × 5 convolution of the original third branch with a 3 × 3 convolution. The expansion coefficients of rate = 1, rate = 2, and rate = 3 were used to increase the receptive field of multi-scale objects and further improve the detection accuracy of the model. The improved RFB module is shown in Figure 4. This multi-branch convolution module sampled the input features into four mutually independent channels. Within the shortcut channel, the feature map was not additionally processed. In the previous three channels, the convolutions of different numbers and expansion rates were superimposed according to the design to express the feature information of various receptive fields.

2.4. Coordinate Attention Mechanism

The coordinate attention (CA) mechanism is an attention mechanism that enhances the perceptual ability of neural networks by embedding spatial coordinate information into them to better capture the correlation between different locations. CA is divided into two steps: embedding coordinate information and generating coordinate attention, which encodes channel relationships and long-term dependencies using precise location information to fully capture the region of interest and the relationship between channels [33]. The mechanism aggregates the input feature maps along the X and Y directions through two global average pooling operations and then encodes the information through dimension transformation. Finally, the spatial information and channel features are weighted and fused, considering the channel and location information. Therefore, CA can better focus on the object of interest, as shown in Figure 5.

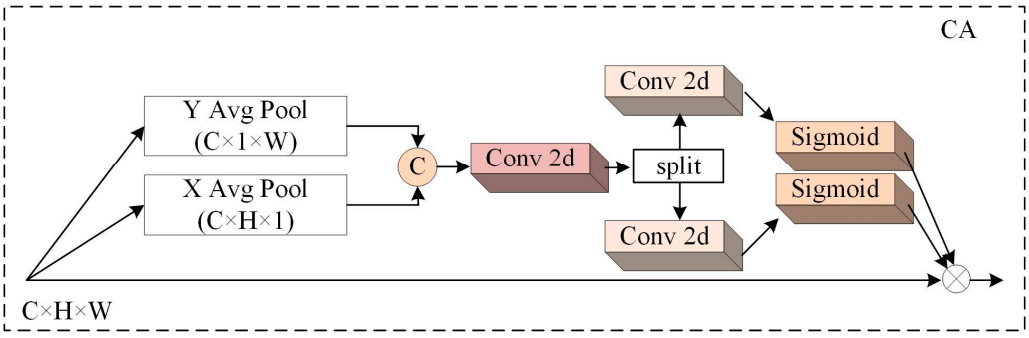

Figure 5. CA module.

In the actual recognition process of remote sensing images, due to the complexity of the image scene, the existing network often cannot eliminate redundant interference information, and the object to be detected is small and densely distributed. In the detection process, the convolutional network needs to process the cells divided by each image. Additionally, many calculations cannot perceive the object well, resulting in missed and false detection problems. Therefore, based on the improved feature extraction network in the previous section, this paper introduced the CA module before the decoupled head. The features could cover more parts of the object to be identified, reduce the interference of background information, make the network focus on essential details of interest, and enhance the expressiveness to improve detection accuracy.

2.5. Loss Function Improvement

The loss function of YOLOX consists of IoU loss (L_{IoU}), category loss (L_{Cls}), and confidence loss (L_{Obj}), which can be expressed as $L_{Loss} = L_{IoU} + L_{Cls} + L_{Obj}$.

Among them, *IoU* refers to the intersection and union ratio, a commonly used indicator in object detection, reflecting the detection effect of the predicted and real detection boxes. The calculation formula is shown in (1):

$$IoU = \frac{A \cap B}{A \cup B} \tag{1}$$

In Formula (1), A represents the prediction box and B represents the real box. IoU is the concept of ratio. In the calculation process using the IoU function, if the prediction box and the real box do not intersect, the degree of coincidence between the two cannot be reflected. In the process of prediction region regression, when the IoU value between the prediction box after the regression and the real box is zero, the problem of target miss rate is caused by the failure of the prediction region to return. In contrast, generalized intersection over union (GIoU) [34] satisfies the basic requirements of the loss function by being concerned not only with the overlapping regions but also with other non-overlapping regions, which can better reflect the coincidence degree between the two objects and accelerate the convergence rate of the model. GIoU first finds the minimum shape A_c to surround the prediction box and the real box. In order to compare two specific geometric shape types, A_c can come from the same type. Finally, the ratio between the area occupied by A_c is calculated and then divided by the total area occupied by A_c, as shown in Formula (2). Therefore, this paper replaced the IoU loss function with the *GIoU* loss function.

$$GIoU = IoU - \frac{|A_c - U|}{|A_c|} \tag{2}$$

In Equation (2), A_c represents the minimum closure area of the prediction box and the real box, and U represents $A \cup B$. For the GIoU loss function, L_{GIoU} can be expressed as:

$$L_{GIoU} = 1 - GIoU = 1 - IoU + \frac{|A_c - U|}{|A_c|} \tag{3}$$

The category loss contains the category information of the remote sensing images, and the confidence loss includes the background information of the image. The category loss and confidence loss are calculated using the bcewithlog_loss function to speed up the model convergence. The loss function is finally shown as (4):

$$L_{Loss} = L_{GIoU} + L_{Cls} + L_{Obj} \tag{4}$$

3. Experiment

3.1. Experimental Environment

The experimental operating system was Windows 10, the GPU was NVIDIA GeForce RTX 3060, and the memory was 12 G. The deep learning framework was Pytorch 1.7.1 and Cuda 11.6. The training had two stages: the freezing stage and the thawing stage. The SGD optimizer was used to adjust the learning rate using the cosine annealing strategy while using pre-training weights.

3.2. Data Set

This experiment uses the NWPU VHR-10 [35] and DIOR [36] datasets.

The NWPU VHR-10 is a high-resolution remote sensing image dataset with spatial resolution ranging from 0.5 m to 2 m. It contains 10 categories of objects and 800 images, with a total number of 3651 target instances. The short names, C1–C10, for our experiment categories were: Tennis court, Harbor, Ground track field, Basketball court, Airplane,

Storage tank, Baseball field, Ship, Vehicle, and Bridge. Figure 6 shows the remote sensing images and objects of the NWPU VHR-10 dataset, where the boxed positions are the objects.

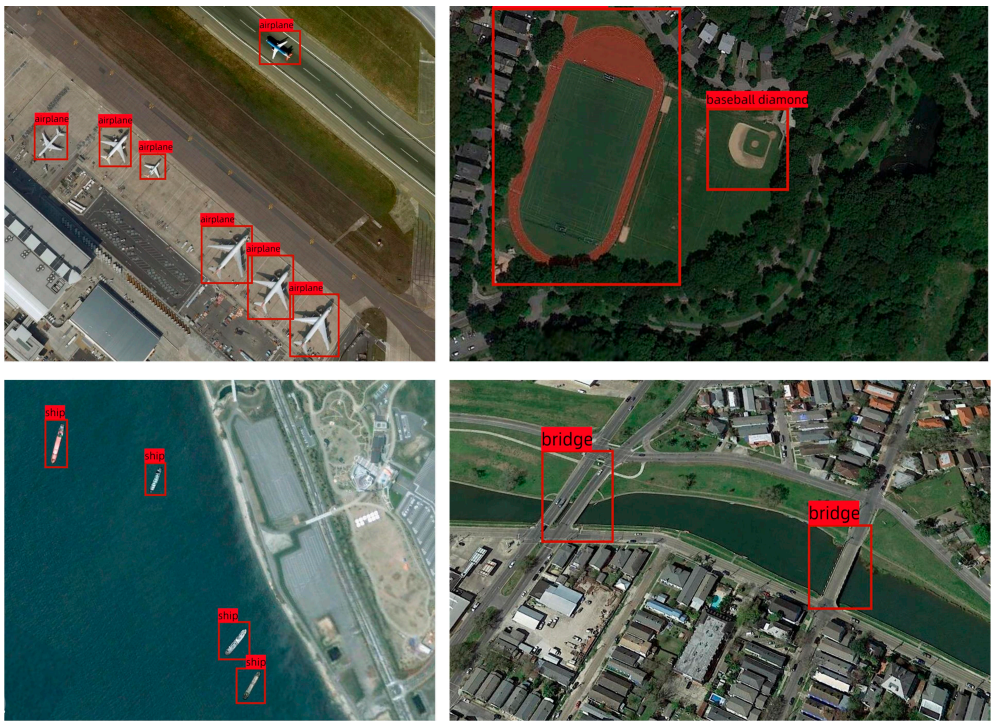

Figure 6. Remote sensing images and objects of NWPU VHR-10 dataset.

DIOR is a large-scale benchmark data set for object detection in optical remote sensing images. It is divided into 20 object classes, including 23,463 remote sensing images and 190,288 instances. It has high similarity and diversity in different imaging conditions, weather, seasons, and image quality. The short names C1–C20 for categories in our experiment were defined as Airplane, Airport, Baseball field, Basketball court, Bridge, Chimney, Dam, Expressway service area, Expressway toll station, Golf field, Ground track field, Harbor, Overpass, Ship, Stadium, Storage tank, Tennis court, Train station, Vehicle, and Windmill. Figure 7 shows the remote sensing images and objects of the DIOR dataset, where the boxed positions are the objects.

3.3. Evaluation Metrics

To accurately evaluate the effect of the proposed method on remote sensing image detection, this study selected the mean average precision (*mAP*), precision rate (*P*), recall rate (*R*), and frame per second (*FPS*) as evaluation indicators. The calculation formula is shown in Formulas (5)–(7). When the accuracy and recall rates are compared separately, ambiguity will occur. Therefore, the experiment used the *mAP* to evaluate the model's effectiveness by comprehensively considering the precision and recall rates. *FPS* refers to the number of frames detected per second to measure the real-time performance of the model.

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{7}$$

In Formulas (5) and (6), *TP* represents the number of correct predictions, *FP* represents the number of false predictions, and *FN* represents the number of missing predictions; in Formula (7), *k* is the category, and the calculation formula of average precision (*AP*) can be expressed as:

$$AP = \int_0^1 \mathrm{p(r)dr} \tag{8}$$

Figure 7. Remote sensing images and objects of DIOR dataset.

3.4. Ablation Experiment

In the improvement strategy for the backbone, if it is a direct addition of RepVGG modules, it may not necessarily have the desired effect. To explore the effectiveness of the RepVGG addition position, this paper added the RepVGG block to the backbone to determine the RepVGG block addition position. The different addition positions are shown in Figure 8.

Table 1 shows that the best mAP of remote sensing image detection was achieved using the RepVGG block instead of the fourth convolutional layer in the backbone. Compared with the experiments conducted by Relu, the Silu function was more stimulating to the feature extraction performance of the model and avoided some neuron necrosis. The different results with different addition positions are because the number of composite convolutional layers and the amount of information reorganization are different, thus bringing different gains to the model. After an experimental demonstration, the RepVGG block was used instead of the fourth convolutional layer in the backbone.

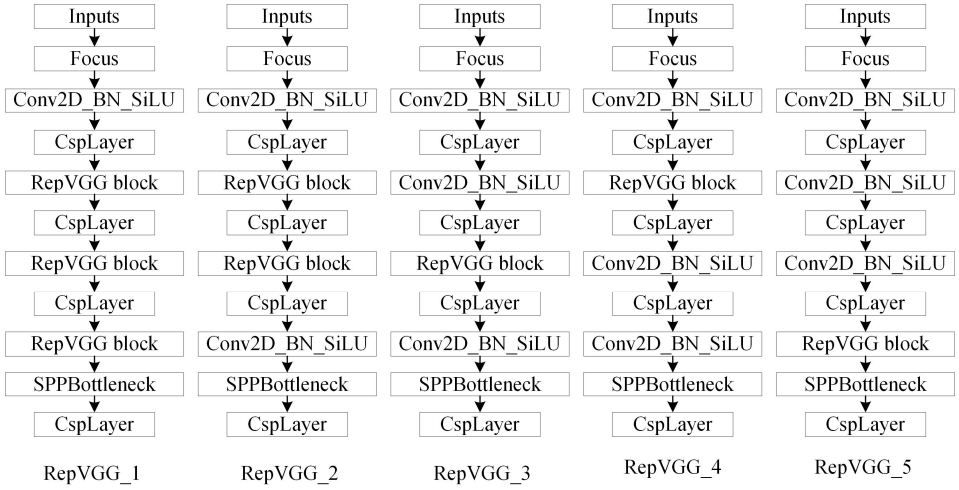

Figure 8. The RepVGG block of backbone is added at different locations.

Table 1. Experimental results of RepVGG at different locations. P stands for precision rate, R stands for recall rate, FPS stands for frame per second, and mAP stands for mean average precision.

Location	P/(%)	R/(%)	mAP/(%)	FPS/(f/s)
RepVGG_1(Relu)	91.55	86.40	90.14	45.64
RepVGG_1	91.94	87.20	92.19	46.89
RepVGG_2	92.91	88.07	93.07	46.37
RepVGG_3	91.03	90.58	93.34	48.39
RepVGG_4	91.17	91.93	92.99	47.45
RepVGG_5	90.67	89.78	93.53	48.50

To explore the effectiveness of the improved module, ablation experiments were conducted on the YOLOX-s-based NWPU VHR-10 dataset for the RepVGG block, Q, multi-branch convolution, CA attention, and GIoU loss function, respectively. The experimental projects were carried out by sequentially adding each proposed module; the results are shown in Table 2.

Table 2. Ablation experiment of the improved module. P stands for precision rate, R stands for recall rate, FPS stands for frame per second, and mAP stands for mean average precision.

RepVGG	Q+ Res-RFBs	CA	GIoU	P/(%)	R/(%)	mAP/(%)	FPS/(f/s)
-	-	-	-	89.68	90.37	92.23	48.02
√	-	-	-	90.67	89.78	93.53	48.50
-	√	-	-	93.45	91.66	94.98	33.61
-	-	√	-	90.51	93.28	94.14	41.26
-	-	-	√	93.12	90.65	93.56	35.61
√	√	√	√	94.09	94.94	96.63	30.09

As shown in Table 2, the mAP of using the RepVGG block was increased by 1.3%, and the FPS was increased by 1% compared with the base network, which indicated that the system performance was further improved. As seen in Table 1, the experiments obtained better results using the Silu activation function than Relu. The RepVGG block made extensive use of 3 × 3 convolution. It used a multi-branch network for training by structural reparameterization and fused the multi-branch into a single branch for prediction, facilitating network acceleration. Figure 9a showed the original input image, and a new layer of 160 × 160 input channels was introduced in the PAFPN, which could express more small object information compared with the original network, and more object information was obtained compared with that shown in Figure 9b,c. After adding Res-RFBs on the basis of introducing Q (160 × 160) input channels, the receptive field could be further expanded to enhance the detailed expression of the model effectively. As seen in Figure 9, the object feature information in the image was not effectively detected in the baseline feature heat map, resulting in a scattered region of interest for the network and object features that could not be extracted effectively. In the multi-scale feature heat map, it could be clearly observed that the features of different objects were enhanced after adding multi-scale convolution especially for ships and dense storage tanks. Information was enhanced in the image of the object, thus proving the effectiveness of multi-branch convolution for improving the features of the object. The results are shown in Figure 9d,e, which effectively enhanced the detection ability of multi-scale objects with dense distribution, and mAP was increased by 2.75%. The reason for the increase of 1.91% using the CA attention mechanism over the baseline was that the CA module considers both channel and direction-related location information to further strengthen the neural network's ability to perceive remote sensing objects and focus more on the object. The improved loss function increased the mAP by 1.33% and improved the model's performance. The reason was that the increased penalty measure of the GIoU function facilitates the network in making accurate judgments on remote sensing objects and compensates for the non-overlapping regions of the detection objects in the IoU loss function, which effectively reduces the target miss rate. After adding the RepVGG block, Q+ Res-RFBS, CA, and GIoU loss function, the detection accuracy reached 95.50%, which was 3.27% better than the baseline. The final detection accuracy of 96.63% was obtained after multiple training iterations. Overall, the improved modules enhanced detection accuracy, and the use of the above improvement strategies eventually brought a gain of 4.4 percentage points to the model, which proved the effectiveness of the improvement strategies.

3.5. Comparison with Other Algorithms

To further verify the effectiveness and rationality of the improved YOLOX for object detection in remote sensing images, this experiment used YOLOX, MFANet, and mainstream algorithms to train and test the detection accuracy of each algorithm in the NWPU VHR-10 and DIOR datasets. The experimental results are shown in Tables 3 and 4. In the NWPU VHR-10 dataset in Table 3, the methods of Faster RCNN, YOLOv4-tiny, YOLOv5, and YOLOX-s, Laban's [20], SCRDet [37], Fan's [38], Zhang's [39], and Xue's [40] networks were selected for comparison. The results showed that, compared with other models, the MFANet proposed in this paper had the best mAP, 96.63%, which was 17.15%, 7.49%, 4.88%, 3.23%, 1.04%, and 0.93% higher than Faster RCNN, YOLOv5, SCRDet [37], Fan's [38], Zhang's [39], and Xue's [40], respectively, and had better detection performance.

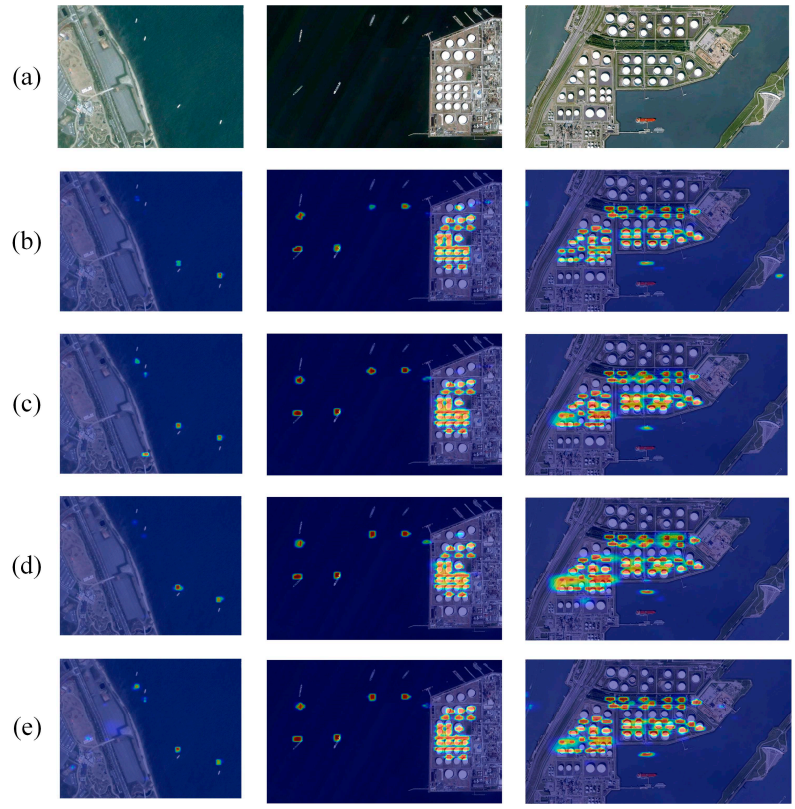

Figure 9. Visualization results of improved feature detection: (**a**) input image; (**b**) visualization results for YOLOX; (**c**) visualization results of adding 160 × 160 channels for YOLOX; (**d**) visualization results of adding Res-RFBs for YOLOX; (**e**) visualization results of adding 160 × 160 channels and Res-RFBs for YOLOX.

Table 3. Experimental results in the NWPU VHR-10 dataset. FPS stands for frame per second and mAP stands for mean average precision.

Method	mAP/(%)	FPS/(f/s)
Faster RCNN	79.48	10.59
Laban's [20]	78.00	-
YOLOv4-tiny	84.14	81.36
YOLOv5	89.14	54.91
SCRDet [37]	91.75	-
YOLOX-s	92.23	48.02
Fan's [38]	93.40	-
Zhang's [39]	95.59	30.07
Xue's [40]	95.70	-
MFANet	96.63	30.09

Table 4. Experimental results in the DIOR dataset. FPS stands for frame per second and mAP stands for mean average precision.

Method	mAP/(%)	FPS/(f/s)
Faster RCNN	57.35	10.42
AOPG [41]	64.41	-
LO-Det [42]	65.85	60.03
Li's [43]	66.71	-
YOLOv4-tiny	66.77	56.68
ASSD [44]	71.80	21.00
Yao's [45]	75.80	-
SCRDet++ [46]	77.80	-
YOLOv5	80.96	51.41
SPB-YOLO [23]	81.10	-
YOLOX-s	82.23	47.99
Zhou's [47]	84.30	-
YOLOX [24]	85.70	-
Ye's [48]	86.55	-
MFANet	87.88	29.45

It can be seen from Figure 10 that in the experiment of the NWPU VHR-10 dataset, when there was a complex scene to be detected, the object distribution was dense, or the object had a low resolution in the image, the effect of YOLOX-s detection was not good, and it was prone to problems such as missed detection and false detection. In Figure 10b, due to the interference of more background information in the image, the baseline network missed and made false detections of small objects, such as ships and vehicles. Compared with Figure 10b,c, the improved network improves the detection efficiency of objects, and the problems in Figure 10b are basically solved. The detection effect of this algorithm is significantly better than that of YOLOX. To compare the model detection effects in one step, this paper selected the AP and mAP results of YOLOv4-tiny, YOLOv5, YOLOX-s, Fan's, Zhang's, Xue's, and MFANet, and the results are shown in Table 5. For the object bridge, although the detection accuracy of MFANet was a little bit lower than the methods of Xue's and Zhang's results, it was higher than that of YOLOv4-tiny, YOLOv5, and YOLOX-s. It can also be seen that MFANet shows a significant improvement in detecting ships and vehicles compared to YOLOv4-tiny, YOLOv5, YOLOX-s, Fan's, Zhang's, and Xue's. It was 1% higher than Xue's when testing ships and 4% higher than Fan's when testing vehicles. In addition, the MFANet achieved better detection results on objects such as the Tennis court, Harbor, and Basketball court. When there is serious background interference in the target, such as the similarity in appearance between the refuse collection point and the parked vehicles alongside the road, the base detection network will miss the detection. In contrast, the improved network effectively improved the inspection accuracy of vehicles and avoided object misdetection.

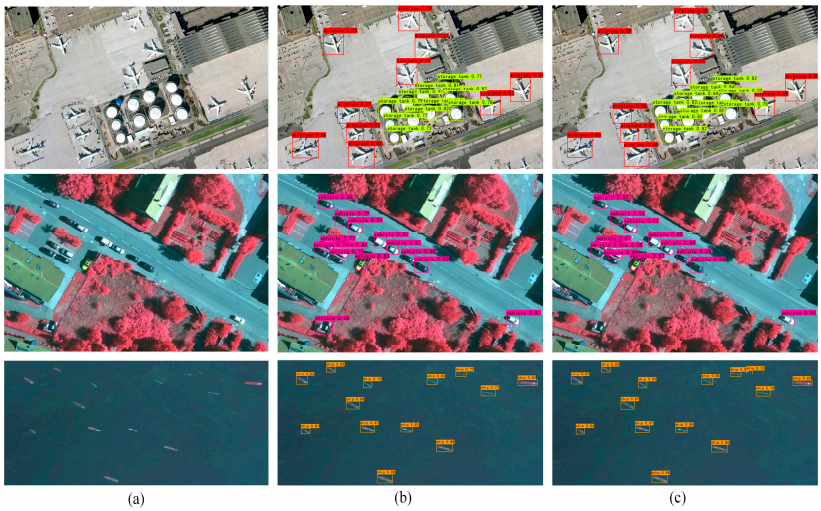

Figure 10. Detection results for the NWPU VHR-10 dataset: (**a**) input image; (**b**) for YOLOX-s; (**c**) for MFANet.

Table 5. AP and mAP of the different algorithms acting on multiple object categroies from NWPU VHR-10 data. AP stands for average precision and mAP stands for mean average precision.

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	mAP
YOLOv4-tiny	0.91	0.98	0.99	0.66	1.00	0.71	0.99	0.76	0.82	0.59	0.84
YOLOv5	1.00	1.00	0.98	0.83	1.00	0.99	0.98	0.91	0.90	0.33	0.89
YOLOX-s	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.94	0.89	0.40	0.92
Fan's	1.00	1.00	0.91	0.91	1.00	0.90	0.94	0.91	0.90	0.91	0.93
Zhang's	1.00	1.00	1.00	1.00	1.00	0.90	1.00	0.90	0.87	0.90	0.96
Xue's	0.90	0.96	1.00	1.00	1.00	0.88	1.00	0.96	0.89	0.99	0.96
MFANet	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.97	0.94	0.77	0.96

The DIOR dataset has a significant difference in spatial resolution and cross-object scale, and its high inter-class similarity and class diversity increase the difficulty of detection. We can see that the YOLOv4-tiny part of the detection object was higher than YOLOX and MFANet, but MFANet still led the overall multi-object detection effect. YOLOX-s missed and mis-checked at the background of complex scenes that contained diverse categories of feature elements (Figure 11c). Compared with Figure 11d, it can be seen that the detection effect of the improved algorithm was significantly improved, and multi-scale objects were effectively detected. The improved model has strong robustness.

To compare the model detection effects in one step, this paper selected the AP and mAP results of ASSD, Yao's, SCRDet++, YOLOv5, YOLOX-s, Zhou's, and MFANet, which can be seen in Table 6. For the object Golf field, although the MFANet detection accuracy is not as good as Zhou's, it was higher than YOLOX-s, YOLOv5, SCRDet++, Yao's, and ASSD, by 3%, 15%, 1%, 6%, and 5%, respectively. It can be seen that in the detection of bridges and vehicles, compared with YOLOX-s, the MFANet had a significant improvement. In addition, we found that when the road and harbor samples were similar, the baseline network missed detection due to insufficient resolution and failed to identify the port object effectively. The optimized network with enhanced multi-scale feature extraction could effectively detect most objects and complete the object detection task.

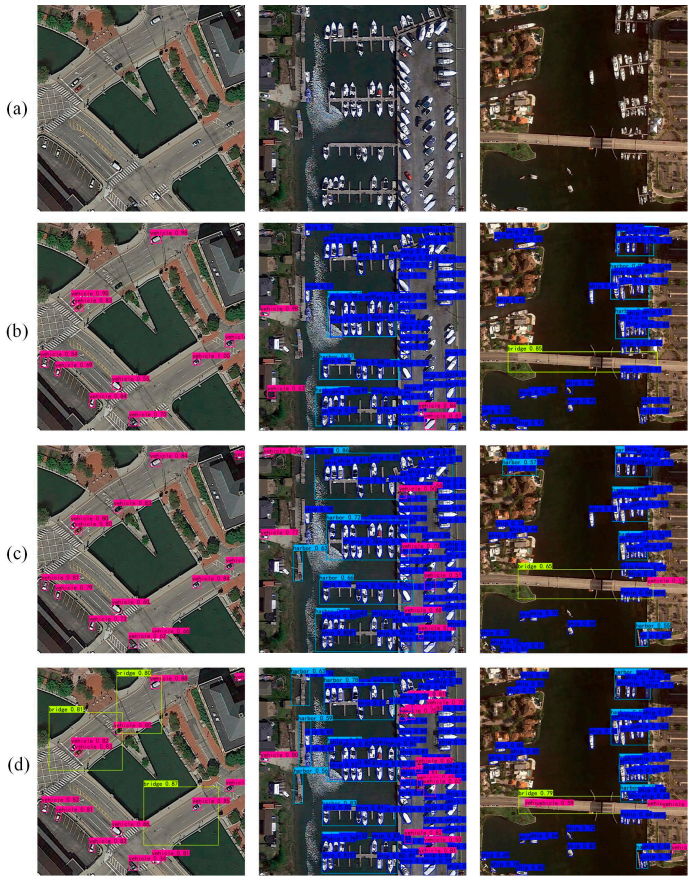

Figure 11. Detection results for the DIOR dataset: (**a**) input image; (**b**) for YOLOv4-tiny; (**c**) for YOLOX-s; (**d**) for MFANet.

Table 6. AP and mAP of the different algorithms acting on multiple object categroies from DIOR data. AP stands for average precision and mAP stands for mean average precision.

Method	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
ASSD	0.86	0.82	0.76	0.90	0.41	0.78	0.65	0.67	0.62	0.81
Yao's	0.91	0.75	0.93	0.83	0.47	0.92	0.63	0.68	0.61	0.80
SCRDet++	0.81	0.88	0.80	0.90	0.58	0.81	0.75	0.90	0.83	0.85
YOLOv5	0.96	0.86	0.97	0.86	0.48	0.86	0.75	0.86	0.77	0.71
YOLOX-s	0.96	0.88	0.96	0.83	0.48	0.78	0.78	0.94	0.79	0.83
Zhou's	0.98	0.90	0.95	0.93	0.62	0.91	0.68	0.96	0.86	0.87
MFANet	0.97	0.93	0.97	0.86	0.59	0.88	0.87	0.97	0.90	0.86
mAP	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
0.71	0.79	0.62	0.58	0.85	0.77	0.65	0.88	0.62	0.45	0.76
0.76	0.83	0.57	0.66	0.80	0.93	0.81	0.89	0.63	0.73	0.78
0.78	0.84	0.63	0.67	0.73	0.79	0.70	0.90	0.71	0.59	0.90
0.81	0.92	0.67	0.71	0.95	0.89	0.86	0.96	0.63	0.60	0.93
0.82	0.90	0.69	0.71	0.96	0.96	0.87	0.95	0.65	0.62	0.92
0.84	0.91	0.63	0.73	0.96	0.92	0.90	0.96	0.57	0.71	0.94
0.87	0.93	0.75	0.79	0.97	0.97	0.92	0.97	0.79	0.73	0.94

4. Discussion

The ablation experimental results show that the improved YOLOX in this paper can improve the model's multi-scale object recognition rate, and the mAP was improved by 4.4% compared to the YOLOX. Figure 9 shows that introducing a new layer of 160 × 160 input channels in the PAFPN can express more information about small objects than the original network. The addition of Res-RFBs was based on the introduction of detail-enhanced

channels, which enhanced feature multiplexing and expanded the perceptual field, thus improving the detection accuracy of multi-scale objects by 2.75% compared to the mAP of the baseline. The results in Tables 1 and 2 show that the RepVGG block uses structural reparameterization to improve the extraction of multi-scale object features, and mAP was increased by 1.3%. The results in Table 2 and Figure 10 show that CA enhances the ability of the neural network to perceive remote sensing objects, with a 1.91% improvement in mAP. The results in Table 2 and Figures 10 and 11 show that the GIoU loss function reduces the target miss rate. The comparison experimental results show that the proposed method had a higher accuracy rate when compared with other mainstream object detection algorithms. On the DIOR dataset, mainstream algorithms such as Faster RCNN [12], YOLOv5, and YOLOX are used in this paper, while models such as AOPG [41], Li's [43], Yao's [45], SCRDet++ [46], Zhou's [47], and Ye's [48] are selected for comparison. The results in Table 4 show that the improved YOLOX model proposed in this paper had a better mAP than other models (30.53%, 6.92%, 6.78%, 3.58%, 2.18%, and 1.33% higher than Faster RCNN, YOLOv5, SPB-YOLO [23], Zhou's [47], YOLOX [24], and Ye's [48]), achieving advanced detection and classification performance. The NWPU VHR-10 dataset shows that the MFANet obtained a lower detection speed than YOLOv4-tiny, YOLOv5, etc., but a higher detection speed than Faster RCNN, Zhang's, etc. In addition, on the DIOR dataset, compared with LO-Det [42] and YOLOv5, although the detection speed was lower, the detection accuracy of the improved network in this paper was much higher than theirs: 22.03% and 6.92% higher than LO-Det and YOLOv5, respectively. Compared with ASSD [44], MFANet is leading in detection accuracy and speed. Comparing with Table 6, we find that for large objects, such as Airport, Expressway service areas, etc., with improved algorithm detection, the AP improved by 5% and 3%, respectively; for medium-sized objects, such as Harbor, Chimney, etc., with improved algorithm detection, the AP improved by 6% and 10%, respectively; for small objects, such as Bridge and Storage tank, the AP was enhanced by 11% and 5%, respectively, after improved algorithm detection. Overall, the experimental results verify the effectiveness of the improved network in detecting multi-scale objects.

At the same time, we find that in the area of small object distribution shown in Figure 11, some images of the small objects are blurred and carry too little feature information, resulting in the detector failing to effectively detect them, which affects the detection results. In Table 6, we can see that the AP for vehicles was lower than boats and airplanes, which is probably because the less contextually available feature information of small objects. In addition, the FPS of the improved algorithm proposed in this paper reached 30.09 and 29.45 on the NWPU VHR-10 and DIOR datasets, respectively, which were much higher than Faster RCNN. However, the FPS decreased compared to the original network. The reason for this is that the improved PAFPN makes the network structure complex, introducing many parameters and increasing the computational time consumption, thus slowing down the detector. A linear discriminant can cluster objects [49], and eliminating redundancy constraints can improve detection speed [50], providing a method for object detection in remote sensing images. Therefore, to improve some shortcomings of the algorithm in this paper, the following aspects can be considered. Firstly, using discriminant analysis and migration learning to improve the generalization of the network. Secondly, reducing the number of redundant parameters in the model while maintaining high efficiency and using deeper contextual feature information to achieve high-quality small object detection.

5. Conclusions

Aiming at the complex problem of multi-scale detection of remote sensing images, this research proposed the MFANet based on YOLOX. The MFANet used RepVGG to build a new backbone, and the detection accuracy of the backbone after reparameterization increased from 92.23% to 93.53%, which proved the effectiveness of its prediction; at the same time, the Silu activation function was selected and the detection accuracy increased by 2.05%. The choice of detail channel and multi-branch convolution should be combined with different datasets to determine the best object extraction performance. In this paper, the Q

channel and three multi-branch convolutions were selected in PAFPN to achieve the best detection effect. In addition, this paper proved the role of the CA module in improving the image detection network by adding the CA module to the improved network, effectively reducing background interference and increasing the detection accuracy by 1.91%. Finally, the GIoU function was used to optimize the loss and the detection accuracy was increased by 1.33%, effectively avoiding missed object detection. The experiment was carried out on the NWPU VHR-10 and DIOR datasets. Compared with current object detection algorithms, the MFANet achieved higher detection accuracy. MFANet demonstrated a high mean average precision of 98.78% for 9 classes of objects in the NWPU VHR-10 10-class detection dataset and 94.91% for 11 classes of objects in the DIOR 20-class detection dataset. The overall performance of mAP was 96.63% and 87.88% for the NWPU VHR-10 and DIOR datasets, respectively. In summary, the combination of multi-branch feature fusion and an attention model is a superior approach to improving the accuracy of multi-scale object detection in remote sensing images. In the future, the feature extraction mechanism in MFANet can be further deepened and optimized, especially when many object categories are contained in remote sensing images.

Author Contributions: Y.C. and W.W. designed the study. Y.C. and W.W. conducted the analysis and wrote the manuscript. W.Z., L.Y., J.W., H.N., T.G., J.H., Y.G. and N.N.T. offered advice to improve the manuscript. Y.C., W.W. and W.Z. interpreted the results and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Natural Science Foundation of China under Grant no. 41975183 and Grant no. 41875184.

Data Availability Statement: The data (NWPU VHR-10 dataset and the DIOR dataset) used to support the results of this study are available from the respective authors upon request and can be found in the references [35,36].

Acknowledgments: We thank the anonymous reviewers for their comments and suggestions that improved this paper.

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

References

1. Li, W. Detection of Ship in Optical Remote Sensing Image of Median-Low Resolution. Master's Thesis, National University of Defense Technology, Changsha, China, November 2008.
2. Wang, Y.; Ma, L.; Tian, Y. State-of-the-art of Ship Detection and Recognition in Optical Remotely Sensed Imagery. *Acta Autom. Sin.* **2011**, *37*, 1029–1039.
3. Rajendran, G.B.; Kumarasamy, U.M.; Zarro, C.; Divakarachari, P.B.; Ullo, S.L. Land-Use and Land-Cover Classification Using a Human Group-Based Particle Swarm Optimization Algorithm with an LSTM Classifier on Hybrid Pre-Processing Remote-Sensing Images. *Remote Sens.* **2020**, *12*, 4135. [CrossRef]
4. Zhang, W.; Zhang, B.; Zhu, W.; Tang, X.; Li, F.; Liu, X.; Yu, Q. Comprehensive assessment of MODIS-derived near-surface air temperature using wide elevation-spanned measurements in China. *Sci. Total Environ.* **2021**, *800*, 149535. [CrossRef] [PubMed]
5. Nie, G.; Huang, H. A survey of object detection in optical remote sensing images. *Acta Autom. Sin.* **2021**, *47*, 1749–1768.
6. Parameshachari, B.; Gurumoorthy, S.; Frnda, J.; Nelson, S.C.; Balmuri, K.R. Cognitive linear discriminant regression computing technique for HTTP video services in SDN networks. *Soft Comput.* **2022**, *26*, 621–633. [CrossRef]
7. Wang, R.; Wu, X.; Kittler, J. SymNet: A simple symmetric positive definite manifold deep learning method for image set classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 2208–2222. [CrossRef] [PubMed]
8. Gao, X.; Niu, S.; Wei, D.; Liu, X.; Wang, T.; Zhu, F.; Dong, J.; Sun, Q. Joint Metric Learning-Based Class-Specific Representation for Image Set Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–15. [CrossRef] [PubMed]
9. Parameshachari, B.; Panduranga, H. Medical image encryption using SCAN technique and chaotic tent map system. In *Recent Advances in Artificial Intelligence and Data Engineering*; Springer: Singapore, 2022; pp. 181–193.
10. Zhou, F.; Jin, L.; Dong, J. Review of Convolutional Neural Network. *Chin. J. Comput.* **2017**, *40*, 1229–1251.
11. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]

13. Zhu, M.; Xu, Y.; Ma, S.; Li, S.; Ma, H.; Han, Y. Effective airplane detection in remote sensing images based on multilayer feature fusion and improved nonmaximal suppression algorithm. *Remote Sens.* **2019**, *11*, 1062. [CrossRef]
14. Shivappriya, S.N.; Priyadarsini, M.J.P.; Stateczny, A.; Puttamadappa, C.; Parameshachari, B.D. Cascade Object Detection and Remote Sensing Object Detection Method Based on Trainable Activation Function. *Remote Sens.* **2021**, *13*, 200. [CrossRef]
15. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
16. 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. 6517–6525.
17. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
18. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
19. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
20. Laban, N.; Abdellatif, B.; Ebeid, H.M.; Shedeed, H.A.; Tolba, M.F. Convolutional Neural Network with Dilated Anchors for Object Detection in Very High Resolution Satellite Images. In Proceedings of the International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 17 December 2019; pp. 34–39.
21. Hong, Z.; Yang, T.; Tong, X.; Zhang, Y.; Jiang, S.; Zhou, R.; Han, Y.; Wang, J.; Yang, S.; Liu, S. Multi-scale ship detection from SAR and optical imagery via a more accurate YOLOv3. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6083–6101. [CrossRef]
22. Zhou, H.; Guo, W. Improved YOLOv5 Network in Application of Remote Sensing Image Object Detection. *Remote Sens. Inf.* **2022**, *37*, 23–30.
23. Wang, X.; Li, W.; Guo, W.; Cao, K. SPB-YOLO: An Efficient Real-Time Detector For Unmanned Aerial Vehicle Images. In Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Jeju Island, Republic of Korea, 13–16 April 2021; pp. 099–104.
24. Han, X.; Li, F. Remote Sensing Small Object Detection Based on Cross-Layer Attention Enhancement. *Laser Optoelectron. Prog.* 2022, pp. 1–19. Available online: https://kns.cnki.net/kcms/detail/31.1690.TN.20220722.2132.050.html (accessed on 17 February 2023).
25. Wu, Q.; Zhang, B.; Xu, C.; Zhang, H.; Wang, C. Dense Oil Tank Detection and Classification via YOLOX-TR Network in Large-Scale SAR Images. *Remote Sens.* **2022**, *14*, 3246. [CrossRef]
26. Yang, L.; Yuan, G.; Zhou, H.; Liu, H.; Chen, J.; Wu, H. RS-YOLOX: A High-Precision Detector for Object Detection in Satellite Remote Sensing Images. *Appl. Sci.* **2022**, *12*, 8707. [CrossRef]
27. Guo, Q.; Yuan, C. Leveraging Spatial-Semantic Information in Object Detection and Segmentation. *Ruan Jian Xue Bao/J. Softw.* 2022, pp. 1–13. Available online: http://www.jos.org.cn/jos/article/abstract/6509 (accessed on 17 February 2023). (In Chinese).
28. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-style ConvNets Great Again. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13728–13737.
29. Shang, W.; Sohn, K.; Almeida, D.; Lee, H. Understanding and improving convolutional neural networks via concatenated rectified linear units. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 2217–2225.
30. Ramachandran, P.; Zoph, B.; Le, Q. Swish: A Self-Gated Activation Function. *arXiv* **2017**, arXiv:1710.05941.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
32. Liu, S.; Huang, D.; Wang, Y. 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. Available online: https://doi.org/10.48550/arXiv.1711.07767 (accessed on 28 August 2022).
33. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
34. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
35. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
36. 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]
37. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8231–8240.
38. Fan, X.; Yan, W.; Shi, P.; Zhang, X. Remote sensing image target detection based on a multi-scale deep feature fusion network. *Natl. Remote Sens. Bull.* **2022**, *26*, 2292–2303.
39. Zhang, J.; Wu, X.; Zhao, X.; Zhuo, L.; Zhang, J. Scene Constrained Object Detection Method in High-Resolution Remote Sensing Images by Relation-Aware Global Attention. *J. Electron. Inf. Technol.* **2022**, *44*, 2924–2931.
40. Xue, J.; Zhu, J.; Zhang, J.; Li, X.; Dou, S.; Mi, L.; Li, Z.; Yuan, X.; Li, C. Object Detection in Optical Remote Sensing Images Based on FFC-SSD Model. *Acta Opt. Sin.* **2022**, *42*, 138–148.

41. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-Free Oriented Proposal Generator for Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625411. [CrossRef]
42. Huang, Z.; Li, W.; Xia, X.; Wang, H.; Jie, F.; Tao, R. LO-Det: Lightweight Oriented Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
43. Li, W.; Chen, Y.; Hu, K.; Zhu, J. Oriented reppoints for aerial object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 1829–1838.
44. Xu, T.; Sun, X.; Diao, W.; Zhao, L.; Fu, K.; Wang, K. ASSD: Feature Aligned Single-Shot Detection for Multiscale Objects in Aerial Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607117. [CrossRef]
45. Yao, Y.; Cheng, G.; Xie, X.; Han, J. Optical remote sensing image object detection based on multi-resolution feature fusion. *Natl. Remote Sens. Bull.* **2021**, *25*, 1124–1137.
46. Yang, X.; Yan, J.; Liao, W.; Yang, X.; Tang, J.; He, T. SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2384–2399. [CrossRef] [PubMed]
47. Zhou, L.; Zheng, C.; Yan, H.; Zuo, X.; Liu, Y.; Qiao, B.; Yang, Y. RepDarkNet: A Multi-Branched Detector for Small-Target Detection in Remote Sensing Images. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 158. [CrossRef]
48. Ye, Y.; Ren, Y.; Gao, X.; Wang, J. Remote sensing image target detection based on improved YOLOv4. *J. Optoelectron. Laser* **2022**, *33*, 607–613.
49. Zhu, F.; Gao, J.; Yang, J.; Ye, N. Neighborhood linear discriminant analysis. *Pattern Recognit.* **2022**, *123*, 108422. [CrossRef]
50. Zhu, F.; Ning, Y.; Chen, X.; Zhao, Y.; Gang, Y. On removing potential redundant constraints for SVOR learning. *Appl. Soft Comput.* **2021**, *102*, 106941. [CrossRef]